Adapting the Practice of Measurement to the Demands of Test-Based Accountability

A working paper of the Education Accountability Project at the Harvard Graduate School of Education
http://projects.iq.harvard.edu/eap

Daniel Koretz
Harvard Graduate School of Education

Revised, December 5, 2013

© 2013 by the author. All rights reserved.

The research reported here was supported by the Institute of Education Sciences, U.S. Department of Education through Grant R305A110420 to the President and Fellows of Harvard College. The opinions expressed are those of the authors and do not represent views of the Institute or the U.S. Department of Education.
Adapting to accountability

Abstract

Accountability has become a primary function of large-scale testing in the U.S. The pressure on educators to raise scores is vastly greater than it was several decades ago. Research has shown that high-stakes testing can generate behavioral responses that inflate scores, often severely. I argue that because of these responses, using tests for accountability necessitates major changes in the practices of educational measurement. The needed changes span the entire testing endeavor. This paper addresses implications for design, linking, and validation. It offers suggestions about possible new approaches and calls for research evaluating them.
Over the past several decades, accountability has become a primary function—arguably, the single most important function—of large-scale educational testing in the U.S. The transition has been gradual, dating back at least to the minimum-competency testing movement of the 1970s, and the nature of the accountability systems and the characteristics of the assessments used for this purpose have varied through several waves of policy initiatives (Koretz & Hamilton, 2006). Nonetheless, the pressure on educators to raise scores has increased markedly from one wave of initiatives to the next. The current situation, in which rewards and sanctions for schools based on scores are ubiquitous and consequential evaluations of teachers based on students’ scores are in place or planned in a large number of states, would have been almost inconceivable forty years ago.

The premise of this paper is that the use of tests for accountability requires major changes to the practices of educational measurement. The needed changes span the full sequence of measurement practices, beginning with test design, continuing with the activities needed to maintain testing programs, in particular, linking, and ending with validation.

In this paper, I focus primarily on behavioral responses to test-based accountability (TBA) and their implications for the validity of score-based inferences. To keep this discussion reasonable in length, I will not discuss the evaluation of other effects of TBA, although I agree with Haertel (2013) that more extensive evaluation of impact is essential. Therefore, for the sake of simplicity and clarity, I will use the term validity only to refer to the extent to which score-based inferences are warranted. I will not discuss
Adapting to accountability

implications for further research except insofar as they are directly relevant to the changes I discuss, e.g., additional validation.

**Traditional Psychometrics**

Although high-stakes testing existed long before the science of psychometrics, much of the development of standardized educational testing occurred during a period of relatively low stakes and was premised on an implicit assumption of low stakes. First, traditional psychometrics gave relatively little attention to the consequences of testing. Second, traditional psychometrics did not focus substantially on behavioral responses to testing, other than administrative behaviors and the *intended* responses of students and teachers. The field gave little attention to behavioral responses to testing that might threaten validity, such as inappropriate narrowing of instruction or coaching students to capitalize on incidental attributes of items.

Warnings that behavioral responses to testing might undermine validity are nonetheless not new. For example, more than 60 years ago, E. F. Lindquist, one of the most important figures in the early development of standardized achievement tests, offered this warning:

> The widespread and continued use of a test [that is a proxy for an unattainable measure of criterion behaviors] will, in itself, tend to reduce the correlation between the test series and the criterion series for the population involved. Because of the nature and potency of the rewards and penalties associated in actual practice with high and low achievement test scores of students, the behavior measured by a widely used test tends in itself to become the real objective of instruction, to the neglect of the
(different) behavior with which the ultimate objective is concerned

(Lindquist, 1951, 152-153).

Lindquist offered this warning in an era of very low stakes. Today, when the well-being of educators and sometimes students depends directly on scores, and when the scores required to avoid sanctions are often high relative to current performance, the incentives to behave as Lindquist warned are far stronger.

Despite early warnings and the dramatic increase in pressure to respond to testing in ways that can undermine validity, most of the practice of educational measurement has proceeded without substantial attention to this problem.

**Changes since the 1970s**

Recent decades have seen an increasing focus on the effects of testing in theoretical discussions of validity. The notion that consequences of testing are relevant to validity stretches back at least to the middle of the last century (Kane, 2006; Shepard, 1997). However, this notion became more central to validity theory over the past several decades (e.g., Haertel, 2013; Linn, 1997; Messick, 1989; Shepard, 1997). While much of this discussion has focused on the effects of testing as important outcomes in their own right, some scholars also began to focus on the risk that high-stakes testing might induce behavioral responses that bias inferences about achievement. Madaus (1988) and Shepard (1988) both warned of the risk that high-stakes testing would generate inflated test scores. Linn, Baker, and Dunbar (1991) argued that:

considering validity in terms of consequences forces our attention on aspects of the assessment process that may not be intended or anticipated by the designers of the instruments. We know from experience that results
Adapting to accountability 4

from standardized tests can be corrupted [i.e., inflated], and we have clear examples of some of the factors that lead to that corruption (p. 17).

Kane (2006) wrote:

In a low-stakes context, performance in various subsets of the target domain may represent performance in the target domain fairly well. In a high-stakes context, standardization to a subset of the target domain may lead to instruction and test-preparation activities aimed specifically at the test, thus making test performance less representative of performance in the target domain as a whole (p. 38).

Koretz & Hamilton (2006) warned that

An overarching psychometric issue raised by …[recent] trends in test use…is the validity of scores obtained under high-stakes conditions. Test scores can become inflated…when teachers or examinees respond to testing in certain ways, and these responses are likely to be particularly severe when the consequences for scores are substantial (p. 542).

The growth of TBA has also spurred empirical research on the effects of high-stakes testing. This work includes studies of educators’ administrative and instructional responses to testing and evaluations of the validity of score-based inferences under high-stakes conditions. This research has confirmed Lindquist’s prediction that some individuals respond to high-stakes testing in ways that inflate scores, and it has shown that in some instances, the resulting bias has been very large.

