How is Testing Supposed to Improve Schooling?¹

Edward Haertel

Stanford University

¹ This paper is based on the 2012 NCME Career Contributions Award address presented by the author on April 15, 2012 at the annual meeting of the National Council on Measurement in Education, in Vancouver, British Columbia.
Abstract

Validation research for educational achievement tests is often limited to an examination of intended test score interpretations. This paper calls for expansion of validation research along three dimensions. First, validation must attend to actual test use and its consequences, not just score meaning. Second, validation must attend to unintended as well as intended testing consequences. Third, validation must attend to indirect as well as direct testing effects. Indirect effects include the effect of score-based incentives in prompting actions intended to raise test scores (directing student effort or focusing the system) as well as messaging effects associated with a testing program per se but not dependent on specific scores (shaping perceptions). This expanded program of test validation can best be accomplished by measurement professionals working in collaboration with scholars from other social science disciplines.
How is Testing Supposed to Improve Schooling?²

Edward Haertel
Stanford University

Measurement professionals are obviously concerned with the question of how testing is supposed to improve schooling, although they might frame it more formally, asking first, with regard to a specific testing program, "What are the intended test score uses or interpretations?" Then, matters of test validity might be framed by asking, "What are the interpretive arguments and supporting validity arguments for those intended uses or interpretations?" These questions direct attention first to the intended use or purpose of the testing, then to the chain of reasoning from test scores to decisions or implications, and finally, to the evidence supporting or challenging the propositions in that chain of reasoning.

This validation framework has served the profession well in organizing the work of scientific inquiry into score meaning. Nonetheless, there have long been concerns that helpful as it has been, this familiar framework is somewhat incomplete. Twenty-five years ago, Cronbach (1988) called for something more in his examination of "Five perspectives on validity argument." The first of these, his “Functionalist Perspective,” directed attention toward questions of value or worth, reaching beyond a more comfortable, narrowly scientific focus on the truthfulness of score interpretations. He illustrated his functionalist perspective with the proposition that, “Tests used in the schools ought to encourage sound distribution of instructional and study time,” and cited cases where particular forms of testing had been found

² This paper is based on the 2012 NCME Career Contributions Award address presented by the author on April 15, 2012 at the annual meeting of the National Council on Measurement in Education, in Vancouver, British Columbia.
to have untoward effects on curriculum and instruction or on student learning (Cronbach, 1988, p. 5).

Messick (1989a, 1989b) also called for the expansion of test validation beyond the logic and evidence required to justify a specific, intended score interpretation. Echoing Cronbach, he called out “the functional worth of the testing in terms of its intended and unintended consequences” as one of four interrelated aspects of validation, and characterized validity, broadly speaking, as “an inductive summary of both the existing evidence for and the actual as well as potential consequences of score interpretation and use” (Messick, 1989a, p. 5).

Kane (2006) reaffirmed this expanded view of test validity. In his thoughtful analysis of potential fallacies in validity arguments, he called out in particular the unwarranted leap from evidence justifying a particular interpretation of test scores to the assertion that a particular use of test scores was thereby justified. This paper offers another way of framing these concerns and perhaps provoking some further reflection.

**Direct and Indirect Mechanisms of Action in Educational Testing**

Testing uses might be organized in various ways. One traditional category scheme distinguished uses for description, prediction, and explanation. Here, uses of educational achievement testing are organized instead according to seven broad purposes, as shown in Table 1. In this scheme, mechanisms of testing effects are first classified as *direct* versus *indirect* and then, within each of these categories, further classified according to what it is that the test scores are intended to describe. Many testing applications serve various purposes that might be located in two or more of these seven categories.
The direct categories are quite familiar. They encompass uses or interpretations that rely directly on the information scores provide about measured constructs. Measurement professionals have worked hard to understand and improve the work of validation for these categories, although as will be argued later, their focus has been more in score interpretations than on score uses. The indirect categories are also familiar, but less clear and not always explicit. As a working definition, indirect test uses or interpretations are mechanisms of action, leading to intended or unintended consequences, that do not depend directly on particular test scores. I distinguish two subcategories of indirect effects. First, there are incentive effects, which flow from deliberate efforts to raise test scores. Among other examples, these would include students' efforts to raise their own scores as well as teachers' efforts to raise their students' scores. Second, there are messaging effects, whereby testing is used to shape perceptions or understandings, again in ways that do not depend directly on the actual scores. Indirect testing effects typically fall outside the purview of professional efforts at test validation, even when they are included among intended, even explicit, theories of action whereby testing is expected to influence educational processes.

To illustrate these distinctions, consider the example of an elementary school teacher who gives a weekly spelling test covering that week's spelling word list. After scoring the test, he notes the words a lot of children missed, so that he can put those words back on the list again a few weeks later. He also notes each child's total score in his grade book. The children's
papers are returned to them so that they can see how well they did overall, and which words they spelled incorrectly. One reason the children study the weekly spelling word lists is because they know there will be a test, and they want to do well. This brief example mentions testing uses from several categories. Direct purposes include grading, providing diagnostic feedback to students, and planning future instruction. These uses all rely on test scores or summaries of item responses. Grading would be classified here as a use for measuring learners, and the others as uses for measuring learning. One indirect purpose is encouraging students to study their spelling words because they want to do well on the test. Their studying is of course a deliberate effort to raise test scores, hence an incentive effect, classified here as directing student effort. The teacher might also intend to send a message, shaping students' perceptions or understandings by using testing to convey the idea that knowing how to spell is important. In a general way, the amount that students choose to study might depend on their prior test performance, but one might argue that these indirect purposes depend only weakly, if at all, on particular outcomes of testing.

