Adapting Educational Measurement to the Demands of Test-Based Accountability

Adapting Educational Measurement to the Demands of Test-Based Accountability

Daniel Koretz
Harvard Graduate School of Education.

January, 2015

Author's final

Published version:

Koretz, D. (2015). Adapting the practice of measurement to the demands of test-based accountability. *Measurement: Interdisciplinary Research and Perspectives, 13*, 1-25. http://dx.doi.org/10.1080/15366367.2015.1000712.

The research reported here was supported by the Institute of Education Sciences, U.S. Department of Education through Grant R305AII0420 to the President and Fellows of Harvard College. The opinions expressed are solely those of the author and do not represent views of the Institute, the U.S. Department of Education, or their staffs.

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, represents a tremendous change from the typical uses of tests 40 years ago.

The premise of this paper is that the current uses of tests for accountability require major changes to several aspects 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 instructional responses to test-based accountability (TBA), including simple test preparation activities, and their implications for the validity of score-based inferences about student achievement. 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

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

Educational Measurement in the Decades After World War II

In this paper, I focus only on changes in large-scale testing that have occurred since the middle of the last century. For a more comprehensive overview and for more detail about the points summarized here, see Koretz & Hamilton (2006).

Much of the large-scale testing in the first decades after World War II was relatively low-stakes—that is, scores had limited direct consequences for students and schools, and consequently, there was relatively modest pressure to raise scores as an end in itself. For example, large-scale commercially produced norm-referenced tests, such as the Iowa Tests of Basic Skills, were marketed as a tool for diagnosing students' strengths and weaknesses, and the materials sold with the test still stress that function (e.g., Hoover et al., 2003). The shift toward higher-stakes tests began gradually but took a critically important turn with the rise of "measurement-driven instruction" and minimum-competency testing in the 1970s. Successive waves of reform broadened and heightened the pressure to raise scores. For example, in the 1990s, many states, such as Kentucky, imposed sanctions and financial rewards for schools based on cohort-to-cohort changes in test scores. The implementation of the No Child Left Behind Act in 2002 incorporated a modification of this system into federal policy, making it mandatory for states accepting the largest portion of aid under the act. More recently, many states have shifted the focus to sanctions for teachers, implementing score-based evaluations using value-added models, student growth percentiles, or other metrics.

For present purposes, the specific sanctions or rewards attached to scores—or even the presence of tangible sanctions and rewards—are unimportant. Rather, the issue is simply educators (and sometimes students) perceiving substantial pressure to raise scores. Indeed, the first empirical study identifying score inflation (Koretz, Linn, Dunbar, & Shepard, 1991) was conducted in a district in which teachers perceived strong pressure to raise scores but did not face tangible sanctions of the sort that are common today.

The development of educational testing during the period of relatively low stakes reflected that environment in a number of ways. First, the educational measurement field gave relatively little attention to the consequences of testing. Second, the field 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, in particular, to instructional responses such as inappropriate narrowing of instruction or coaching students to capitalize on incidental attributes of items.

Warnings that educators' behavioral responses to testing might undermine the validity of inferences about student achievement 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 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).

By "criterion series," Lindquist meant what we now call the target of inference. 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 increases in scores required to avoid sanctions are often large, 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 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 in inferences about achievement gains has been very large. Several illustrative studies are noted below.

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

TBA has a powerful impact on educators' practices. For example, 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).

The focus of this paper is the validity of inferences about students' achievement, and while the documented effects of TBA are diverse, I focus here only on those that bias these inferences. Test scores are currently used in many ways, and there may be some inferences that are not biased even when the scores of individual students are inflated. For example, if inflation were uniform across teachers, it would not bias inferences from

normative value-added evaluations of teachers. (Our ongoing research indicates that inflation is not uniform and that it varies systematically with characteristics of students and schools [e.g., Koretz, Yu, & Braslow, 2013], but that empirical question is beyond the point of this paper.) Even if other unbiased inferences exist, they would not moot the argument presented here. Few would argue that severe bias in the scores of individual students on large-scale assessments would be acceptable if other score-based inferences were not comparably biased.