However, neither the heightened theoretical focus on the effects of testing nor the growing empirical evidence of undesirable responses to high-stakes testing has brought
about substantial changes in the actual practices of measurement. The premise of this paper is that current uses of large-scale testing make such changes in practice imperative.

**Research on Behavioral Responses to Testing**

There is no doubt that TBA has a powerful impact on educators’ practices. Hamilton, Stecher, & Yuan (2008) noted:

Studies of relationships between high-stakes testing and classroom practices have produced one consistent finding: High-stakes testing systems influence what teachers and administrators do. Some of the changes would generally be considered beneficial (e.g., providing additional instruction to low-performing students; taking steps to align the school curriculum across grades), whereas others raise concerns about possible negative effects on the breadth and quality of instruction (e.g., shifting resources from untested subjects or topics to tested subjects or topics; focusing on specific items styles or formats) (p. 3).

Although the effects of TBA documented by research are diverse, the ones of particular importance here are those that threaten the validity of inferences about individuals. These are primarily although not entirely instructional responses, including explicit test preparation. Koretz and colleagues (Koretz & Hamilton, 2006; Koretz, McCaffrey, & Hamilton, 2001) suggested a taxonomy of seven types of test preparation that is useful for the discussion of test design in the following section. Advocates of test-based accountability hope for three types of responses to TBA: that teachers will allocate more time to instruction, work harder, and find more effective ways to teach. All of these are likely (at least to a point) to produce meaningful gains in achievement and are not
pertinent here. At the other extreme, educators may simply cheat, a problem that has recently grown in salience if not in actual prevalence (e.g., Severson, 2011). Cheating obviously undermines validity, but one could argue cheating can arise almost regardless of the standard practices in measurement.

More relevant to the present discussion are the final two categories of responses, which Koretz et al. (2001, 2006) labeled reallocation and coaching. Reallocation refers to shifting resources to better align instruction with the substantive content of the test used for accountability. Reallocation between subjects can arise, for example, when educators increase time spent on tested subjects at the expense of untested ones. However, this is not relevant to the present discussion. Regardless of whether it is desirable as a matter of policy, reallocation between subjects does not bias inferences about achievement from scores in tested subjects. In contrast, reallocation within subjects is critically important. It arises because tests typically sample sparsely from the target of inference. Within-subject reallocation can be desirable, e.g., if teachers to shift resources to portions of the target that are more important for the intended inference. However, reallocation can inflate scores if teachers shift resources to material that is emphasized by the test (e.g., so-called “power standards”) at the expense of material that is untested or de-emphasized in the test but is nonetheless important for the intended inference.

Koretz et al. (2001, 2006) used coaching to refer to test preparation that focuses on narrow, specific, often incidental attributes of the test. These specifics may include format and other aspects of item style, details of scoring rubrics, and details of content that are not substantively important. What these attributes share is that their selection in test design is not governed by the target of inference. Test-taking tricks, such as process
of elimination for multiple-choice items, are also considered coaching. In rare cases—for example, if a test confronts students with a format that is so unusual that it could cause a downward bias in performance—coaching may be desirable, but in general, it is likely to inflate scores or simply waste instructional resources.

Reallocation and coaching inflate scores by different mechanisms. This difference has substantial consequences for test design, and I will return to it below. To clarify this difference, it is helpful to conceptualize a test as comprising a composite of performance elements (Koretz, McCaffrey, & Hamilton, 2001; see also Koretz & Hamilton, 2006). This subsumes all aspects of the test that influence both performance on a test and inferences about it. Many of these elements are substantive, i.e., explicitly or implicitly linked to the target of inference. Others are nonsubstantive, i.e., they are not determined by the inference but may influence performance. Item format is often a nonsubstantive element. An element’s test weight is the sensitivity of the test score to performance on that element. The target of inference is also a composite of performance elements, and inference weights indicate the importance of these elements to the inference. Test and inference weights may not coincide. Because most large-scale tests are small, many elements with substantial inference weights will be given zero test weights in any given form. These can be called implicit elements. Conversely, nonsubstantive performance elements may have an appreciable impact of scores despite being of no importance to the inference. These might be labeled superfluous elements.

In these terms, reallocation entails shifting resources from substantive elements given little or no weight on the test to those with higher test weights. In contrast, coaching is primarily a focus on nonsubstantive elements.
Inflation from reallocation does not require any bias in the estimated performance on individual elements. It generates bias in the creation of a composite—the test score—because performance on the emphasized elements overstates performance on those unmeasured or de-emphasized and therefore exaggerates mastery of the target as a whole. Many examples of the corruption of performance indicators in other fields conform to this pattern of bias in creating a composite measure (e.g., Rothstein, 2008), and the extensive discussion of the principal-agent problem in economics has a similar focus (e.g., Holmstrom & Milgrom, 1991).

In contrast, coaching can undermine the validity of inferences about performance on individual elements. For example, suppose that a test includes items that use Pythagorean triples (integer ratios that conform to the theorem, such as 3:4:5). Suppose that test preparation includes a suggestion that students “solve” these problems by memorizing the most commonly used triples (as in the case of one test-preparation book aimed at the Massachusetts assessments; Rubinstein, 2000). Knowing how to use the triples is clearly insufficient to indicate mastery of the Pythagorean theorem, so performance on that element would be exaggerated.

Research provides ample evidence of reallocation in response to TBA, including forms of reallocation that are likely to contribute to score inflation, such as de-emphasizing or omitting from instruction content that is not tested (e.g., Hamilton, Stecher, Marsh, McCombs, Robyn, Russell, Naftel, & Barney, 2007; Koretz, Mitchell, Barron, & Keith, 1996; Pedulla, Abrams, Madaus, Russell, Ramos, & Miao, 2003; Romberg, Zarinia, & Williams, 1989; Shepard & Dougherty, 1991). Research bearing on coaching is more rudimentary. Few of the relevant studies focused on the nonsubstantive
details of specific tests that would provide the basis for many coaching strategies, and coaching is often subsumed under a broader category of test preparation. Nonetheless, the literature provides clear evidence of coaching. For example, numerous studies have found that teachers adapt instruction to mirror the item formats, scoring rubrics, or other nonsubstantive aspects of the tests used for accountability (e.g., Hamilton et al. 2007; Pedulla et al., 2003; Shepard & Dougherty, 1991; Stecher & Mitchell, 1995).

