Tests, Test Scores, and Constructs

Edward H. Haertel

To cite this article: Edward H. Haertel (2018) Tests, Test Scores, and Constructs, Educational Psychologist, 53:3, 203-216, DOI: 10.1080/00461520.2018.1476868

To link to this article: https://doi.org/10.1080/00461520.2018.1476868

Published online: 29 Jun 2018.

Submit your article to this journal

Article views: 381

View Crossmark data

Citing articles: 1 View citing articles
In the service of educational accountability, student achievement tests are being used to measure constructs quite unlike those envisioned by test developers. Scores are compared to cut points to create classifications like “proficient”; scores are combined over time to measure growth; student scores are aggregated to measure the effectiveness of teachers, schools, and school districts; indices are created to measure college and career readiness. These and other new uses rely on derived scores created to measure new constructs. The field of educational and psychological measurement has largely ignored these significant, consequential measurement applications. The conceptual frameworks and analytical tools of educational and psychological measurement should be used to study such derived scores and the validity of their uses and interpretations.

I have been trying for a long time to understand how meaning is imparted to test scores. As a graduate student at the University of Chicago, I assisted my doctoral advisor, John Bormuth, in his work on formal methods for deriving test questions from texts, based on linguistic analysis. Our goal was to devise a rigorous, rule-governed procedure for creating tests from reading texts, leaving no room for item writer subjectivity. The hope was that with such tests, an examinee’s raw score would have a lawful relationship to the source text’s comprehensibility for that examinee, described by a mathematical function. If such a function could be determined, then a test could be derived from any text and used to determine whether that text was too easy, too difficult, or just right for any examinee. That would be an empirically grounded, criterion-referenced score interpretation.

Meanwhile, Benjamin Bloom, also a member of my doctoral committee, was promoting formative testing to guide day-to-day instructional decisions, but with no firm basis for deciding what score should mark the boundary between “nonmaster” and “master.” I also took courses with Benjamin Wright, who interpreted Rasch scaling as providing a kind of score objectivity not attainable with any other item response model (Wright, 1967). All these different ideas led to one of my early publications, in which I wrestled with the problems of formulating and validating criterion-referenced score interpretations (Haertel, 1985).


I was still troubled by what seemed to me a shaky foundation for criterion-referenced score interpretations, especially standards-based interpretations. In 2004, almost twenty years after my first paper on criterion-referenced testing, my student William Lori and I revisited the validation of criterion-referenced score interpretations in the context of standard setting. We distinguished several different ways of criterion referencing scores and set forth an interpretive argument for cut score validation (Haertel & Lorié, 2004).
Beginning around 2005, I tried repeatedly to organize and classify different kinds of test uses in education (e.g., Haertel & Herman, 2005; Pullin & Haertel, 2008). I finally settled on a scheme that distinguished among measurements targeting learning (as with formative tests in classrooms), learners (as with annual standardized achievement tests), methods (as with outcome measures in educational program evaluations), and actors or organizations (as with accountability testing). At the same time, I distinguished between theories of action relying on direct versus indirect effects of testing (Haertel, 2013a). Up until that point, however, I was still focusing mainly on intended test score uses and interpretations as conceived by test developers.

Most recently, I’ve turned my attention to uses of test scores devised by legislators and other policy makers, and to the lack of systematic theorizing and empirical research on these increasingly important test uses. It is to that topic that I turn in the remainder of this article. My message is that, as a profession, we have some work to do. New testing applications, especially in educational policy, have gotten out ahead of our field’s accepted practices in validating test score uses and interpretations. Testing is no longer being employed merely to measure the effects of policy interventions—rather, it has itself become the intervention, as policymakers have sought to harness efforts to raise test scores in the service of educational reform. There is a pressing need to scrutinize these new, policy-driven test uses through the lens of measurement theory.

In this article, evolving notions of validation are briefly summarized, and several examples are then offered to show how interpretation and use arguments (IUAs) can guide validation efforts. These ideas are next applied to some high-stakes applications of testing for educational accountability that have emerged within the past 30 years, illustrating new test score uses very different from those envisioned by the original test developers. These new uses require new IUAs. As is shown, they often rely upon the creation of test scores intended to tap constructs quite distinct from individual students’ proficiencies at a single point in time. The article concludes with an appeal for more research using the conceptual frameworks and analytical tools of our discipline to call out misuses of tests and to guide wiser testing practice.

**Evolving Notions of Test Validation**

Validation is at the heart of our work as measurement specialists and as educational psychologists concerned with testing and test use. Our conceptions of validity have evolved over time. We ask more questions about tests themselves, how test scores are used, how that use is supposed to be beneficial, and what might go wrong. As our list of questions has grown, so too has the set of research methods used to address them.

In broad overview, the field of educational and psychological measurement has moved from a narrow view of validation as the collection of test–criterion correlations, still sometimes referred to as “validity coefficients,” to a view of validation as scientific inquiry into the meanings given to test scores, and then on to a still further expanded view of validation as the evaluation of both test score interpretations and test score uses, marshalling theory and empirical evidence in a connected argument. Our challenge today is to bring this framework to bear on varieties of test scores typically not even thought of as being test scores at all. Applications of testing for accountability have stretched the notion of derived test scores well beyond alternative characterizations of single test performances by individual examinees. Many of these accountability applications rely on incentive effects, incorporating into theories of action the direct efforts of students, teachers, or administrators to raise test scores. Framing and pursuing rigorous programs of test validation for these new test uses and interpretations will require a further expansion of our methodological repertoire.