Kane (2006, p. 57) gives an example of the fallacy of "begging the question" that serves nicely as a second example:

The arguments for [current accountability] testing programs tend to claim that the program will lead to improvements in school effectiveness and student achievement by focusing the attention of school administrators, teachers, and students on demanding content. Yet, the validity arguments developed to support these ambitious claims typically attend only to the descriptive part of the interpretive argument (and often to only a part of that). The validity
evidence that is provided tends to focus on scoring and generalization to the
content domain for the test. The claim that the imposition of the accountability
requirements will improve the overall performance of schools and students is
taken for granted. (Kane, 2006, p. 57)

Once again, both direct and indirect purposes are alluded to here. Validity evidence
tends to focus on direct purposes, but "[improving] school effectiveness and student
achievement by focusing ... attention ... on demanding content" refers to an indirect mechanism
of action.

Validation of Test Uses As Well As Test Interpretations

In addition to neglect of interpretive arguments and supporting validity evidence for
indirect testing effects, there is a second, related way to describe how the measurement field
tends to fall short in validation efforts. Kane (2006) lays out four broad stages in most
interpretive arguments, namely scoring, generalization, extrapolation, and decision or
implication. Various technical aspects of alignment to test specifications, differential item
functioning (DIF), scaling, norming, and equating, support the interpretive argument for the
scoring stage; and studies of score precision, reliability, or generalizability address the
generalization stage. Measurement professionals typically devote considerably less attention to
stage three, extrapolation beyond the true score, universe score, or trait estimate, although
studies of test-criterion correlations would be located here. There is often little or no attention
to the stage four questions of how test scores are actually used or interpreted.

It might be argued, of course, that Kane's (2006) decision or implication stage falls
beyond the purview of test validation, or at least beyond the purview of the field of educational
measurement. Some testing professionals might argue that testing effects, whether direct or indirect, are simply not our concern. Despite the guidance provided in test manuals concerning appropriate and inappropriate test uses, the *Standards for Educational and Psychological Testing* (American Educational Research Association [AERA], American Psychological Association, & National Council on Measurement in Education, 1999, p. 111) acknowledge that “appropriate test use and sound interpretation of test scores are likely to remain primarily the responsibility of the test user.” Test users share responsibility for appropriate test use and sound test interpretation especially if a test is used for some purpose other than that for which it was validated (AERA, et al., 1999, p. 18, Standard 1.4).

Nonetheless, I believe that deeply understanding, as well as thoughtfully advocating for, appropriate test use as well as sound test interpretation is very much a part of the role of testing experts, and that more attention should be devoted to indirect as well as direct testing purposes. There is no bright line between technical concerns and practice or policy concerns in testing. A full consideration of validity requires that the interpretive argument be carried all the way through to the end, and that the full range of intended purposes be considered. It may not work very well for testing experts to see to scoring and generalizability and perhaps extrapolation, but then to hand off the test for someone else to worry about whether score-based decisions or inferences are truly justified, whether the test solves the problem it was intended to solve, and whether it creates unforeseen problems.

**Mechanisms of Action for Educational Testing**

The direct and indirect testing purposes in Table 1 are further divided into a total of seven broad functions or mechanisms whereby testing is intended to improve educational
outcomes. In the Direct Action column are listed (1) Instructional Guidance, (2) Student Placement and Selection, (3) Informing Comparisons Among Educational Approaches, and (4) Educational Management. For each of these purposes, test scores are taken as indicators of some underlying construct, and on that basis the scores are used to guide some decision or draw some implication. How accurately scores reflect underlying constructs is of the essence here. The four purposes are distinguished by what it is that test scores are used to describe: individual student's short-term learning progress (Learning), more stable student aptitudes or cumulative achievement (Learners), the effectiveness of different educational materials or approaches (Methods), or the effectiveness of teachers or schools (Actors or Organizations).

In the Indirect Action column are listed (5) Directing Student Effort, (6) Focusing the System, and (7) Shaping Perceptions. These are incentive and messaging effects distinguished according to what it is that testing and its associated consequences are intended to influence: student effort (Learners), curriculum and instruction (Methods), or perceptions (Actors). These indirect purposes all refer to testing consequences, but they should not be mapped onto Messick's (1989b) "consequential basis of test interpretation and use." Messick was drawing attention to normative considerations in testing, including the value implications of test interpretations and the social consequences of test uses. Normative issues arise in connection with indirect testing mechanisms, of course, as well as direct mechanisms. These issues are certainly important, but attention is here directed instead to effects of testing, both intended and unintended, that have no direct dependence on the information particular scores provide about underlying constructs.
As will become clear, still finer distinctions would be possible within these seven categories. It will also be obvious that testing and test scores are often used for several different purposes at the same time. In the remainder of this paper, these seven broad purposes are considered in greater detail.