Although studies of score inflation are not numerous—relevant data are limited, and policymakers usually have no incentive to make them feasible—it is nonetheless clear that inflation is common and is often very large. Most of the studies gauge inflation by evaluating the disparity in score trends between a high-stakes test and a lower stakes audit test, often the National Assessment of Educational Progress (NAEP). The logic of these studies is that because validity requires generalization from the tested sample to the domain, it also requires reasonable generalization in score gains from one tested sample to another, as long as the tested samples are designed to support similar inferences. A number of other factors could cause disparities in trends in scores—for example, differences in the alignment of the two tests to the intended curriculum, independent of educators’ and students’ responses to testing (e.g., Koretz & Barron, 1998). While most studies are insufficient to differentiate among these sources of divergence, most use audit tests intended to support quite similar inferences, so a sufficiently large divergence indicates that score inflation is likely.

The first empirical study of inflation, a cluster-randomized experiment conducted in a context that by current standards was quite low-stakes, found that third grade mathematics scores were inflated by half an academic year four years after a change in
testing programs (Koretz, Linn, Dunbar, & Shepard, 1991). Other studies of inflation have not been experiments, but most compare performance of either the same students across tests or of randomly equivalent samples. In the early 1990s, Kentucky implemented a high-stakes testing program, using primarily constructed-response and performance assessment formats, that included sanctions and rewards. The state legislature mandated that the state tests reflect the NAEP frameworks, making NAEP an ideal audit test. Two years after the implementation of the program, fourth-grade reading scores had increased by roughly three-fourths of a standard deviation on the state test but not at all on NAEP (Hambleton, Jaeger, Koretz, Linn, Millman, & Phillips, 1995). After three years, gains in fourth-and eighth-grade mathematics scores were roughly four times as large on the Kentucky state test as on NAEP (Koretz & Barron, 1998). A number of other studies have found similar disparities (e.g., Jacob, 2005; Klein, Hamilton, McCaffrey, & Stecher, 2000). Studies examining multiple states have shown that disparities in trends between high-stakes state tests and NAEP are common, although not ubiquitous (e.g., Fuller, Gesicki, Kang, & Wright, 2006; Ho, 2007; Jacob, 2007).

**Implications for Test Design**

To lessen these problems, we must start with the features of educational tests that facilitate inappropriate test preparation and score inflation. The core of the problem is that the sampling from the target used in creating most large-scale assessments is both sparse and predictable over time.

Holcombe, Jennings, & Koretz (2013) examined the mathematics tests administered in several grades in New York and Massachusetts from 2006 to 2008 to evaluate the predictability of sampling. They noted that substantive narrowing can arise
at several stages in the construction of a test: in the specification of the standards, the
selection and weighting of standards to test, and the selection of content from within
tested standards. They also observed that test construction requires sampling of
representations:

We use the term *representation* to refer to both unimportant details of
content and what we term *item style*. Item style is broader than item
format, in the usual sense. For example, it includes the type of visual
representation in the item, if any, the magnitude and complexity of the
numbers used, and so on. Consider a hypothetical standard stating that
students should understand the concept of slope in the context of simple
linear equations. Problems involving slope could be presented verbally,
graphically, or algebraically, or they could require translation among those
representations. If the problems are presented graphically, they could be
presented only with positive slopes in the first quadrant, or with a mix of
positive, negative, and zero slopes in all four quadrants (p. 172).

Holcombe et al. (2013) found that predictable sampling substantial at each of
these stages of narrowing. For example, the proportion of standards tested over three
years varied markedly, from 58% in the New York eighth-grade test to 90% in the
Massachusetts eighth-grade test. These figures understate the importance of systematic
sampling of standards because some standards are given far more emphasis than others.
For example, in the Massachusetts tenth grade test, the most highly weighted standard
consistently contributed 15 percent or more of test points, while the second most highly
weighted standard contributed 7 to 17 percent.
Sampling at finer levels of detail is often less straightforward than sampling at the level of standards. For example, in the New York State testing program, one of the two most frequently tested eighth-grade mathematics standards was the following:

8.N.4 Apply percents to: tax, percent increase/decrease, simple interest, sale price, commission, interest rates, gratuities.

One might argue that the wording of the standard in itself narrows the focus of the test, but even taking the standard as given, the test items administered further narrowed the content. Over the first five years of the testing program, 12 items addressing this standard were administered. All but one of the 12 was of one of two forms:

1. The item provides the base quantity and the percent. The student provides the product.
2. The item provides the base quantity and the percent. The student calculates the product and adds it to or subtracts it from the base to get a new total.

The single exception was an item that required that students calculate a percent change. Given these predictable forms, it would be easy to develop methods to coach students to answer these 11 items mechanically, particularly if one could find key words or phrases (such as “total” and “new”) that allow them to differentiate the two classes of items. 1

In extreme cases, the sampling of representations produces items that are near clones. An example from Holcombe et al. (2013) is shown in Figure 1. Four items testing this standard were administered over the first five years of the testing program. One of

---

1 During the years in question, New York State released all of its test items after use. They can be retrieved from http://www.nysedregents.org/intermediate.html. A document containing the 12 items referenced in this paragraph is available on request from the author.
the additional items was essentially the same as those in Figure 1; the fourth was very similar but used longer distractors. One could train students to answer these items by simply matching “measure” and “mass” in the stem with “scale” in the answer, without teaching them any of the relevant concepts. In most cases, of course, items are not as similar as these, but they often show similarities that afford opportunities for the types of coaching that undermine validity.