*Educational Measurement* has served for decades as a major handbook for our field. Its successive editions have offered snapshots of the field’s conception of validity when each was published. Edward E. Cureton’s (1951) chapter “Validity” appeared in the first edition. Cureton defined validity as “the correlation between the actual test scores and the ‘true’ criterion scores” (p. 623), in other words, the test–criterion correlation corrected for attenuation due to unreliability in the criterion measure. His concern was predictive validity in specific testing situations. At that time, predictive validity was pretty much the whole story, one testing application at a time. The methodology for test validation was essentially calculating correlation coefficients. Of course, calculating a correlation coefficient was a lot more work in 1951 than it is today.

Lee J. Cronbach’s (1971) chapter “Test Validation” appeared in the second edition. Cronbach greatly expanded upon Cureton’s view, making room for description and explanation alongside prediction. He emphasized that “one validates, not a test, but an interpretation of data arising from a specified procedure” (Cronbach, 1971, p. 447) and that test validation required the integration of different kinds of evidence, depending on the circumstances. He described a range of analytical methods, especially in connection with construct validity. Correlations still had a place, but beyond “validity coefficients,” they were now used to examine convergent and discriminant validity (Campbell & Fiske, 1959) and in factor analysis. Variance components calculated using random-effects analysis of variance models were discussed, and experimental designs made an appearance, including Cronbach’s “duplicate
construction experiment” to see if two teams working from the same test specification would come up with anything like parallel tests. For Cronbach in 1971, validation was essentially scientific inquiry into score meaning. Nonetheless, he did raise issues calling for value judgments. In discussing content validity, he stated that the importance of the content measured should be considered. He also urged test developers to consider whether the names given to their test score scales invited overly broad score interpretations.

Samuel Messick’s (1989b) chapter “Validity” expanded the scope of test validation still further. “Scientific inquiry into score meaning,” which for Cronbach (1971) was essentially the whole of test validation, appeared in Messick’s framework as inquiry into the adequacy of inferences based on test scores. With that as a starting point, Messick then broadened the scope of validation to cover scientific inquiry into both the adequacy and the appropriateness of both inferences and actions based on test scores. Thus, as shown in Table 1, test use (“actions”) had a place alongside test interpretation (“inferences”), and consequential basis (“appropriateness”) appeared as an equal partner with evidential basis (“adequacy”).

Messick (1989b) argued strongly for more attention to values in test validation, including what he referred to as the social consequences of test interpretation and use. Cronbach’s nod toward considering the importance of the content tested and the implications of test titles was expanded by Messick to include discussions of relevance and representativeness with respect to curriculum and instruction (in achievement testing) or job requirements (in employment testing).1 Notions of construct-irrelevant variance and construct underrepresentation were called out, and significant space was devoted to the problem of determining the relevance of prior validation research to a test use in a new setting (validity generalization). Messick expanded the range of inquiry methods required for test validation to include much greater emphasis on logical argumentation and even philosophical inquiry, and he mentioned dialectical and rhetorical methods. He emphasized that validity was not all-or-none but rather a matter of degree and that it required an “integrated, evaluative judgment” (Messick, 1989b, p. 13) to pull together both empirical evidence and theoretical rationales.

Finally, Michael T. Kane (2006) took on the challenge of writing the chapter titled “Validation” for the most recent, fourth edition of Educational Measurement. Kane followed Messick in the view that what is validated is a test score interpretation or use, and he embraced Messick’s explicit consideration of values in validation. One major contribution of Kane’s chapter was his guidance for organizing the work of validating test score uses and interpretations. Even by the time of Messick’s chapter in the late 1980s, many kinds of questions and analyses had been subsumed under the heading of test validity. By the time Kane was writing, ideas of test validity had expanded still further (e.g., Borsboom, Mellenbergh, & van Heerden, 2004; Gipps, 1999; Haertel & Lorié, 2004; Messick, 1994, 1998; Moss, 2003; Ryan, 2002; Shepard, 1993). Messick’s (1989b) formulation of validity was rich and comprehensive but said little about actually designing and carrying out a program of validity research (Shepard, 1993). By the time of the fourth edition of Educational Measurement, the need for such practical guidance was pressing.

Kane (2006) took a pragmatic approach, delving into the process of test validation and offering a road map, of sorts, for the work validators do. He began by dividing that process into two phases. During the development phase, the test developer builds a case for the way the test is supposed to work. For this phase, a confirmationist bias is acceptable. Test development is often an iterative process, identifying and fixing problems to arrive at a measurement tool that works as intended. During the appraisal phase, after development is complete, the validation takes a more neutral or even critical turn. Evidence is assembled and scrutinized concerning claims about the meanings of test scores and about appropriate test score uses. Not just the test developer but also test users and other stakeholders may participate. It is during this phase that validators look for hidden assumptions and consider alternative explanations for good or poor test performance.

TABLE 1
Samuel Messick’s Framework for Test Score Validation

<table>
<thead>
<tr>
<th>(Inferences) Test Interpretation</th>
<th>(Actions) Test Use</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Adequacy) Evidential basis</td>
<td>Construct validity (CV)</td>
</tr>
<tr>
<td>(Appropriateness) Consequential basis</td>
<td>CV + Value Implications (VI)</td>
</tr>
</tbody>
</table>


1By the latter half of the 1980s, Cronbach had also moved in the direction of a broader conception of test validity, as had others in the field (e.g., Anastasi, 1986; Cronbach, 1988).
developmental, phase, the primary focus is an interpretive argument. This is the chain of logic or the network of propositions connecting test performances to conclusions or decisions. For the second, appraisal, phase, the focus shifts more to the validity argument. The validity argument is an evaluation of the interpretive argument. It examines the coherence of the interpretive argument and weighs the evidence for and against the various propositions the interpretive argument entails. Kane gave several illustrations of interpretive arguments organized into four broad stages, although he stressed that other organizations could also work. His general scheme began with the scoring stage, followed by generalization, then extrapolation, and finally use or interpretation (see Table 2).