The behavioral responses that create bias in individuals' scores are primarily instructional responses, including explicit test preparation. My colleagues and I have suggested a taxonomy of seven types of test preparation that is useful for the discussion of test design in the following section (Koretz & Hamilton, 2006; Koretz, McCaffrey, & Hamilton, 2001). 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 necessarily 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, so it is not germane here.

More relevant to the present discussion are the final two categories of responses, which we labeled *reallocation* and *coaching* (Koretz et al., 2001, 2006). 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; whether or not it is seen as desirable, 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.

Note that the key issue is representation of the target of inference, not representation of a given set of standards. As discussed below, the target is often broader than a particular set of standards.

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 to which I will return below. To clarify this difference, it his helpful to conceptualize a test as comprising a composite of *performance elements* (Koretz, McCaffrey, & Hamilton, 2001; see also Koretz & Hamilton, 2006). This term 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. Conversely, nonsubstantive performance elements may have an appreciable impact on scores despite being of no importance to the inference. While these elements are superfluous for representation of the target, they are often necessary for construction of a test. What is often not necessary is that they be similar over time.

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 entail bias in creating a composite measure (e.g., Rothstein, 2008). The extensive discussion in economics of the principal-agent problem in employment has a similar focus (e.g., Holmstrom & Milgrom, 1991). In this literature, the principal is the owner of the firm, and the agents are employees. Ideally, incentives for agents should induce behaviors that increase value for the firm, but because most employees have multidimensional roles that will be measured only incompletely, incentive systems may induce distortions that run counter to the interests of the principal.

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 in many studies, coaching is 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 such studies 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 or randomly equivalent samples, so they circumvent the primary weakness of evaluations using observational data. For example, In the early 1990s, Kentucky implemented a high-stakes testing program that included sanctions and rewards, using primarily constructed-response and performance assessment formats. 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 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 over a period of several years contributed 7 to 17 percent in specific years.

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.[1] Given these predictable forms, it would be easy to develop methods to coach student 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.

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 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 concepts relevant to the intended inference.

Figure 1 about here

In most cases, of course, items are not as similar as these across years, but they often show sufficient similarities to afford opportunities for the types of coaching that undermine validity. 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

---

[1] During the years in question, New York State released all of its test items after a single use. They can be retrieved from http://www.nysedregents.org/intermediate.html.

students to select the largest number. Once this presentation was expected, it would be easy to coach students to answer these items correctly without adequately teaching them the skill the items are intended to measure. 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 that they are exponents or what exponents 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.

Figure 2 about here

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, which arises simply from the fact of sampling rather than its systematic nature, is measurement error, which has been a primary focus of psychometrics for generations. The second consequence, which does reflect the systematic nature of the sampling, is differences in results among tests with similar purposes. Infrequently, these differences are 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 of TIMSS. However, under low-stakes conditions, differences across tests are usually more modest, and they never became a major focus of 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, as evidenced by the numerous studies of reallocation noted earlier. 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. Findings of score inflation indicate that educators have found effective tools for using this information. In addition, recent research has found that performance gains are larger on items assessing items that had high weights in previous years (Jennings & Bearak, 2014).

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. 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 [depth of knowledge] 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, emphasis added).

The effects of task models on performance have not been well explored empirically. One study (Morley et al., 2004) found that performance on "close variants" was generally higher than on items with similar content but less similar representations, which is consistent with the argument by Holcombe et al. (2013) that similarities in representation and details of content may result in score inflation. It is important to note that neither the importance nor the complexity of the elements shared among items because of a task model is relevant to the argument about inflation posed here. What matters is solely predictability that allows students to improve performance on administered items more than their performance would improve on items selected randomly from the universe of acceptable content and representations.