However, items do not need to be near clones to afford the opportunity for coaching. For example, one of New York’s seventh grade mathematics standards before the adoption of the Common Core was “Compare numbers written in scientific notation.” During the first four years of the testing program, five items testing this standard were administered. Four of the five were similar to the one presented in Figure 2. Specifically, all presented four numbers with the same base (10) and asked students to select the largest number. Once this presentation was expected, it would be easy to coach students to answer these items correctly without teaching them the content. One would tell them to follow the following steps:

- Pick the numbers with the largest exponent. (The student only needs to know which numbers to compare, not what they are.) If there is only one number with that exponent, you are done.
- If there are two numbers with the largest exponent, pick the one with the largest coefficient. Again, one need not understand what the coefficient is.
In the low-stakes framework that dominated the development of educational testing, systematically incomplete sampling had two primary effects, neither of which posed a fundamental threat to the validity of score-based inferences. The first effect was of course measurement error, which has been a primary focus of traditional psychometrics. The second consequence was differences in results among tests with similar purposes. These differences are occasionally large. For example, during the 1970s, Iowa students showed nearly twice as rapid a decline in performance on the ITBS as eighth-graders than the same cohorts showed on the ITED one grade later (Koretz, 1986, p. 54). Over the past two decades, the upward trend in fourth grade mathematics has been considerably faster on the main NAEP than the long-term NAEP and far faster on NAEP than in the U.S. sample on TIMSS. However, under low-stakes conditions, differences across tests are usually more modest, and they never became a major focus of traditional psychometrics.

In the current era of TBA, however, the consequences of this predictable sampling are far greater and far more threatening to valid inference. Educators and students have strong incentives to focus on the specifics of the tested sample, at the expense of untested or de-emphasized portions of the target of inference. To the extent that they do so, the relationship between scores and mastery of the target will be corrupted. As the research cited above shows, the resulting biases can be very large.
For this corruption to occur, two things beyond predictability of sampling are needed: educators need to be aware of the predictability, and they need effective mechanisms—either reallocation, coaching, or both—to take advantage of them. Certainly, many educators do recognize this predictability to some degree. They learn about it through their own experience with tests, through the now ubiquitous and widely encouraged use of old test items in instruction, and from test preparation materials, including both commercial products and materials frequently provided by state and local education agencies. The extent to which they have found effective tools for using this information is suggested by findings of score inflation.

**Factors Encouraging Predictability**

Why are tests predictable? In addition to technical reasons, I suggest that there are practical ones. The first is cost. It is less expensive and quicker to modify items that have performed well in the past, or use them as templates, than to devise totally new items that avoid the sampling details of the old ones. A second is that unpredictability in high-stakes tests can generate intense protests from educators and parents. (I have been in discussions with state officials about the design of new tests in which this concern was prominent.)

For present purposes, however, the technical considerations encouraging predictability are more relevant, as they are directly in the field’s purview. The technical benefits of predictability are primarily matters of error. For example, maintaining similarity of items will reduce noise in estimates of change over time, and it will reduce the risk of model misfit. Thus, generating similar items has often been an explicit goal of measurement experts. One example is the extensive efforts at the Educational Testing Service to develop item models that can “generat[e] items that are isomorphic, that is,
equivalent in content and equivalent psychometrically” (Bejar, Lawless, Morley, Wagner, Bennett, & Revuelta, 2003, Abstract [no page number]). As Morley, Bridgeman, & Lawless (2004) clarified, these isomorphs share what Holcombe et al. (2013) called representations as well as content, and they found that the similarity in representations influences performance. A second example is the use of task models by the Smarter Balanced Assessment Consortium, one of the two multi-state consortia developing new assessments to align with the Common Core standards:

> Often in statewide assessment, item writers are given the objectives and asked to develop a certain number of items of different types and at different DOK levels to address the objectives. In the case of Smarter Balanced, item/task models have been developed for each target to assure a greater degree of consistency and replicability of the items or tasks addressing a target across developers and across years (Smarter Balanced Assessment Consortium, 2012, p. 11).

While the technical benefits of item similarity have been the focus of extensive work, the technical costs of this similarity have received far less attention. These costs are primarily matters of bias—score inflation—and undesirable effects on instruction rather than error. Thus, to oversimplify, the consequence of ignoring the problem of predictability in the context of TBA is to generate more precise but often badly misleading estimates of achievement.
Possibilities for Changes in Design

The notion of making tests less predictable to lessen score inflation is not entirely new, even though it has gained little traction in the measurement field. For example, Hanushek (2009) suggested:

Having a large test bank would permit providing each student with a random selection of questions, minimizing any chance of cheating. Indeed with a large test bank covering the range of relevant material, it would even be possible to make questions available beforehand, with the notion that “teaching to the test” could actually be considered productive (803).

The challenge, however, is determining how much and in what ways items must vary in order to lessen the problem of score inflation and then to address the psychometric issues that an increase in item heterogeneity will generate. To date, the measurement field has devoted little attention to either of these challenges.

Three steps are needed to begin addressing the first of these challenges. First, one will often need more specific agreement about the target of inference than we generally have, because without that it is not feasible to specify the nature and limits of the tasks students should be given. Second, to lessen inflation from undesirable reallocation, one will need variation over time in the sampling of substantive performance elements, which may be what Hanushek meant by “the range of relevant material.” Third, to lessen bias from coaching, one will need variation in the sampling of nonsubstantive performance elements, including representations, substantively unimportant details of content, and nonsubstantive response demands, e.g., substantively unimportant aspects of scoring rubrics.
Some might argue that agreeing on standards is sufficient specification of at least the substantive content of the target of inference, but for several reasons, it is not. First, important groups of users of scores may have different desired or intended inferences. For example, Porter, McMaken, Hwang, & Yang (2011) found that the alignment between the Common Core standards and the National Assessment of Educational Progress (NAEP) is only modest. If policymakers accept the Common Core as limiting the domain from which test writers can sample, they have in effect declared that content included in the NAEP frameworks but excluded from the Common Core is not part of their targets of inference, but it is not clear that stakeholders used to relying on NAEP trend data will agree. Second, as noted above, the construction of a test requires sampling from within the domain defined by standards, either sampling of standards or sampling of content from within standards, as well as assigning weights to these substantive elements.