Scoring refers to the connection between a particular examinee performance and the resulting test score. The test score should be an accurate and undistorted reflection of the relevant qualities demonstrated by a particular test performance. To the extent that the test score also reflects other qualities, construct-irrelevant variance may be a concern. So, for example, the story problems on a math test should not be too difficult to read so that math scores are not distorted by variation in students’ reading proficiency that is irrelevant to the intent of the measurement. Checks for item bias and for test bias might also be incorporated here.

Generalization is the connection between the actual test score and some hypothetical universe of scores that might have been obtained instead, on different occasions, say, or with different test forms, or with different raters. It is at this stage that score reliability or generalizability is considered.

Extrapolation moves out further still, to the broader domain of performances or situations where the measurement ought to matter, beyond test taking per se. It is at this stage, primarily, that construct underrepresentation might come into play. So, for example, a reading test that only requires selecting responses to factual questions may leave out important reading skills needed in real-world settings.

Use or interpretation gets to the question of what is actually done with the test scores—their relevance and usefulness, as well as the appropriateness of decisions or other actions they support. What this means, exactly, and how the rationale for testing is framed, varies from one situation to another. I give some examples next.

Kane (2006, 2013) made the point that in practice, test validation has generally focused heavily on scoring and generalization but has largely neglected extrapolation and use or interpretation. He cited state-level school accountability programs as one relevant example, observing that the arguments for these … 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 … [focusing] 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 [emphasis added]. (Kane, 2006, p. 57)

This is a problem. The chain of argument is as strong as its weakest link. Robust support for some parts of the interpretive argument—high test reliability, for example—cannot make up for lack of support for other critical parts of the argument.

In summary, Kane laid out a general scheme for interpretive arguments, reaching all the way from a test performance to the ultimate decision, description, prediction, or inference a test use is intended to support. During the appraisal phase of test validation, it is the job of test developers, test users, and other interested parties to build a validity argument that critically examines all the key propositions in that interpretive argument. Many kinds of studies could yield evidence supporting or challenging the propositions included in the interpretive argument, and the mix of studies chosen in designing a validation plan will vary from one testing application to another. Validation may proceed iteratively, with successive decisions informed by prior findings. Cronbach (1988) drew attention to political and economic perspectives that may

<table>
<thead>
<tr>
<th>Stage of IUA</th>
<th>Description of Stage</th>
<th>Some Relevant Validity Concerns</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Scoring</td>
<td>From structured observations to scores</td>
<td>Bias? Construct-irrelevant variance?</td>
</tr>
<tr>
<td>2. Generalization</td>
<td>From scores to (hypothetical) replications (universe score)</td>
<td>Random error?</td>
</tr>
<tr>
<td>3. Extrapolation</td>
<td>From universe score to broader domain of application</td>
<td>Relation to other measures of construct? Construct underrepresentation?</td>
</tr>
<tr>
<td>4. Use or interpretation</td>
<td>Actual use of score in some testing application</td>
<td>Relevance and utility? Theory of action?</td>
</tr>
</tbody>
</table>

inform validation design choices, as well as more narrowly scientific (explanatory), utilitarian (functional), and domain sampling (operationist) perspectives. He noted that “tests that honestly report facts may disserve student development and social progress” and went on to assert that “tests that impinge on the rights and life chances of individuals are inherently disputable” (Cronbach, 1988, pp. 5–6, italics omitted). Messick (1989b, 1994) likewise offered helpful guidelines, but no prescriptions, for assembling evidence of score validity for specified uses and interpretations.

In this article I focus mainly on interpretation and use arguments, not validity arguments, but I can at least suggest some varieties of studies that might support or challenge propositions at each stage, with the strong caution that these are merely illustrative and may not be appropriate in all circumstances. That said, propositions at the scoring stage might call for direct observation and modeling of the processes underlying test responses, perhaps using think-aloud protocols or examinee debriefs (stimulated recall). For this stage, studies of differential item functioning, test bias, factor analyses of tests’ internal structure, and expert reviews of content relevance and representativeness might also be used. For the generalization stage, evidence might come from traditional analyses of test reliability as well as from generalizability studies. Score stability over time, interrater reliability, and comparisons across test forms would be relevant here. For the extrapolation stage, correlations or regression analyses might be used to probe the relation of test scores to alternative measures of the same or different constructs. Campbell and Fiske’s (1959) multitrait-multimethod matrices might be used to organize and interpret patterns of correlations. Finally, for the stage of use or interpretation, empirical studies of test bias as well as evaluations or case studies might be called for. Throughout, validation efforts would ideally attend to significant examinee subpopulations as well as the population as a whole.

Unfortunately, as stated previously, validation efforts often fall short. The evidence assembled often addresses only propositions at the beginning of the chain—the scoring and generalization stages—to the neglect of propositions toward the end of the argument, including extrapolation beyond test taking per se, but especially the soundness of rationales for the actual uses or interpretations of the test scores. Kane (2013) later renamed his interpretive argument the Interpretation and Use Argument, further calling attention to the need for validation of the entire logical chain from the examinee’s test performance all the way to the score-based decision or action ultimately taken. Kane (2006, 2013) was concerned primarily with interpretive arguments pertaining to single test performances by individual examinees, but as argued next, the same framework can be helpful in considering scores derived from multiple observations of one examinee, such as individual growth measures; or from scores of multiple examinees, such as some school or teacher effectiveness measures.