**Instructional Guidance**

The purpose of instructional guidance is most often associated with narrowly focused achievement tests used to inform day-to-day instructional decisions within classrooms. This is the realm of formative testing, of the quizzes, unit tests, midterms and final examinations that teachers create or select for use in their own classrooms, used to assign grades and to guide the pacing of instruction. The primary users of test information for this purpose are teachers and students themselves. The constructs measured are usually narrowly focused and closely tied to the curriculum. Interpretations are mostly criterion-referenced, and mostly at the individual student level. Testing what has just been taught in order to provide feedback to students about their own learning and to help teachers plan instruction may sound straightforward, but of course that does not mean that it is simple or easy to do well. An interpretive argument is helpful in organizing potential concerns.

**Interpretive argument.** Following Kane, we might begin with scoring. There must be some way to quantify examinee performances to capture accurately the dimensions of interest. If scoring keys or rubrics credit only the particular answers a teacher intended, for example, and discount alternative answers that are also defensible, then there is a problem at that very first step. Obviously, scoring should be free from bias. Except on tests of handwriting, papers should not be marked down for poor penmanship alone, for example. Apart from obvious
flaws, perhaps the most critical scoring concerns here are whether the collection of questions adequately samples the intended content domain and whether the questions elicit the kinds of reasoning deemed important. It has been recognized at least since the time of the Bloom taxonomy (Bloom, Engelhart, Furst, Hill, & Krathwohl, 1956) that teachers tend to write low-level items calling for factual recall or routine application of learned procedures, rather than more complex questions demanding nonroutine problem solving or so-called higher-order thinking.³

The second broad step in Kane's interpretive argument is generalization. This is the traditional concern of test reliability, asking to what extent the scores obtained from the observed performance on a particular occasion are representative of the many scores that might have been obtained with alternate forms, with different scorers, on different occasions, and so forth. For classroom tests serving the instructional guidance function, there may be no need to standardize test administration or scoring across different times or places, but it is still important to consider whether tests are long enough and well enough designed that scores have adequate precision.

The third step, extrapolation, asks whether test scores are accurate indicators of proficiency or performance in some larger domain of situations, beyond the test or testing replications. We might ask, for example, whether the test scores are predictive of performance on different kinds of tasks that call for the same skills, including situations outside the classroom. The relation of tested content and skills to prior instruction becomes important

³ Writing multiple-choice questions that demand complex reasoning is a challenge not only for classroom teachers. The prevalence of low-level questions in the work of professional item writers was noted by Frederiksen (1984), and similar concerns were echoed in a recent report examining the mathematics exercises from the National Assessment of Educational Progress (Daro, Stancavage, Ortega, DeStefano, & Linn, 2007).
How is Testing here. Test tasks must not be completely familiar to examinees. It is difficult to imagine a cognitive item that could not be reduced to a test of rote recall by coaching students on that specific question. More subtly, classroom teachers may fool themselves and their students about their depth of understanding if the students are asked to demonstrate their knowledge using the same words or examples as they encountered during instruction. At the same time, a consequential examination should not be the students’ first opportunity to practice applying material learned to new problems or situations. Ideally, students will have had some prior opportunity to practice with whatever patterns of reasoning are called for on unit tests or final examinations.

The fourth step, decision or implication, addresses the ways test scores are actually used: What is going to be done differently or understood differently based on the score? Black and Wiliam (1998) emphasize that by their definition, assessment becomes formative only when it is actually used to adapt teaching work to meet learning needs. Decisions to provide extra review and practice or to advance a student to a higher-level course would be examples. Other possible decisions might include reteaching some material, skipping ahead in the syllabus, or contacting a student's parents concerning academic progress. If there is a formal decision rule determining what is to be done with the scores, involving a passing score, for example, then that rule also should be evaluated.

In practice, of course, it may be quite unclear what different actions are indicated by alternative testing outcomes. Data are of little use if teachers simply don't know what to do with them, or if they lack the time and other affordances needed to take action. "Implication" refers here to some inference or conclusion to be drawn. This might perhaps be conveyed by
assigning a grade. Again, there is a range of concerns, including the appropriateness of the decision rule. The point here is that there is little value in providing teachers with reports of student test performance, however nicely formatted, if they do not know and cannot figure out how to make use of those reports, or if they lack the capacity to take constructive action.

**Measurement-driven instructional systems.** Instructional guidance also encompasses more formal testing systems where the material to be learned is broken down into many narrow learning objectives with a prescribed sequence, often with several test forms available to assess each learning objective. Students' progress through the curriculum is controlled by their success on the corresponding tests. This has been tried many times. An early example was the *Winnetka Plan*, created by Professor Carleton Washburne and his colleagues at the University of Chicago in the early 1920s (Washburne, 1925). This was an elementary mathematics program that supplemented a textbook with mimeographed worksheets and tests. Pupils could take self-administered tests of each sequential objective, and then when they were ready, they could request a teacher-administered test. The date on which each pupil mastered each objective was recorded.