A simple thought experiment shows the risk of mistaking standards for the target of inference. Until the acceptance of the Common Core standards, most states had their own content standards, and these often differed markedly from state to state. Consider a metropolitan area that spans a state border, such as New York City or Washington, D.C. Imagine an employer in the Virginia suburbs of Washington who hires entry-level employees from Virginia, Maryland, and the District of Columbia. Would the employer expect different mathematical competencies from employees drawn from the three jurisdictions because of their differing standards? Clearly not; most employers will expect the same competencies of employees regardless of the state in which they were educated. For most employers, the most important question is not a specific set of standards, but rather students’ ability to use the knowledge approximated by the
standards in the variety of forms in which the student will later confront it. The adoption of shared standards, such as the Common Core, will partially obscure this limitation of standards, but it will do nothing to eliminate it. Similarly, the current focus on ‘college and career readiness’ is in substantial part rhetorical and is not a substitute for systematic evidence about the targets of other stakeholders.

Some of the choices about inclusion and weighting reflect explicit decisions of policymakers (again, not necessarily decisions of other stakeholders), but others appear to be incidental. For example, the Partnership for Assessment of Readiness for College and Careers (PARCC), one of the two major consortia developing assessments aligned with the Common Core standards, has divided the standards into three categories reflecting PARCC’s view of their importance: Major Clusters, which should constitute the majority of the test, Supporting Clusters that will receive less weight, and Additional Clusters that will receive little weight (Partnership for Assessment of Readiness for College and Careers, 2012). In contrast, it appears that decisions about test weights are often made without a decision about relative importance by policymakers. For example, the core curriculum implemented in 2005 by New York State—since replaced by the Common Core standards—included the following two eighth-grade algebra standards:

8.A.12 Apply algebra to determine the measure of angles formed by or contained in parallel lines cut by a transversal and by intersecting lines.

8.A.15 Understand that numerical information can be represented in multiple ways: arithmetically, algebraically, and graphically (New York State Education Department, 2005).

The curriculum document did not provide information about the relative importance of these standards, but the first was tested six times in the first four years of the testing program, while the
second was not tested at all. Similarly, over a five year period, approximately 12 percent of the possible raw score points in New York’s eighth-grade English Language Arts tests were assigned to items representing this standard:

Interpret characters, plot, setting, theme, and dialogue using evidence from the text.

In contrast, the items representing the following standard were assigned about two percent of possible points:

Evaluate the validity and accuracy of information, ideas, themes, opinion, and experiences in texts to identify multiple levels of meaning.

This is not unique to New York’s assessments.

To lessen bias from coaching will require close attention to varying many details of the test other than substantive content. These may include, for example, substantively unimportant details of content, representations, other aspects of item style, unnecessary regularities in task demands (e.g., some aspects of scoring rubrics), and superfluous performance elements. Test preparation focused on any of these aspects of items can improve scores without increasing the knowledge or skill the item is intended to represent.

These three aspects of item sampling—definition of the target, substantive sampling, and nonsubstantive sampling—are often difficult to disentangle, and identifying problems with them will often require detailed examination of tests as well as standards. For example, standards often specify that students should be able to calculate the volumes of simple polyhedra and apply this skill to real-world problems, but it would be a mistake to infer from this that various tests assess similar skills. For example, until 2011, the Massachusetts standards for 10th grade included this:
10.M.2. *Given the formula* [emphasis added], find the lateral area, surface area, and volume of prisms, pyramids, spheres, cylinders, and cones… (Massachusetts Department of Education, 2000, p. 75).

As students are given the formula, items testing this standard entail nothing more than substitution in the formula and simple arithmetic. In contrast, items in the Singapore Primary School Leaving Examination, taken at the end of the sixth grade, test whether students can solve such problems without being given the formula (Singapore Examinations and Assessment Board, 2009). Moreover, the items assessing this content differ in other important ways as well. Example items are shown in Figure 3. The Massachusetts items focusing on rectangular prisms generally refer to an object that is completely full, which eliminates the need to use any mathematical knowledge or skills other than substitution and arithmetic. In contrast, the Singapore item is far more complex, requiring the student to apply the formula (not given) to two partially full objects of different dimensions. Here again, one must ask: what is the likely target of inference for various stakeholders? For users who expect the skills assessed in the Singapore item, the Massachusetts standard and items represent an unwarranted narrowing with respect to the target.

Thus, even identifying unwanted predictability and narrowing as a precursor to designing a more varied set of tasks is a major undertaking. One possible approach would be to pair routine examination of regularities in test items (similar to that done by
Holcombe, et al., 2013) with less frequent monitoring of test preparation materials and activities to identify the regularities that contribute most to inappropriate preparation for a specific test.

**Options for test design.** Only a few options have been suggested for incorporating more diverse items into tests, and there is as yet very little work exploring their advantages and disadvantages. This is an area in which there is a pressing need for research.

Neal (2013) suggested administering two different tests, one for accountability and a second for other purposes, such as describing the progress of individual students. He argued that the item similarities required for linking and scaling encourage educators to respond in ways that inflate scores. He also suggested that one needs only ordinal data and reliable ranking, and neither information about change nor putatively interval scales, to support an effective performance incentive system. Therefore, he suggested that the test used to provide incentives to educators should incorporate items that vary in content and format and should not be scaled or linked. A second, low-stakes test would be used to provide information about students. This second test would be subject to the usual psychometric constraints governing test construction, would have less item variation, and would be scaled and linked.