HISTORICAL EXAMPLES OF IUAs

Put simply, an IUA is a theory of action that explains how testing is supposed to be of benefit in some particular situation. Three brief examples may serve to clarify this notion. The first uses tests to measure stable student attributes for the purpose of sorting or selection, the second uses tests to measure narrow elements of learning for instructional management, and the third uses tests to measure achievement outcomes in order to compare alternative curricula or instructional treatments. Additional examples are presented elsewhere (Haertel, 2013a, 2013b; Haertel & Herman, 2005; Haertel & Ho, 2016; Haertel & Lorié, 2004; Kane, 2006).

IQ-Based Tracking

In the early 20th century, one popular use of testing was to give intelligence tests to schoolchildren and place them in separate tracks, some higher, some lower, according to their “Intelligence Quotient” (IQ) test scores (Chapman, 1988). That turned out to be a bad idea for several reasons, and it was largely abandoned long ago. Nonetheless, it serves as a good IUA illustration. Proponents of IQ-based tracking generally embraced assumptions that IQ tests are objective and accurate (scoring and generalization stages) and that the quality they measure is predictive of learning ability in school settings (extrapolation stage). Support for assumptions about IQ tests’ predictive power might be both empirical (e.g., correlations of IQ test scores with school grades) and theoretical (e.g., psychological theories positing that IQ is innate and largely fixed at birth). In addition, IQ-based tracking proponents implicitly assumed that education was essentially direct instruction, whereby a teacher transmits information to students, ideally at a rate matched to the students’ capacity for absorbing that information (use or interpretation stage). These assumptions justified the grouping of children with similar IQ test scores so that teachers could work more efficiently, with more homogeneous student groups, and so individual students would not be bored or frustrated by an instructional pace that was either too slow or too fast for them. Research challenging or supporting these and other propositions in the IUA would form the validity argument.

Of course, much more needs to be said. Among other factors, the simple IUA just sketched ignores the influence of culturally transmitted knowledge on IQ test performance, the variability of IQ test performance over time, the social
context of school learning, the multidimensional character of human cognitive abilities, the range of schooling outcomes beyond factual knowledge acquisition, and the range of pedagogical approaches beyond whole-class direct instruction. As considerations like these have come to be better understood, IQ-based tracking has fallen out of favor.

Measurement-Driven Instruction

A distinctive IUA emerged in the mid-20th century, with various test-driven instructional management systems. I have referred to this elsewhere as the use of tests to measure learning rather than learners (Haertel, 2013a). An early example was the Winnetka Plan for teaching arithmetic (Washburne, 1925). This was a self-paced instructional management system that used tests of narrow learning objectives to track individual students’ progress through the curriculum. Students could take a nonsecure practice test covering some learning objective whenever they felt ready, and then if they did well, they could ask the teacher for a secure test to confirm their readiness to go on to the next objective and repeat the process. Later examples based on a similar logic included programmed instruction (e.g., Lumsdaine & Glaser, 1960) and Benjamin Bloom’s (1976) Mastery Learning. This IUA was most prominent during the criterion-referenced testing movement (e.g., Popham, 1971), which made heavy use of tests keyed to narrow behavioral objectives. Like IQ-based tracking, these testing uses relied on a set of common assumptions. As argued by Resnick and Resnick (1992), two of these assumptions concerned the decomposability and the decontextualization of learning—ideas that material to be learned can be broken down into small pieces (decomposability), which can then be taught in isolation and later recombined to accomplish complex tasks in different contexts (decontextualization). The decomposability assumption that narrow criterion-referenced tests in fact measure discrete abilities might be examined at the scoring stage of the IUA, and the decontextualization argument that these narrow skills are relevant in other settings might be examined at the extrapolation stage. Once more, research on these and other assumptions would support, or challenge, the IUA undergirding measurement-driven instructional management systems.

Program Evaluation

Yet another kind of interpretive argument is seen in the use of tests as outcome variables for program evaluations or for comparisons of curricula or instructional methods. This kind of testing use came into prominence in the 1960s, with the mandate for program evaluations written into the Elementary and Secondary Education Act (ESEA, 1965), as well as federally funded evaluations of new science curricula sponsored by the National Science Foundation after the Soviet Union’s 1957 launch of Sputnik (Walker & Schaffarzick, 1974). It continues today with the What Works Clearinghouse. This interpretive argument requires a series of assumptions about the qualities measured by the outcome tests used, their precision, and their broader import (scoring, generalization, and extrapolation stages of the IUA) that together help assure a fair comparison of different programs or curricula. The use or interpretation stage of the IUA for such testing applications might bring in additional assumptions that outcome measures are equally well aligned with each of the alternatives being compared (perhaps requiring multiple tests, one favoring each alternative); that the groups tested are equivalent except for their participation in one instructional treatment versus another; and that the evaluation itself does not prompt changes in instruction, like teaching to the test. Once more, the validity argument would provide evidence for or against these and other propositions forming the IUA.

**DIRECT VERSUS INDIRECT EFFECTS OF TESTING**

An obvious feature of the IUAs just described is that each relies directly upon information provided by test scores. One obtains a score, then takes some action or performs some calculation based on that score’s numerical value. However, testing can also have effects that are not mediated by actual scores. When a classroom teacher says, “This will be on the test,” for example, that announcement is intended to influence student behavior even before the test is given. This is an example of an indirect testing effect. It is a path of influence that does not depend directly on actual test performance. Two kinds of indirect testing effects may be distinguished: incentive effects and messaging effects (Haertel, 2013a).

Incentive Effects

Like the teacher’s announcement that something will be on the test, incentive effects use testing to spur actions intended to raise scores. Another example would be using high school exit exams to push some students to study harder. Testing may also create incentives for changes in teacher behavior, such as greater efforts to cover the curricular content included on a high-stakes examination. This power of testing to influence curriculum and instruction has not gone unnoticed. Policymakers have repeatedly sought to exploit testing incentive effects in the service of educational reform but with very limited success.