Decades later, this same pattern was found in *Programmed Instruction*, where content was presented in frames, each followed by test questions with branching to one frame or another according to whether the test score indicated a need for more work versus readiness to move on. Still later, Benjamin Bloom's (1968) *Mastery Learning* model followed a similar format, using what he called formative tests, similar to the criterion-referenced tests appearing from the late 1960s into the early 1980s. The Pittsburgh Learning Research and Development Center's Individually Prescribed Instruction Mathematics Project (Lindvall, Dillard, Hosticka,
LaPresta, & Light, 1976) offered yet another illustration. Measurement-driven instruction worked best for those areas of the curriculum that could be readily analyzed into a sequence of small independent units that built upon earlier units. Teaching basic arithmetic or phonics are good examples. It was less successful teaching more advanced content, where the precise patterns of reasoning required are less predictable.

Interpretive argument. These more formal instructional guidance models mostly finessed scoring, the first step of the interpretive argument, by making the constructs measured more-or-less isomorphic with answering the test questions. This fit well with the perspective of behaviorist psychology, which sought to avoid inferences about the psychological processes underlying observable behaviors. Robert Mager's influential book, Preparing Instructional Objectives (Mager, 1962, 1975, 1984) coached teachers on how to formulate instructional objectives in observable terms, using verbs like "arrange," "define," or "label" rather than verbs like "knows" or "understands." Well-written instructional objectives were readily translated into criterion-referenced tests.

Generalization, likewise, was not much of a problem. If a test consists of highly similar items asking students to do pretty much the same thing over and over, then any given student's performance is likely to be quite consistent from one test form to another, or even from one item to another on the same test. U-shaped score distributions were seen on some criterion-referenced tests, clearly distinguishing students who had mastered some narrow objective from
those who had not. That said, reliability coefficients for criterion-referenced tests were often low simply because there was little observed-score variance.\(^4\)

Extrapolation, the third step, was more problematical for criterion-referenced testing. As explained in Resnick and Resnick’s (1992) influential chapter on "Assessing the Thinking Curriculum," the measurement-driven instructional model and its behaviorist underpinnings relied on assumptions of decomposability and decontextualization. There was often insufficient attention to the question of whether students could summon up the right pieces of learning in new contexts and put them together in different ways to arrive at solutions to unfamiliar problems.

Decision or implication, the fourth step, typically relied on some simple cut score, usually defined by a required percent correct, to classify students as having mastered versus not having mastered the tested objective. There was some research on the problem of formulating these decision rules (Block, 1972), but general guidelines like "Mastery is 80% correct" were often adopted uncritically (Glass, 1978). And, as with the example from Kane (2006) cited earlier, the efficacy of the use of these tests for purposes of instructional guidance was too often taken for granted, with validation efforts focused instead on score interpretation.

**Measurement-driven instruction today.** Educational psychology has evolved since the era of behaviorism, and it would probably be a mistake to try and resurrect these testing and instructional approaches. Nonetheless, measurement-driven instructional models did have some positive features. They did not work well for all kinds of learning objectives, but were

\(^4\) Reliability coefficients fit well with norm-referenced test interpretations, but are often less useful with criterion-referenced tests. In generally, the standard error of measurement is a more appropriate statistic for describing score precision.
How is Testing

useful in promoting some important learning outcomes. The tests used in these systems were closely aligned to a specific curriculum, and it was always clear that the curriculum came first and the tests, second. It was also clear what the tests were supposed to be used for. Students either moved on directly to the next small unit of instruction or else they were given help before moving on. These many, precise instructional decisions for each pupil may have been overly mechanical, but the systems were workable. Communications around test performance were primarily about mastery, not about ranking or comparing one student to another. We could do worse. The instructional guidance purpose lives on in interim or benchmark assessments, of course, but these are generally designed by working backwards from the summative test, and are administered at fixed intervals (Perie, Marion, & Gong, 2007). That is a very different model.

Student Placement and Selection

Whereas testing for instructional guidance focused on measuring learning, testing applications for purposes of student placement or selection involve inferences about learners. This category includes testing for purposes of guiding decisions about student ability grouping, determining entry into and exit from classifications like "English learner," and making college admissions decisions, among other uses. It also includes certification tests like high school exit examinations, as well as Advanced Placement and International Baccalaureate tests and examinations. Compared to formative tests for instructional guidance, tests designed for placement and selection purposes generally measure constructs broader in referent generality (Coan, 1964; Snow, 1980). Aptitude tests make an appearance, and where achievement tests are employed, they are intended to represent broader content domains and are used, like
aptitude tests, to predict future performance or readiness to profit from some instructional setting considerably more remote than the next brief instructional unit.

The shift in focus from learning to learners necessarily directs attention away from individual students’ progress through the curriculum and toward individual differences among learners. Formal systems for measurement-driven instruction were often predicated on the notion that almost all students could master the curriculum if given the supports they needed. When measuring learning, growth is expected as learning progresses. Testing applications measuring learners, for purposes of placement or selection, focus on differences among students at a given point in time, and may be predicated on the notion that children differ significantly in their capabilities relevant to school learning. The characteristics measured are often regarded as stable over time—Even as all children progress through the curriculum, their ranking according to academic achievement may remain much the same.