This is an intriguing idea, but it raises numerous issues that have yet to be addressed. For example, Neal has not yet clarified how the frameworks or targets for the two tests would be similar and different. Presumably, they would have to be reasonably similar, as few stakeholders would be satisfied with a system in which students are evaluated with respect to substantially different material than teachers are expected to
teach, but the more similar the two tests are, the less effective separating the two would be. Testing time would presumably roughly double, which would be difficult for many educators and parents to accept given the already burdensome amount of testing. The test designed for students, while intended to be low-stakes, might not in fact remain as such, as there are mechanisms other than explicit sanctions and rewards that can make a test somewhat high stakes. (The first empirical study that found identified, score inflation [Koretz et al., 1991] was conducted in an environment in which there were no explicit sanctions and rewards.) Neal’s approach also requires consistent ranking of students, and it is not clear how stable rankings would be across the as yet unspecified variation in the tests he proposes for accountability. Administering a test that is consequential for teachers but not for students also runs the risk of motivational biases affecting the test used for accountability.

It might be possible to implement a variation of Neal’s approach with a single test by modifying a frequently used design in which common items are used to score students, while both common and matrix-sampled items are used to score schools. In theory, one could exclude the common items from scores for schools and teachers and free the matrix items from some of the design constraints imposed on the common items. This might require longer testing time than current tests, but it would avoid some of the problems inherent in administering tests that are consequential for teachers but not students.

Koretz & Beguin (2010) suggested a third approach, self-monitoring assessments (SMAs). They suggested maintaining current design principles for much of an operational test while adding audit components. These audit components would comprise
Adapting to accountability

items designed to maintain coverage of important performance elements while undoing predictable patterns that encourage inappropriate test preparation and score inflation. Inflation would still occur on the conventional items, and discrepancies in performance between the operational and audit components would signal likely inflation. While the primary purpose of an SMA is to detect score inflation, the inclusion of audit items might reduce the incentives for inappropriate test preparation and hence might lessen score inflation. Koretz & Beguin (2010) suggested five specific SMA designs that differ in terms of when items are introduced and how linking is done.

Although some initial pilot testing suggests that an SMA design can identify some score inflation (e.g., Ng, Koretz, & Jennings, 2013), this approach, like Neal’s, is largely untried and poses many unresolved issues. It remains unclear, for example, in what ways and to what degree audit items should differ from others to function well as audits while maintaining fidelity to the target of inference. Providing a sufficiently large number of audit items is likely to require longer tests, although in this case as well, computer adaptive testing might compensate. Some SMA designs also confront a serious analytical limitation. Ideally, one would initially place both audit items and operational items on a single scale. This would provide estimates of inflation that are absolute relative to the scale. However, this requires that audit items be designed at the outset of a testing program, and that is not necessarily entirely feasible. Two or even three iterations of a test may be needed before some of the predictable patterns become apparent, and if so, it would be necessary to add additional audit items at that time in order to capture the inflation arising from these patterns. However, it is not likely to be practical to place these additional audit items on the operational scale because of likely corruption of the
extant linking items. In that case, one would be limited to a substantially weaker, difference-in-differences analytical approach that can only detect variations in inflation (Koretz & Begin, 2010).

One of the critical issues that would be raised by an increase in the diversity of items is measurement error. Increasing the heterogeneity of items over time without changing other aspects of test design would likely create greater instability in the relative performance of individuals and groups. This instability could take several forms, but to my knowledge, there is as yet no research directly addressing this question. For example, it is possible that internal consistency reliability (e.g., Cronbach’s alpha, or the person-by-item interaction in a cross-sectional generalizability analysis) might decrease. In the absence of data, it seems likely that a more substantial effect might be an increase in the person-by-occasion (or school-by-occasion) interaction across years, as students or schools affected by larger amounts of inappropriate test preparation would fall in relative position compared with students or schools affected by less. Any substantial increase in instability may cause difficulties for educators, policymakers, parents, and students. However, this may be an unavoidable cost of reducing the sometimes large bias from score inflation. Moreover, it is important to recognize that the decrease in measurement error obtained by not varying items, as in current designs, is artifactual, arising from the sampling of knowledge and skills that has been made inadequate by the pressures of accountability.

An obvious response to any increase in measurement error would be to lengthen the test, but in the current environment, increases in testing time may be unacceptable. By reducing time spent on ineffective items, computer adaptive testing might permit greater
stability in estimates for students and schools without an increase in testing time. Matrix sampling could lessen the instability of estimates for schools. However, responses to any increases in measurement error will have to wait for research analyzing which facets of measurement error most are most affected by the needed changes in test design.

**Implications for Linking**

Most large-scale testing programs now rely on IRT non-equivalent groups anchor test (NEAT) designs to link scales over time. A modest number of secure anchor items are included in at least two successive forms, and performance on those items is used to link the scale across forms. The nonequivalent groups are the students in the same grade in successive years. NEAT linking can be done in numerous ways. For example, one can calibrate the two forms separately in the two samples of students and then adjust the mean and standard deviation of the item difficulty parameters in the new sample to match the old. When the IRT NEAT approach is justified, it has numerous important advantages. For example, it avoids the need for an additional test administration and the motivational effects that might bias linking done in the context of a field test.

However, a key assumption required for NEAT linking to provide valid inferences is not tenable under the high-stakes conditions created by TBA. In NEAT linking, the anchor items provide the constraint that allows one to make the scale comparable from one year to the next, and they provide the only basis for judging the change in achievement in the tested population as a whole. An essential assumption underlying this approach is that the relationship between the latent trait and observed performance on the linking items is constant, apart from error and scaling artifacts that can be removed by the transformation of scale effected by the linking procedure.
Adapting to accountability

This assumption is not warranted under high-stakes conditions. If test preparation has improved performance on the linking items more than improvement on the latent trait warrants—that is, if test preparation has made the items easier than changes in mastery alone would do—NEAT linking will build score inflation into the scale (Koretz & Barron, 1998; Koretz, 2007). Keeping linking items secure is not sufficient protection against this bias. Linking items may be sufficiently memorable that teachers can focus instruction on them directly, anticipating the appearance of similar items in the future. Moreover, it is not necessary that specific linking items be remembered by teachers. It is only necessary that linking items be sufficiently similar to other items that inflation generalizes substantially to the linking items. Given that linking items are typically chosen to be representative of the test as a whole (Kolen & Brennan, 2004), this degree of similarity is likely.