---

2See https://ies.ed.gov/ncee/wwc/.
Incentive effects and the limited success of past efforts to harness that power, various initiatives promoted performance assessment as a way to improve curriculum, instruction, and learning outcomes. A slogan popular at the time was, “What You Test Is What You Get.” High-stakes multiple-choice tests were blamed for driving schooling toward an emphasis on low-level, decontextualized information, and performance assessment was seen as the key to driving schooling in the opposite direction (Haertel, 1999; Kirst & Mazzeo, 1996; Resnick & Resnick, 1992; Simmons & Resnick, 1993). Alas, performance assessments proved to be expensive to develop, time-consuming to administer, and less reliable than more familiar paper-and-pencil tests. Moreover, hoped-for improvements in student learning seemed slow to materialize. Enthusiasm for performance assessment soon waned, but reliance on high-stakes testing as a tool of educational reform persisted.

State accountability systems focused on school performance began appearing in the early 1990s. These represented a new educational reform model that sought to leverage testing incentive effects by using scores aggregated to the level of schools and districts. One early example was the Kentucky Instructional Results Improvement System created in response to the Kentucky Educational Reform Act of 1990 (Guskey, 1994). Other states soon followed, spurred in part by school accountability requirements for Title I students that were included in the 1994 reauthorization of the ESEA, titled the Improving America’s Schools Act (IASA, 1994). Since that time, tests have been used to measure not only learners, learning, or for program evaluation but also to measure and compare the effectiveness of schools and school districts in promoting student achievement (Haertel, 2013a). This explicit reliance on an indirect testing effect marked a highly significant change: Instead of using tests to evaluate educational treatments, testing itself had now become the treatment.

In his comment concerning current accountability programs, noted earlier, Kane (2006) explained, “The arguments for these … 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” (p. 57). His observation nicely illustrates the idea of incentive effects. Stating that testing will lead to improvements by focusing attention on demanding content is much like a teacher’s focusing students’ attention with an announcement that “this will be on the test.” Of course, this is just one among several mechanisms whereby accountability testing is expected to promote educational improvement, but when such indirect testing effects are an explicit part of testing rationales, I would argue that they should be called out in the IUA and subjected to validation. Studying such effects is particularly important because it is here that many unintended testing consequences will be found. Incentivizing attention to valued content may also incentivize neglect of untested content or of additional learning outcomes even more important than those tested (Koretz, 2017).

Messaging Effects

In addition to incentive effects, there is a second, more subtle category of indirect testing effects, namely, messaging effects. These include mechanisms of influence whereby media coverage of test results may shape popular perceptions of public schooling or of the teaching profession. Elliott Eisner, for many years my colleague at Stanford, remarked from time to time that the fastest way to provoke an educational crisis was to give a test and then announce that half the children scored below average. The long-standing, popular narrative of public schools in crisis has been fueled by accounts of poor U.S. test performance relative to other nations and relative to cut scores designating “basic” or “proficient” performance on the National Assessment of Educational Progress (NAEP), for example, as well as by flat or declining test score trends over time (Barshay, 2018; Berliner & Biddle, 1995; Carnoy & Rothstein, 2013; National Commission on Excellence in Education, 1983).

Messaging effects may also contribute to elected officials’ enthusiasm for new testing programs. Alluding to the typical pattern of score inflation over the first few years that a new, high-stakes test is in use, Linn (2000) observed that “results are visible [and] can be reported to the press. … [The] 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” (p. 4). Thus, testing per se can convey the message that elected officials are effectively addressing educational concerns.3

3The positive public relations message at the time a new testing regime is announced would clearly be an indirect testing effect. The impression of improvement a few years later would depend on scores having risen, and so might not be regarded as a purely indirect effect.
DERIVED SCORES

A school measurement is a formal procedure for assigning numerical values to schools in a manner intended to reflect some property of each school—some dimension on which schools can then be evaluated or compared. In school accountability systems, these school measurements are calculated starting from student test scores. In other words, student test scores are used to create a new set of derived scores, one for each school (Haertel & Ho, 2016).

The idea of derived scores is familiar. A derived score is created whenever an existing score is reexpressed on a different scale. This might be done simply to make alternative test forms comparable, using scale scores to account for slight variations in form difficulty. Often, however, derived scores are used to add meaning, not just to improve comparability. For example, percentile ranks or grade equivalents add meaning by locating a score relative to some comparison group.

Most familiar derived scores, including scale scores, percentile ranks, and grade equivalents, are functions of a single raw score for an individual examinee. Derived scores can also be more complicated, however. A pretest score might be subtracted from a posttest score to produce a gain score, or a more elaborate procedure might produce a Student Growth Percentile or some alternative growth score might be subtracted from a posttest score to produce scores can also be more complicated, however. A pretest percentile ranks, and grade equivalents, are functions of a score might be subtracted from a posttest score to produce a gain score, or a more elaborate procedure might produce a Student Growth Percentile or some alternative growth measure (Betebenner, 2008; Castellano & Ho, 2013). Student test scores can also be combined to create derived scores describing their teachers or schools, as with teacher “value-added” models, or school accountability systems (Haertel, 2013b; Haertel & Ho, 2016).