A major change currently underway in large-scale testing that is related to the use of task models is a growing emphasis on Evidence-Centered Design (ECD; Mislevy, Steinberg, & Almond, 2003). ECD provides a detailed framework for the entire assessment process. Most relevant to the present discussion are the first two stages, domain analysis and domain modeling. As explained by Misley & Haertel (2006):

> The Domain Analysis layer concerns gathering substantive information about the domain to be assessed. If the assessment being designed is to measure science inquiry at the middle school level, domain analysis would marshal information about the concepts, terminology, representational

> forms, and ways of interacting that professionals working in the domain use and that educators have found useful in instruction (p. 7).

The domain modeling stage entails determining the test-based behaviors that will be used as evidence of claims about students' mastery of elements of the domain, as well as specifying the *warrants* that indicate why these behaviors are evidence of that mastery.

There is as yet no evidence indicating whether the ECD approach will lessen or exacerbate the forms of predictability that are problematic under high-stakes conditions. It seems likely that this will depend on the nature of the behaviors specified in the domain-modeling stage. If the application of ECD focuses on "broad claims" (Herman, 2013), it may not impede appropriately varied sampling of substantive elements. However, focusing ECD on the wording of the specific standards is likely to produce undesirable narrowing, which in turn is likely to generate score inflation.

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—scores that are inflated with respect to the target of inference—and undesirable effects on instruction rather than error. Thus, 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, although it does not appear to have led to substantial changes in large-scale assessment programs. 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 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 will have implicitly deemed content included in the NAEP frameworks but excluded from the Common Core as not part of their targets of inference. However, 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 stakeholders. Presumably, employers hiring individuals in different occupations and college programs of various sorts will have differing expectations about what "readiness" means.

Some choices about inclusion and weighting reflect explicit decisions of policymakers, such as staff of state education agencies or of other organizations responsible for testing programs, 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.

Such variations in weights are not unique to New York's assessments (Jennings & Bearak, 2014).

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, and other unnecessary regularities in task demands, such as some aspects of scoring rubrics. 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,

consider two items testing knowledge of the volume of rectangular prisms. 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 into the formula and simple arithmetic. However, even if this simplification were eliminated, the items assessing this standard are narrowed in important ways. 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, items about this skill in the Singapore Primary School Leaving Examination, taken at the end of the sixth grade, are far less narrow; they require students to apply the formula, which is not given, to partially full items of different sizes and compare them (Singapore Examinations and Assessment Board, 2009). Illustrative Massachusetts and Singapore items are shown in Figure 3.

The implications of these differences among items depends on stakeholders' targets of inference. If stakeholders only infer that students can apply the formula to completely full rectangular prisms, the Massachusetts items are reasonable. On the other hand, if stakeholders infer that success on the item indicates an ability to apply the formula to a variety of real-world problems, the Massachusetts standard and items represent an unwarranted narrowing with respect to the target.

Figure 3 about here

Specific options for test design. While calls for sampling from a broad range of items are not uncommon (e.g., Hanushek, 2009), I am aware of only two specific suggestions in the published literature for how this might be done, one calling for the use of two separate tests and the second suggesting embedding less predictable items in the test used for accountability. There is as yet almost no empirical work exploring the advantages and disadvantages of these (and of as yet unspecified alternative) approaches. This is an area in which there is a pressing need for research.

Neal (Barlevy & Neal, 2012; Neal, 2013) suggested administering two different tests, one to be used solely for accountability and a second for other purposes, such as describing the progress of individual students. Consistent with the argument made here, he asserted that the item similarities required for linking and scaling encourage educators to respond in ways that inflate scores. He also suggested that to support an effective performance incentive system, one needs neither information about change nor putatively interval scales, rather only ordinal cross-sectional data and reliable ranking. 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. Neal did not specify the nature of this variation. 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. Because this test would not be used to hold educators accountable,

educators would have weaker incentives to prepare students for it in ways that inflate scores.