Today, users of tests for placement or selection are likely to recognize the powerful role of environment and opportunity to learn in explaining both individual and group differences in aptitude test scores or achievement outcomes. However, a century ago such differences were more often seen as reflecting differences in innate capacity.

Student IQ testing in the early 20th century. Perhaps the extreme view of aptitude as fixed and immutable was represented in the student IQ testing used early in the last century. Following on the perceived success of the Army Alpha during World War I, IQ tests were widely used for tracking students (Chapman, 1988). Tracking was intended to make schooling more efficient by giving teachers more homogeneous groups of students to work with. That way,
slower children would not be pushed to the point of frustration, and quicker children would not be held back.

The validity of IQ tests may have been accepted in part because their results seemed to confirm prevailing stereotypes about racial and ethnic differences in intelligence. Group differences in test scores provided an all-too-convenient explanation for what are today referred to as between-group achievement gaps. Chapman (1988) points out that tracking was already widespread before IQ tests became available, but the tests supported the practice, made it seem more objective and scientific, and fit comfortably with prevailing beliefs about intelligence and individual differences.

Whereas tests used for instructional guidance are often carefully integrated into the school curriculum, IQ tests were intended to have zero dependence on the school curriculum. They were designed to tap into different aspects of general mental ability, using progressively more difficult tasks of various types to determine the examinee’s “mental age,” from which an IQ could be calculated. Scoring rules were carefully worked out, and generalization was shown by high correlations between alternate test forms as well as high stability over time. Extrapolation was assumed on the basis of a strong theory about the importance of intelligence in determining fitness for different occupations and other aspects of adult life, as well as success in school. At the time, however, the implications drawn from IQ test scores and the justification for decisions based on those scores were largely unexamined. It was simply assumed that intelligence was largely inborn and little affected by age or experience.

Just as psychology has moved beyond the behaviorist notions underlying elaborate formal models for instructional guidance, it has also moved beyond the strongly hereditarian
views that justified IQ testing. Today, environmental influences on IQ test scores are better understood, and the idea of "culture-free" testing has long since been abandoned. It is also more generally recognized that tracking can lead to a self-fulfilling prophecy whereby the groups offered a stronger, faster-paced curriculum progress more quickly than their peers in lower academic tracks. Nonetheless, this model too may offer some useful lessons. Every teacher knows that some students tend to catch on more quickly than others. In addition to students’ different patterns of strengths and weaknesses, there are also general tendencies whereby those stronger in one subject area are likely to be stronger in others. There seems to be no acknowledgement of this reality in the current rhetoric of school reform. Of course all children can learn, and nearly all have the potential to become literate, successful members of society. But in U.S. educational policy over the past two decades, the mantra of "all children can learn" has led to an insistence on very high common standards for all children. The seeming denial of large individual differences in aptitude for school work has distorted education policy, whatever the origins of those differences might be.

Contemporary placement and selection testing applications. The era of routine student IQ testing is past, but tests are still widely used for student placement and selection. Current applications feature stronger connections between tests and curricula, with the Advanced Placement and International Baccalaureate testing programs as major examples. The ACT, used for college selection, is designed to measure of academic achievement as defined by a broad curriculum syllabus. Even the SAT, which was traditionally regarded as an aptitude measure, is now described as a test of proficiency in academic work, developed through formal schooling over a period of years.
Informing Comparisons Among Educational Approaches

Instructional guidance had to do with measuring *learning*, and student placement and selection had to do with measuring *learners*. The third category listed in Table 1 includes tests used in evaluations of curricula or of alternative instructional approaches. These have to do with measuring the outcomes of different instructional *methods*. This purpose can be traced back at least to the Project Head Start evaluations mandated by the Elementary and Secondary Education Act of 1965 and the large-scale evaluations of curricula created with National Science Foundation funding in the post-Sputnik era, around that same time (Walker & Schaffarzick, 1974). More recently, the National Diffusion Network, in existence from 1974 to 1995, attempted to synthesize evaluations documenting successful educational approaches, and the What Works Clearinghouse, established in 2002, has a similar mission (Haertel & Herman, 2005).

These evaluations may be regarded as testing applications because they use achievement tests as outcomes. The purpose in each case is to compare methods in order to discover which works best. This is one case where the interpretive argument really is carried through to the end, or *almost* to the end. There is usually a final, taken-for-granted assumption that the approach found superior in the sites involved in the evaluation will also give superior results if it is widely disseminated, and of course that cannot be guaranteed. These "informing comparisons" purposes are best implemented with randomized controlled trials, but the category would include testing applications in studies with a wide range of research designs.5

---

5 The study by Dee and Jacob (2011) of the impact of the No Child Left Behind Act on student achievement offers an elegant example in which NAEP data were used for the purpose of evaluating the NCLB approach to school accountability.
Educational Management