Various recent, consequential testing applications have created and employed complex derived scores with little attention to their validity (Haertel & Ho, 2016). In addition to school scores for accountability systems, examples include the designation of students as English learners, indices intended to predict college and career readiness, teacher effectiveness scores derived using value-added models, and Adequate Yearly Progress (AYP) determinations under the No Child Left Behind Act (NCLB, 2002). The binary classification of students as either English learners or not creates a kind of derived score. In practice, this classification depends not only on English language proficiency test scores but also on complex and variable rules determining when students exit from English learner status (National Research Council, 2011a; Robinson, 2011). College and career readiness indicators (e.g., U.S. Department of Education, 2010) invite predictive interpretations that are typically under supported (Haertel & Ho, 2016). Serious concerns with the validity of teacher evaluations based on value-added models are discussed elsewhere (Haertel, 2013b).

ADEQUATE YEARLY PROGRESS

One of the prime examples from Haertel and Ho (2016) was the classification of schools as either making AYP or being “In Need of Improvement,” as mandated by the NCLB (2002). This annual determination can be thought of as a binary score, like assigning each school a 1 for AYP or a 0 for “In Need of Improvement.” The process of creating and using these scores was in some ways the culmination of all the school accountability systems that started cropping up in the 1990s. The AYP illustrates some mechanisms whereby complex derived scores have been built up layer by layer from simpler scores. Assumptions woven into their interpretive arguments cry out for validation.

Thinking of AYP status as a derived score prompts questions like, “What construct does it measure?” “How reliable is it?” “What might be some sources of construct irrelevant variance?” and “How about construct underrepresentation?” These and similar questions can be addressed systematically by applying Kane’s (2006, 2013) interpretive argument schema. Spoiler alert: The validity argument is not strong. It is not at all clear just what school characteristics the “AYP” label actually captures, and as explained next, escalating “Annual Measurable Objectives” (AMOs) systematically changed the meaning of AYP over time. In any case, there was no doubt but that making AYP was better than being found “In Need of Improvement.” Under NCLB, schools “In Need of Improvement” year after year faced escalating sanctions, up to and including staff dismissal and school reconstitution.

Haertel and Ho (2016) identified two additional derived scores created en route from student test scores to school AYP determinations, namely, student achievement level designations (e.g., “basic” or “proficient”) and school-level percent-proficient statistics. Each of these arguably qualifies as a derived score because it is reported, interpreted, and used as a measurement reflecting a characteristic of either students or schools. In broad outline, the steps along the path from student test scores through student achievement levels and school percent-proficient statistics and then on to school AYP determinations were as follows.

Student-Level Test Scores

Under NCLB, in each state, reading and math tests were administered to all public school students in Grades 3–8. Typically, validity evidence for these scores would be provided by the original test developers. Except for an additional review to document the tests’ alignment to the state’s academic content standards, this validation would address traditional, student-level test score uses and interpretations—important, but just the beginning.
Student Achievement Level Designations

Student achievement test scores were then compared to a series of cut points to classify each student as “below basic,” “basic,” “proficient,” or “advanced” in reading and in mathematics. Cut points were established independently by each state, on whatever accountability tests that state was using. Students were designated “proficient,” for example, if their scores fell between the “proficient” cut point and the “advanced” cut point. These classifications qualify as a derived score because they are supposed to add meaning to the interpretation of student test performances by indicating whether a student is or is not meeting some performance expectation. Proficient and advanced students are regarded as doing okay, whereas basic or below basic students are regarded as falling short.

Validation of achievement level determinations is addressed in Haertel and Lorié (2004), where the following three concerns are among those discussed. First, although the measurement field is far from consensus, various researchers have argued that the standard setting methods used to define cut points like “proficient” are seriously flawed (e.g., Glass, 1978b; Haertel & Lorié, 2004; National Academy of Education, 1993; National Research Council, 1999). Study after study has shown that different standard-setting methods, purporting to yield estimates of the same quantity, in fact produce systematically different results (e.g., Jaeger, 1989, Table 14.1, pp. 498–499). In addition, there has been little attention to the psychological processes at play when standard-setting panelists work toward consensus (Fitzpatrick, 1989). Second, a label like “proficient” carries considerable surplus meaning. The prose describing what proficient students know or are able to do, as well as common-sense understandings of the word proficient, may go well beyond what can actually be supported by empirical evidence or theoretical rationales. Third, because the NCLB legislation required all states to use the label “proficient” even though they used different tests and established different cut scores, the label “proficient” has no uniform meaning (National Research Council, 2017). Its common use therefore invites inaccurate comparisons of proficiency percentages among states and between state tests and NAEP.

School-Level Percent-Proficient Statistics

These difficulties notwithstanding, the next step in the AYP derivation is to tabulate the proportions of “proficient” students in each school, pooling across grade levels, separately for reading and for mathematics. Percent-proficient values are calculated and reported not only for each school as a whole but also for specified student subgroups within each school, provided these subgroups are sufficiently large. Note that this aggregation represents a shift from derived scores for students (their achievement level determinations) to a new set of “percent-proficient” derived scores describing not students but schools. Once more, because these school-level statistics are reported and interpreted, their interpretations also call for validation. They inherit the weaknesses of the student-level derived scores they are based on. Their reliability may also be subject to challenge because in addition to the measurement error in individual students’ test scores they are affected by sampling error. The students present in a school in a given year may happen to include more high scorers or more low scorers than in some other year, and this will affect “percent-proficient” scores, especially for small schools and small student subgroups within schools.

Reliability aside, a major concern with the use of these percent-proficient numbers to describe school quality or school effectiveness is construct-irrelevant variance. A school’s “percent proficient” is largely determined by the characteristics of the students that school serves. A school serving many English learners or students from less affluent families, or a high school with a special mission like dropout recovery, will have a lower “percent proficient” than one serving students from relatively advantaged backgrounds. Even within a single state, reporting percent-proficient statistics for schools invites faulty inferences about school quality and misleading school comparisons. Comparisons of percent-proficient numbers based on the different tests used in different states are essentially meaningless without some correction for the differential stringency of states’ standards (Braun & Qian, 2008).