This is an intriguing idea, but it raises numerous issues that the field has yet to address. 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. However, 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 because there are mechanisms other than explicit sanctions and rewards that can make a test high-stakes. (The first empirical study that 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 variations 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.

Koretz & Beguin (2010) suggested the alternative approach of embedding novel items in a single assessment, which we labeled *self-monitoring assessments* (SMAs). We suggested maintaining current design principles for much of an operational test while adding audit components. These audit components would comprise items designed to

maintain coverage of important performance elements while undoing predictable patterns that encourage inappropriate test preparation and score inflation. Inflation could still occur on the conventional items, and discrepancies in performance between the operational and audit components would signal likely inflation. SMAs should be able to provide estimates of score inflation at the school level, although the reliability of these estimates might be too low to apply to individual schools rather than categories of schools. 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 and show variations in inflation across categories of schools, (e.g., Koretz, Jennings, Ng, Yu, Braslow, & Langi, 2014), 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. The addition of audit items would increase test length, and constraints on length could lead to short audit components and hence unreliable estimates of inflation. 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 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).

Other alternatives may be feasible. For example, 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.

A likely consequence of making items more diverse would be greater instability in the relative performance of individuals and groups. This instability could take several forms depending on the design of the testing system, but to my knowledge, there is as yet no research directly addressing this question. For example, In Neal's design, it is possible that internal consistency reliability in the accountability test (e.g., Cronbach's alpha, or the person-by-item interaction in a cross-sectional generalizability analysis) might decrease compared to a conventional test, but it seems likely that a more substantial effect might be an increase in the person-by-occasion (or school-by-occasion) interaction across years because the substitution of novel items will likely change the alignment of

the test with the implemented curriculum. Designs that provide both operational and audit components, such as the Koretz & Beguin proposal, are likely to provide inconsistencies in school rankings because of bias as well as measurement error because schools affected by larger amounts of inappropriate test preparation would rank lower on the audit component (e.g., Ng & Koretz, 2013). Moreover, these differences in rankings are likely to be unstable over time.

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.

A comparison of some issues raised by the Neal and Koretz & Beguin approaches is shown in Table 1.

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.

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—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 linking items will *displace* the line, resulting in an incorrect transformation of scale. The conventional procedure of checking for items that fall far from this line 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) confirm 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 using 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 cross-sectional 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). Validity theorists have warned for some time that high-stakes uses of tests are likely to generate threats to validity. We now have ample empirical research confirming that this concern is warranted, that is, documenting the problems of score inflation and inappropriate test preparation. The argument of this paper is that the practices of educational measurement

have not been changed sufficiently to address these problems. The purpose of this paper is to spur debate about how various aspects of the testing enterprise—test design, linking, and validation—should be changed in response to this documented problem, that is, to create assessment systems that are less vulnerable to score inflation from inappropriate test preparation.

Some might argue that I am presenting too harsh a view of the field's weak response to these problems, but I can point to very few efforts to adapt measurement practices to address them. Some in the field have made clear and specific suggestions—for example, Haertel's suggestions (2013) about validation—but I cannot identify many responses in large-scale testing programs. As I noted, I can find only two published proposals for design changes specifically tailored to reduce inflation. These are both recent, but it still may be telling that to my knowledge, there has been as yet no trial of Neal's proposal (one is being discussed as I write this), and while there have been four pilot efforts to test the SMA design principle, I have been responsible for all of them. I cannot find any responses to the threats high-stakes uses pose to NEAT linking.

The responsibility for this dearth of responses is of course not entirely that of the measurement field. Some of the largest 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.[2]

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 our proposals for validation of testing programs—to reflect the importance of accountability-oriented uses of tests. Such changes could have an appreciable if long-term impact. 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 tailored to that use, then sponsors of testing programs may begin accepting the needed changes.