My fourth category of purposes, referred to in Table 1 as educational management, refers to uses of achievement tests to describe or to guide decisions about teachers, schools, principals, or other actors and organizations in the educational system. Unlike program evaluations, which used test scores to characterize instructional methods or materials, educational management applications employ student achievement test scores to characterize specific schools, teachers, and so forth. Teacher or school scores are typically used to identify problematical cases requiring intervention, not to determine best practices. Even though these scores may sometimes be used to identify exemplary schools or teachers from which others might learn, the primary aim is usually to establish accountability.6

The No Child Left Behind Act of 2001 (NCLB) is a prominent example. Here, student test scores from annual state-level testing systems are used in a complicated way to calculate scores for schools and then to determine whether each school is meeting a goal called Adequate Yearly Progress (AYP). There is a rational mechanism whereby schools not meeting AYP targets are supposed to receive remediation leading to improvement. Rational reliance on the information provided by school and teacher indices derived from student test scores is one key to this accountability model’s theories of action, but direct mechanisms are only part of the story. Another major piece of the rationale whereby NCLB was supposed to improve schooling would be situated the sixth category in Table 1, that of focusing the system by directing energy and effort toward improvement in measured schooling outcomes. The two purposes of

6 The National Assessment of Educational Progress (NAEP) would be classified here as an Educational Management application designed to characterize achievement in the nation, states, and large school districts. Although a possible audit function for NAEP is cited in the No Child Left Behind Act of 2001, its main use is for description, not accountability.
educational management and focusing the system interact under NCLB, in that the pressure to focus is most intense for schools failing to meet AYP.

Turning to another example, perhaps the most prominent current applications of educational management mechanisms are seen in applications of "value-added" models (VAM) for teacher evaluation. Teacher VAM scores are used to identify specific teachers for follow-up actions, which might include, among others, awarding merit pay, requiring mentoring, or dismissal.

As with nearly all of the testing uses discussed in this paper, validation efforts for teacher value-added models have focused primarily on the first two steps of the interpretive argument. The scoring step is familiar, although with some value-added models, vertical scaling assumptions place more stringent demands on scores for VAM than for other testing uses. The generalization step also has an extra wrinkle. When student scores over time are used in the aggregate to construct scores for teachers, the reliability of individual student-level scores becomes less important while other sources of error, related to student assignment and sampling, must also be considered. Only limited attention has been devoted to extrapolation, the question of how strongly value-added scores relate to broader notions of teaching quality or effectiveness. Likewise, there has been little attention to the decision or implication step, addressing the question of whether actions taken on the basis of value-added scores have the intended effects.

---

7 The recent Measures of Effective Teaching (MET) study funded by the Bill & Melinda Gates Foundation is an important exception. MET study results now appearing do address the relation of teacher value-added scores to various other measures of teacher effectiveness. Reports are available at the MET study website, http://www.metproject.org.
As with the NCLB example, rationales for teacher VAM also invoke indirect category six purposes of focusing the system as well as direct category four purposes of educational management. Obviously, the use of student achievement test scores for high-stakes evaluation of teachers is likely to focus teacher effort on raising test scores. But the purposes for teacher value-added models may also cross over into seventh category in Table 1, that of shaping perceptions. It seems quite likely that highly publicized reports of individual teachers' value-added scores, and the highly publicized reactions to these reports coming from teachers' unions and from various political actors are influencing public perceptions of teachers and the teaching profession. My own view is that in some cases this may be an intended effect of these systems, but that would probably be difficult to establish empirically.

**Indirect Mechanisms of Action in Educational Testing**

The indirect purposes of directing student effort, focusing the system, and shaping perceptions are rarely if ever explicitly set forth as the sole intended mechanisms of action for testing applications, although they are frequently stated as secondary purposes, together with direct uses of test score information. Nonetheless, I believe that these incentive and messaging effects may be more significant than direct uses of score information in the working of many large-scale testing programs, especially test-based accountability systems.

**Incentive effects.** The fifth and sixth purposes, directing student effort and focusing the system, rely on the motivation of students, teachers or other actors in the educational system to improve test scores. The consequences of such efforts, intended or unintended, depend critically on the content of the tests used and the relation of tested content to test takers’ prior instruction. Black and Wiliam (1998) clearly document in their review of formative assessment
that the ways testing is done and test results are communicated can influence what, how, and how much students study. The question, "Is it going to be on the test?" has become a cliché. Likewise, high school exit examinations may influence student effort or behavior, though not uniformly, and not always as intended (Jacob, 2001).

As Cronbach (1988) observed long ago, consequential testing has the power to shape not only student effort, but also curriculum and instruction. Such incentive effects can be of positive value. Participating in test construction or in the scoring of performance assessments is sometimes regarded as a kind of in-service training for teachers, leading them over time to teach in different ways (e.g., Resnick & Resnick, 1992). As another example, performance assessments themselves might exemplify valued but underutilized instructional approaches. Or, high-stakes tests covering just a few school subjects might encourage teachers to spend more time on those subjects and less time on others. Or, inclusion of questions about specific topics on a high-stakes test might be intended to help assure that those topics are addressed in classrooms. At the same time, using high-stakes tests to focus the system on specific subjects or specific kinds of knowledge and skills carries substantial risks of unintended consequences, including score inflation and distortion of curriculum coverage (Koretz, 2008; Shepard, 2000).