In addition to construct-irrelevant variance, when policymakers or the media treat these numbers as telling us all we need to know about school quality, there is also a deeply troubling problem with construct underrepresentation. We pay lip service to the idea that many valued schooling outcomes are left unmeasured, then return to discussions of test scores as if they are all that matter. As Dylan Wiliam (1998, p. 1) said, “Put crudely, we start out with the intention of making the important measurable, and end up making the measurable important.” Finally, as explained by Ho (2008), serious statistical problems arise when these percent-proficient numbers are used to quantify achievement gaps between groups or, worse, changes in achievement gaps over time.

School AYP Determinations

Moving on from these school-level derived scores, under NCLB, the percent-proficient values were next compared to criterion values referred to AMOs. If every one of the school’s various percent-proficient numbers for different groups, for both reading and mathematics, met or exceeded that year’s AMO, then all was probably well. Otherwise, the school was probably in need of improvement. The qualifier “probably” is needed because there were some
additional requirements for making AYP, as well as some additional rules to lessen the AYP burden for lower scoring schools making good progress over time.

The first AMOs, for the year 2002, were determined by a complex formula specified in legislation, based on statewide percent-proficient numbers for subgroups along with the state’s distribution of percent-proficient numbers for schools (NCLB, 2002). States were then required to specify in advance how they would set their AMOs for each year from 2003 through 2014, ramping up from that 2002 initial value to 100% proficient by 2014. There were some constraints on how this could be done, but a lot of states went with a pattern that came to be called the “balloon mortgage” option (e.g., Risberg, 2011, p. 901), whereby the AMOs rose very gradually during the early years of the period and then suddenly shot up to 100% as the year 2014 approached (e.g., Risberg, 2011; Wiley, Mathis, & Garcia, 2005). This seemed like a good strategy back in 2002, because the ESEA was supposed to be up for reauthorization in 2007, at which point the accountability rules were supposed to be revised, at the very least pushing back the deadline for reaching 100% proficiency. The balloon mortgage pattern caused big problems, however, as NCLB remained the law of the land for far longer than expected. Reauthorization was postponed by Congress, year after year, until the passage of the Every Student Succeeds Act (2015–2016). In the meantime, we went from from no student being left behind to virtually no schools making adequate progress.

Consider the implications of rising AMOs for the meaning of the binary school-level AYP determination, conceived as a kind of derived score. Whatever construct AYP might have been measuring in the early years of the NCLB era, that meaning was essentially lost as 2014 approached and the cut points defining adequate percent-proficient statistics rose to unrealistic extremes. In defining their patterns of AMOs from 2002 through 2014, states were in effect determining, a priori, how the meaning of the AYP designation was going to change over time. This notion of a time-dependent construct definition would be quite foreign to most measurements, although it does have a rough parallel in developmental scales like IQ, where a given level of test performance maps to different scores as a function of the examinee’s age. In the case of IQ scores, of course, those mapping functions are grounded in both psychological theory and extensive empirical research, not invented in response to a legal mandate.4

Actual AYP determinations were even more complicated than this brief summary suggests. There was a rule that at least 95% of the students in each school and in each numerically significant subgroup within that school had to participate in testing, which in turn entailed rules for counting students to calculate those testing participation rates. There were special “one percent” and “two percent” rules for students with disabilities; “safe harbor” provisions and “margin of error” adjustments; and often, additional decision rules for determining for how many years students could still be counted as English learners for accountability purposes after they had been reclassified as fully English proficient (NCLB, 2002; Wong, Wing, Martin, & Krishnamachari, 2018).

IUA for AYP

This step-by-step analysis has thus far focused on the mechanics of school AYP determinations, together with some discussion of what Messick (1989b) termed the evidential basis of test interpretation, asking what this binary school-level indicator actually means or signifies. Similar questions were raised about two intermediate scores generated along the way, namely, student achievement-level determinations and school-level reports of percentages at or above proficient. Important as these questions of score meaning are, they are merely the beginnings of the IUAs for these derived scores. In the appraisal stage of validation, it may be even more important to consider the evidential and consequential bases of test use, which are often neglected. If NCLB actually spurred educational improvement, technical shortcomings might be forgiven. Unfortunately, it appears that positive effects were weak and spotty (e.g., Dee & Jacob, 2011; National Research Council, 2011c). Meanwhile, evidence has been found for negative effects like excessive test preparation of dubious educational value, distortions of curriculum in favor of tested content, and educational triage focusing on so-called bubble kids thought to be just below the proficient threshold. These effects were unintended, of course, but widely anticipated (Diamond & Cooper, 2007; Koretz, 2017).

A CALL FOR ACTION

As educational psychologists and measurement specialists, we are rightly concerned when tests are used irresponsibly. We are heirs to a decades-long tradition concerned with the validation of test score uses and interpretations. Our ways of thinking about these matters, and our repertoires of research approaches and of statistical tools for addressing them, have expanded enormously since Cureton wrote about validity coefficients in 1951. This has occurred in part because we have become more sophisticated in our understanding of threats to validity; in part because we have been challenged to examine the

---

4As observed by Andrew Ho (personal communication, February 15, 2018), the measurement field has given disproportionate attention to standard setting for students versus standard setting for schools (which is what AMO setting is). We would never dream of subjecting students to standards defined like the AMO-baseline-setting formula for schools or like states’ subsequent definitions of AMOs through 2014.
ways testing practices have sometimes functioned to justify or support systemic social inequality; and in part because we are, in fact, seeing fundamentally new ways of using tests and testing, calling for new IUAs that in turn demand support from new kinds of empirical evidence and theoretical rationales.