Standard procedures for evaluating linking items will not reveal this problem. Ideally, the relationship between the estimated item difficulty parameters in the two years should be approximately linear, and the deviation of the line fitting this plot from the identity line is assumed to reflect the artifactual differences in scale that the linking process is designed to eliminate. Items that lie far from the equating line are assumed to be functioning differently in the two years and are typically dropped from the linking (e.g. Kolen & Brennan, 204, p. 187 ff.). However, inflationary bias in an appreciable number of linking items will displace the line, resulting in an incorrect transformation of scale. Unless only a few items are affected, conventional examination of plots of item difficulties will not reveal this bias.
When tests are linked with IRT NEAT linking, this bias in the difficulty of linking items is the only way in which score inflation can occur in the tested population as a whole, because the estimate of change for the population as a whole reflects only these items. Therefore, studies of score inflation in assessment programs that used NEAT linking (e.g., Klein et al., 2002; Koretz & Barron, 1998) illustrate that this bias does arise and that it can be very large.

**Possibilities for Changes in Linking**

The impact of TBA on NEAT linking may be the most difficult of the challenges described in this paper. It does not seem that any modest departures from current practices in NEAT linking can eliminate this bias. I agree with Brennan (2007), who suggested that the data currently used in NEAT linking are not sufficient to resolve this problem. I am aware of no efforts to design and evaluate reasonable alternatives.

One option might be to complement NEAT linking with common-persons linking in a population that was not previously administered the particular high-stakes test in question. However, may be impractical under current conditions. First, the sheer volume of current testing makes it unlikely that a reasonable number of schools would agree to participate in additional testing to benefit the testing program in another state. Second, this would require linking in the context of a zero-stakes or very low-stakes field test. This would pose a serious risk of motivational biases, particularly for difficult items, and it seems likely that this risk would be exacerbated by students’ experiences with high-stakes testing. Finally, if, as currently planned, many states adopt common tests, the opportunities for this type of linking will be reduced.
A second alternative might use linking items that are less vulnerable to bias from score inflation, but this would pose serious technical challenges. The first challenge is one of design. To function in this manner, linking items would need to differ sufficiently from the majority of operational items to be largely unaffected by inappropriate test preparation focused on most operational items. At the same time, these items could not be too novel, for two reasons. First, linking items should be representative of the test as a whole (e.g., Kolen and Brennan, 2004). Second, if items are too novel, they may be too salient and therefore too memorable to serve as uncorrupted linking items. There is as yet no research indicating whether this approach is feasible.

A second barrier to using more novel linking items is analytical. Linking functions can vary substantially with the selection of anchor items, even in a traditional design in which linking items are not novel (Fitzpatrick, 2008; Michaelides & Haertel, 2004). Michaelides & Haertel (2004) estimated that linking error from item sampling and person sampling contribute roughly comparable amounts of error to individual scores but that item sampling is likely to be the primary source of linking error for aggregate scores. It seems likely that using more novel items for linking will increase error from item sampling. It is also likely that increasing item novelty will increase the number of items that fail to function well as linking items. Finally, in a testing system in which score inflation is severe, the use of relatively bias-resistant linking items could result in a rapidly growing disparity in difficulty between linking items and conventional operational items. This could draw unwanted attention to the linking items and might also generate complaints about seemingly inappropriate test construction.
Implications for Validation

Clearly, traditional approaches cannot be relied on to provide a reasonable evaluation of the validity of inferences under high-stakes conditions (VIHS). The tests that produced the severely biased scores in studies of score inflation had been validated by traditional means, but the many of the key inferences based on those scores were nonetheless not even approximately correct.

There are several reasons why traditional approaches to validation fail to provide reasonable information about VIHS. First, the evaluation of content generally considers only a single form, not the adequacy of sampling over repeated forms, and relatively little attention is given to omissions from the tested sample or undesirable patterns in test weights. The empirical evidence adduced in traditional validation is typically collected before high stakes have influenced educator and student behavior—that is, before there has been an opportunity for score inflation to occur. This evidence is also primarily cross-sectional and correlational, so it is insensitive to changes in level even if collected repeatedly. (For a concrete example of correlations between high-stakes and audit scores that were stable over time in the presence of severe inflation, see Koretz & Barron, 1998).

Despite the unarguable failure of traditional validation to evaluate VIHS adequately, changes in practice have been meager. With very few exceptions, studies of score inflation and relevant behavioral responses have been conducted independently of the validation programs of the host jurisdictions and their testing contractors. Very few states explicitly consider the risk of score inflation in their validation programs, and many of the discussions of validity used in training educators or psychometricians make little or no mention of VIHS.
In contrast to test design and linking, the changes in validation warranted by the use of tests for accountability seem relatively clear and feasible, albeit expensive and burdensome. The first imperative is to continue validation past the first administration of tests used for accountability in order to capture the effects of undesirable responses to testing. Second, it is necessary to extend the evaluation of content-based evidence. We must evaluate multiple forms over time, considering predictable patterns in test weights, including omissions, and we should evaluate nonsubsantive as well as substantive performance elements.

Third, we need routine evaluation of possible score inflation, which requires a greater emphasis on extrapolation. In increasingly common parlance in the measurement field, generalizability is used to refer to the consistency of performance across exchangeable instances of measurement, while extrapolation refers to consistency between the universe score (across those exchangeable instances of measurement) and the target of inference (Kane, 2006). Evidence bearing on extrapolation—specifically, the degree of generalizability between a high-stakes test and an audit test—is the primary source of data about VIHS. Nonetheless, as Haertel (2013) noted, current validation studies do not give much attention to extrapolation. Of course, one impediment to assessing extrapolation in evaluating VIHS is that one needs appropriate audit measures.