If intended positive effects are to be realized and unintended negative effects are to be avoided, careful attention to test content is critical. This reality poses a pervasive challenge for standardized testing where there is no common curriculum. Reasoning has to be about something. An ideal task for eliciting complex reasoning or problem solving might draw upon a rich base of prior knowledge and skills, but require using that prior learning in new ways. In order to test students' understanding of irony or foreshadowing, for example, it is helpful to
know a particular book they have read. Test questions can then refer to specific events from
different parts of the book and ask how they are connected or why the author placed them in a
particular sequence. If it is known only that the students have read some books, but their
familiarity with a particular book cannot be assumed, then selected-response test tasks must
instead be self-contained packages, each one bundling a question with whatever background
material might be required to arrive at the answer.  

Selected-response questions can be useful, of course, and can measure some important
kinds of learning. But, one can ask much deeper questions knowing which Shakespeare play
students have read than are possible simply knowing that they have read a Shakespeare play.
Likewise in history, knowing what specific events, historical figures, or historical documents
students have studied enables the examiner to pose better questions to measure historical
thinking. In ecology, one can ask deeper questions knowing which particular ecosystem
students have learned about.

So-called content standards cannot substitute for curriculum as a basis for test design.
The Common Core State Standards in English language arts, for example, call for students to
have read a Shakespeare play, but don’t say which one (Common Core State Standards
Initiative, 2010). In many subject areas, curriculum-neutral assessments employing selected-
response items cannot drill down very far on complex reasoning, simply because there is so
little that can be assumed concerning specific, relevant, common background knowledge
students should have had a chance to acquire in the classroom.

---

8 Students’ understanding of irony or foreshadowing could also be evaluated with open-ended constructed
response questions requiring respondents to provide specific examples from whatever material they were familiar
with, but such items would require human scoring, which would be much more expensive than machine scoring.
Messaging effects. The seventh category in Table 1 covers the use of tests to shape perceptions, including perceptions of public education and of the teaching profession. Such perceptions matter because they influence actions like supporting school bond issues or choosing alternatives to regular public schools. It has been argued that test results have been used systematically to bolster perceptions that our public education system functions poorly (Berliner & Biddle, 1995). As another example, it is commonly said that among the principal users of school accountability scores in California are real estate agents who can quantify the effect on neighborhood housing prices of the local school’s Academic Performance Index.\(^9\)

Testing may also be used to convey the message that an elected leader cares about education and is doing something about it. Many presidents and governors have embraced the cause of educational improvement. Robert Linn (2000, p. 4) explained testing’s appeal for public officials as follows:

Test results can be reported to the press. Poor results in the beginning are desirable for policymakers who want to show they have had an effect. Based on past experience, policymakers can reasonably expect increases in scores in the first few years of a program ... with or without real improvement in the broader achievement constructs that tests and assessments are intended to measure. The resulting overly rosy picture that is painted by short-term gains observed in most new testing programs gives the impression of improvement right on schedule for the next election. (Linn, 2000, p. 4)

\(^9\) Perceptions of a particular school might arguably be based on specific scores, which would be a direct, not an indirect, effect. But I decided to include this use as an example of an indirect effect because a school’s reputation over time seems not to be closely tied to any specific individuals' scores on specific tests.
Testing can also be a form of symbolic action. How many more times will we hear about a new testing program to identify low-performing groups so that remediation can be targeted? Are the results ever surprising?

Test Validation and Indirect Mechanisms of Action

It is a commonplace of the measurement profession that validation is directed to the uses or interpretations of test scores, not to tests themselves. Expanding the scope of test validation to include indirect testing effects is challenging, but I would argue that it is nonetheless essential. Indirect mechanisms of action are often critical components of the interpretive arguments explaining the benefits expected and intended from large-scale testing applications. They come at the end of the chain of reasoning, often serving as the ultimate rationale for the entire testing program. Excellent validity studies supporting the scoring and generalization stages of the interpretive argument are important, of course, but a chain of reasoning is only as strong as its weakest link. The interpretive argument must be carried through to its conclusion.

Examining indirect purposes is also important because it is in connection with these purposes that many unintended consequences of testing arise. Using tests to focus the system is a prime example. When curriculum-neutral tests are used to promote key learning objectives, the result may be superficial coverage of just that subset of learning objectives included in the test specification. If there were well worked out, widely understood and widely used methods for studying such effects, such unintended consequences would receive more attention and might in time come to be better understood by policy makers and the public at large.
Challenges in validating indirect mechanisms of action. The difficulties of examining indirect testing purposes are formidable for at least five reasons. First, these purposes may not be clearly articulated. Second, necessary data may not become available until after a testing program has been implemented. Third, the research methods needed to carry out the required studies may be unfamiliar to testing professionals. Fourth, the agencies and actors best positioned to carry out this work have disincentives for doing so. Finally, this work is expensive, and it may not be clear that the answers will really matter. Nonetheless, acknowledging all these challenges, I would argue that measurement professionals can and should do a better job than they have historically.