We have made great strides and have much to be proud of, but I am concerned that our ways of thinking about validity are falling behind some of the new ways tests are now being used; we have some catching up to do. Achievement tests are being put to uses vastly different from those envisioned by their developers. Scores for individual examinees are being interpreted relative to judgmentally determined cut scores. New kinds of derived scores are being created by combining current and prior-year scores to measure growth. Complex functions of student scores are being used as school-level accountability scores. Complex functions of student residualized gain scores are being used for teacher evaluation.

Formulating IUAs for New Test Score Uses

These new test uses cry out for scholarly attention, beginning with the questions of what constructs these new derived scores are intended to assess and whether they do so adequately. We can demonstrate how concepts including score reliability, construct-irrelevant variance, construct underrepresentation, and bias can be brought to bear to illuminate the strengths and weaknesses of the testing applications in which these new derived scores are deployed. Most important, we can work to clarify just how these testing uses are intended to improve schooling, whether they are succeeding, and whether their benefits outweigh any unintended negative consequences.

Current conceptions of validation point the way. The first step is to flesh out the IUAs for new test uses. Evaluating the coherence of the interpretive argument is part of the validation effort, bringing both empirical evidence and theoretical rationales to bear (Cronbach, 1988; Haertel & Ho, 2016; Kane, 2006, 2013; Messick, 1989b, 1998). Key propositions with weak support must be called out for scrutiny so that evidence for and against them can be assembled and weighed. Many individual investigators may focus on one issue or another, contributing to the validation effort without taking on the whole task. Others may assume the role of synthesizers, framing the full argument and assembling and weighing extant evidence from the research literature.

Actual Versus Intended Score Interpretations

Carrying these new IUAs all the way to their conclusions will require attention to actual, not just intended, score interpretations. In a recent publication, O’Leary, Hattie, and Griffin (2017) made a strong case for routinely examining actual interpretations and uses of test scores as part of validation. Although these authors did not discuss standards-based reporting, misinterpretations of descriptors like “proficient” offer prime examples in support of their concerns. Journalists and nonspecialist audiences have been all too willing to take the word proficient at face value. One such case is the reporting of NAEP results in terms of achievement levels. A recent evaluation of the achievement levels found “a disconnect between the kind of validity evidence that has been collected and the kinds of interpretations and uses that are made of NAEP’s reported results” (National Research Council, 2017, p. 9).

That evaluation went on to conclude that “[NAEP] achievement levels are widely disseminated … but the interpretive guidance about the meaning and appropriate uses of those levels … is inconsistent and piecemeal. Without appropriate guidance, misuses are likely” (National Research Council, 2017, p. 13). Consistent with O’Leary et al.’s recommendations, the National Research Council evaluation report went on to call for research on both intended and actual uses and interpretations of achievement levels, as well as clear guidance as to both substantiated and unsubstantiated interpretations.

Explaining Test Limitations

Perhaps the greatest challenge for our profession lies in communicating the limitations of testing and test scores to nontechnical audiences. Policymakers have embraced testing as a reform tool. It is no longer being used simply in evaluating education reforms but has instead become the linchpin of the reform itself. Aspirations for accountability testing have vastly outpaced what such testing can deliver, and policymakers may be reluctant to accept the news that rapidly rising scores over the first few years of some new high-stakes testing regime cannot be taken at face value (Koretz, 2017).

More than 25 years ago, Shavelson, Baxter, and Pine (1992) contrasted “political rhetoric” with “measurement reality” and concluded, “In the final analysis, we suspect that this nation may be placing far too much weight on accountability to achieve its reform agenda” (p. 26). Over 12 years ago, Braun and Mislevy (2005) coined the phrase “intuitive test theory” to refer to a set of naïve popular misconceptions about score comparability, precision, and related technical matters. They wrote bluntly,

We should not shy away from critiquing policies and programs that are based on intuitive test theory. This involves telling lots of people (some of them very important) that what they want to do won’t work and that doing something right is harder or takes longer than they might like. (p. 497)
That challenge is still very much with us. As measurement aspirations rise toward prediction of “College and Career Readiness” and the measurement of “21st Century Skills,” we have got to do a better job of explaining the limitations of what can be measured and just how imprecise our measurements really are (National Research Council, 2011b). As Braun and Mislevy (2005) reminded us, the idea of measurement error “is not a natural part of everyday reasoning about test scores (with the major exception that occurs when someone’s score is lower than he or she expected)” (p. 493).

Need for an Interdisciplinary Approach

As psychologists and measurement specialists, we may find that carrying out this work calls for the use of unfamiliar research methods and tools, as well as collaborations with researchers in other social science disciplines. Document analysis, interviews, and focus groups may be necessary to flesh out interpretive arguments, especially when pinning down anticipated indirect testing effects. The same methods may be used to investigate the interpretability of score reports so as to anticipate likely misinterpretations. Sophisticated econometric models may be used to probe causal claims about new programs for accountability or teacher evaluation. Teacher logs or surveys may be used to document changes in the allocation of instructional time in response to accountability pressures.

There is much work for us and our students to do, taking up the challenges of advancing our knowledge and understanding of how tests are actually being used, investigating the validity of these uses, and actively working to advance public understanding so that educational measurement comes to be used more wisely and well.

ACKNOWLEDGMENTS

This article is based on the author’s 2016 E. L. Thorndike Award address, presented August 5, 2017, at the annual meeting of the American Psychological Association, Washington, DC. I gratefully acknowledge Andrew Ho’s comments and suggestions on a draft of this article. I am of course responsible for deficiencies and any errors that may remain.

REFERENCES