Finally, we need more frequent evaluation of the effects of testing. However, I mean this in a restricted sense. Haertel (2013) recently argued that we should evaluate a broad range of the effects of testing, including indirect effects, and he noted that the use of tests for accountability increases the importance of some of these effects. I entirely agree. Research has shown that TBA has powerful effects, and there is an ethical
obligation to evaluate those effects, particularly as they affect children. However, in this paper, I am addressing only the impact of accountability on the validity of inferences, so here I am referring only to the effects of testing that have a potentially serious impact on VIHS. We need evaluation of behavioral responses to testing, including explicit test preparation and other aspects of instruction. This is needed not only for evaluation, but to inform the test-design decisions sketched above.

**Discussion**

I agree with Neal (2013), who argued that “Education researchers need to directly address the task of designing assessment systems that make it difficult, if not impossible, for educators to form profitable coaching strategies for exams” (p. 17). The purpose of this paper is to explore several of the aspects of the testing enterprise—test design, linking, and validation—that would need to be changed to create assessment systems that are less vulnerable to score inflation from inappropriate test preparation.

Some of the impediments to the creation of testing systems that are more appropriate for accountability are beyond the control of education researchers and measurement experts. First, the stakeholders who commission the creation of testing systems have little or no incentive to demand more appropriate testing programs. In our current accountability programs, all of the participants in the system, from classroom teachers to state superintendents, have the incentive to raise scores as much and as quickly as possible. The system provides no incentives to monitor how gains are obtained or to evaluate or lessen score inflation. Second, some of the needed improvements are expensive, and the state departments of education that would need to fund the improvements are often badly short of funds. These two factors will impede the research
needed to develop and evaluate approaches more appropriate for test-based accountability, as well as the implementation of designs that prove feasible. Third, independent studies augmenting conventional validation are hindered by fact that in education, in contrast to some other areas of public policy, there is no expectation that data should be accessible for research purposes. Researchers proposing potentially threatening studies, such as evaluations of possible score inflation, are sometimes denied access to data.

Despite these obstacles, however, there are important steps that the field can and should take. It is entirely within the purview of the field to set professional standards and expectations for validation. We can modify our discussions of validation—our theoretical discussions and even our proposals for validation of testing programs—to reflect the importance of accountability-oriented uses of tests. We can undertake some of the necessary research and development without waiting for states and others sponsors of tests to agree to large-scale changes in design, linking, or validation. Finally, given the current uses of tests, those in the field with responsibility for dissemination, such as journal editors, should give appropriate weight to the importance of accountability-related issues.

The changes that are within the control of the measurement field could have a substantial if long-term impact on practice. For political and legal reasons as well as a desire for best practice, the sponsors of tests often depend on the claim that their testing programs are operated consistently with professional standards. If the field makes it clear that the design, operation, and validation of tests used for accountability should be
Adapting to accountability

tailored to that use, then sponsors of testing programs may begin accepting the needed changes.

While some of the general directions necessitated by accountability are clear, many of the specifics are not. Some of the options noted above were suggested primarily to spark discussion, and I noted several instances in which we still lack sufficient information about the advantages and disadvantages of them. This paper is not an argument for a specific approach, but rather an argument that we must recognize the need for fundamental changes and a call for the research and development that will make these changes feasible.
References


Adapting to accountability


[http://dash.harvard.edu/handle/1/10880553](http://dash.harvard.edu/handle/1/10880553).


Figures

a) Item 27 in 2008

Which tool is most appropriate for measuring the mass of a serving of cheese?
A. ruler  
B. thermometer  
C. measuring cup  
D. weighing scale

Item 9 in 2009

Which tool would be most appropriate for Natasha to use when finding the mass of a watermelon?
A. scale  
B. inch ruler  
C. meter stick  
D. measuring cup

Figure 1. Near-clone items from a seventh-grade mathematics test. From Holcombe et al. (2013), Figure 7.6, with permission of the authors.
The table below shows the number of computers a company sold in four different years.

<table>
<thead>
<tr>
<th>Year</th>
<th>Computers Sold</th>
</tr>
</thead>
<tbody>
<tr>
<td>2002</td>
<td>$3.2 \times 10^5$</td>
</tr>
<tr>
<td>2003</td>
<td>$8.4 \times 10^3$</td>
</tr>
<tr>
<td>2004</td>
<td>$5.9 \times 10^5$</td>
</tr>
<tr>
<td>2005</td>
<td>$1.2 \times 10^4$</td>
</tr>
</tbody>
</table>

In what year did the company sell the most computers?

A 2002  
B 2003  
C 2004  
D 2005

*Figure 2. A seventh-grade item assessing comparison of numbers in scientific notation.*
a. 10\textsuperscript{th}-grade MCAS (March retest), 2010: formula provided

A right rectangular prism has the following dimensions:

- The height is 5 feet.
- The length is 6 feet.

The volume of the prism is 60 cubic feet.
What is the width of the prism?

A. 2 feet  
B. 3 feet  
C. 5 feet  
D. 6 feet

b. 6\textsuperscript{th}-grade Singapore Primary School Leaving Exam: no formula given

In Figure 1, Tank A is completely filled with water and Tank B is empty. Water is poured from Tank A into Tank B without spilling. The heights of the water level in the two tanks are now equal as shown in Figure 2.

Figure 1

![Figure 1: Tank A and Tank B with water levels]

Figure 2

What is the height of the water level in Tank A in Figure 2?

Answer ________________

Figure 3. Items addressing the volume of a rectangular prism, 10\textsuperscript{th} grade Massachusetts MCAS and Singapore Primary School Leaving Exam (6\textsuperscript{th} grade). MCAS item retrieved from MA Department of Education’s MCAS Question Search (2011).

http://www.doe.mass.edu/mcas/search/ Singapore item from Singapore Examinations and Assessment Board (2009).