Clear statement of indirect mechanisms of action. The lack of clarity around influencing purposes is not insurmountable. Major purposes can generally be discerned, and articulating these purposes more clearly could help. As an example, the procurement for state consortia's comprehensive assessment systems under U.S. Department of Education’s “Race to the Top” funding competition required that applications include "a theory of action that describes in detail the causal relationships between specific actions or strategies in the eligible applicant’s proposed project and its desired outcomes for the proposed project, including improvement in student achievement and college- and career-readiness" (U.S. Department of Education, 2010, p. 71). Similar language can now be found in other federal requests for proposals.

Availability of data. Data availability can be greatly improved simply by planning evaluation studies before new testing initiatives are launched. Baseline data must collected in advance. Such data collections might employ audit tests aligned to outcomes, but could also include surveys on content coverage or allocation of instructional time to different subject
areas, for example. If evaluation is an afterthought, it may be too late to do it properly. Phased implementation is also a helpful strategy. If changes are made everywhere at the same time, it may be impossible to isolate the effects of those changes on schooling outcomes.

**Expanding the range of research approaches used in test validation.** With regard to familiarity with the range of research methods appropriate to the study of indirect testing effects, measurement professionals might look to their colleagues across the social sciences for assistance. The measurement field draws most heavily from the disciplines of statistics and psychology, especially the psychology of individual differences. Training in these fields may leave testing experts better equipped to study phenomena at the level of individual examinees, versus systemic effects on organizations or patterns of interaction among individuals. As a thought experiment, reflect for a moment on the sorts of questions an anthropologist, a philosopher, or a political scientist might ask about standardized achievement testing. For example: Is uniform standardization the best or only way to approach the ideals of fairness and score comparability? How does testing support belief in a meritocratic system where individual effort is rewarded? How has testing shaped popular views of the quality of public education?

Validation might benefit from more cross-disciplinary dialogue about testing, and from the inclusion of distinct perspectives from sociologists, anthropologists, economists, linguists, curriculum specialists, and teacher educators, among many others.

Some progress could be made by simply reasoning through the likely responses of school principals, teachers, and students to the incentive structures shaped by testing. However, more than common sense is called for. As Messick (1989b, p. 13) argued long ago, validity must be supported by both empirical evidence and theoretical rationales. The
measurement field has developed bodies of theory to address the descriptive stages of interpretive arguments—investigating what test scores mean. However, when it comes to the ways testing is actually supposed to function out in the world, the bodies of theory within the purview of measurement per se are impoverished. Measurement may borrow a bit from economics, as when test scores are regarded as providing information essential for rational consumer choice, but insights from organizational theory, social psychology, and other social science disciplines are also required.

**Costs and incentives for better test validation.** My last two challenges had to do with costs and incentives for doing a better job of validating test purposes and uses, especially indirect mechanisms of action. It is all well and good to lament the confirmationist bias in test validation, as in much social science research. It is understandable that testing companies and state education agencies might be reluctant to subject their tests and their policies to serious, critical scrutiny. But evaluation requirements built into testing procurements might help. The procurement for the SMARTER Balanced and PARCC assessment consortia was cited earlier as an example. The evaluative criteria for those proposals allocated 30 points to bidders' Research and Evaluation Plans, which required not only research on the properties of the tests themselves, but also a plan "for determining whether the assessments are being implemented as designed and the theory of action is being realized, including whether the intended effects on individuals and institutions are being achieved" (U.S. Department of Education, 2010, p. 37). Obviously, the two consortia each submitted plans responsive to the RFP's requirements. Similar requirements in requests for proposals from other agencies would be a step in the right direction. With experience, such evaluation requirements might be elaborated and refined to
the point where serious attention to indirect as well as direct purposes of testing, and serious attention to actual uses and their consequences as well as to score interpretations, became an expected part of test validation. That will not happen overnight, but it is a goal worth striving for.

Summary

Testing has been used in many ways to further educational goals. It has become an indispensable tool in guiding day-to-day decision making in classrooms, in communicating learning goals and evaluating progress toward their attainment, in student placement and selection, in identifying best practices, and in monitoring the effectiveness of the educational system. Inevitably, as with any powerful policy tool, testing has also at times brought unintended consequences.

In order to maximize the benefits and minimize any negative effects of educational testing applications, test validation must attend to indirect as well as direct effects of testing, both implicit and explicit. Unintended as well as intended testing consequences must be examined. Scholars from many social science disciplines can contribute significantly to this important work.
References


Table 1. Seven uses for educational tests.

<table>
<thead>
<tr>
<th>Who or What is Described</th>
<th>Direct Action</th>
<th>Indirect Action</th>
</tr>
</thead>
<tbody>
<tr>
<td>Learning</td>
<td>Instructional Guidance</td>
<td></td>
</tr>
<tr>
<td>Learners</td>
<td>Student Placement and Selection</td>
<td>Directing Student Effort</td>
</tr>
<tr>
<td>Methods</td>
<td>Informing Comparisons Among Educational Approaches</td>
<td>Focusing the System</td>
</tr>
<tr>
<td>Actors or Organizations</td>
<td>Educational Management</td>
<td>Shaping Perceptions</td>
</tr>
</tbody>
</table>