---

[2] One reviewer asked for specific examples. I choose not to identify particular jurisdictions or individuals. However, I have personally experienced this in at least three states, in all cases being told explicitly that the reason was the risk of unwanted findings, and the same happened to one of my students in a fourth state just last year.

We can also find opportunities to do the needed research and development without waiting for states and others sponsors of tests to agree to large-scale changes in design, linking, or validation. The two alternative designs discussed above certainly are not the only possibilities for new approaches to design, and I emphasized how little we still know about the practicality, advantages, and disadvantages of these two. There is clearly a need to design and evaluate alternatives. The problem of linking under high-stakes conditions is a particularly difficult one that could benefit from research.

In addition, there are possibilities for research that might afford opportunities to improve conventional designs, without turning to changes in design as substantial as those proposed by Neal (2013) or Koretz & Beguin (2010). For example, systematic evaluation of predictable regularities in test items (similar to that done by Holcombe, et al., 2013) and monitoring of test-preparation materials could provide information about unneeded and problematic patterns that could be addressed in the preparation of new forms of extant tests.

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.

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 debate, research, development, and evaluation that will make needed improvements feasible.

References

Barlevy, G., and Neal, D. (2012). Pay for percentile. *American Economic Review, 102*(5), 1805-1831.

Bejar, I. I., Lawless, R. R., Morley, M. E., Wagner, M. E., Bennett, R. E., & Revuelta, J. (2003). A feasibility study of on-the-fly item generation in adaptive testing. *Journal of Technology, Learning, and Assessment, 2*(3). Available from http://www.jtla.org.

Brennan, R. L. (2007). Personal communication, October 11.

Common Core State Standard Initiative (2012). *Mathematics*. Washington, D.C.; Council of Chief State School Officers. Last retrieved on July 22, 2013 from http://www.corestandards.org/Math.

Fitzpatrick, A. R. (2008). NCME 2008 Presidential Address: The impact of anchor test configuration on student proficiency rates. *Educational Measurement: Issues and Practice, 27*(4), 34-40.

Fuller, B., Gesicki, K., Kang, E., & Wright, J. (2006). *Is the No Child Left Behind Act working? The reliability of how states track achievement*. University of California, Berkeley: Policy Analysis for California Education. Retrieved July 16, 2013, from http://www.eric.ed.gov/PDFS/ED492024.pdf

Haertel, E. (2013). How is testing supposed to improve schooling? *Measurement: Interdisciplinary Research and Perspectives, 11* (1-2), 1-18.

Hambleton, R. K., Jaeger, R. M., Koretz, D., Linn, R. L., Millman, J., and Phillips, S. E. (1995). *Review of the Measurement Quality of the Kentucky Instructional Results Information System, 1991-1994.* Frankfort: Office of Education Accountability, Kentucky General Assembly, June.

Hamilton, L. S., Stecher, B. M., Marsh, J. A., McCombs, J. S., Robyn, A., Russell, J., Naftel, S., & Barney, H. (2007). *Standards-based accountability under No Child Left Behind: Experiences of teachers and administrators in three states*. Santa Monica, CA: RAND.

Hamilton, L. S., Stecher, B. M., & Yuan, K. (2008). *Standards-based reform in the United States: History, research, and future directions.* Santa Monica, CA: RAND. Last retrieved June 26, 2013, from http://www.rand.org/pubs/reprints/2009/RAND_RP1384.pdf.

Hanushek, E. (2009). Building on No Child Left Behind. *Science, 326,* 802-803.

Herman, J. L. (2013). Assessing the new Common Core tests. *Harvard Education Letter, July/August*, 6-8.

Ho, A. D. (2007). Discrepancies between score trends from NAEP and state tests: A scale invariant perspective. *Educational Measurement: Issues and Practice, 26*(4), 11-20.

Holcombe, R., Jennings, J., & Koretz, D. (2013). The roots of score inflation: An examination of opportunities in two states' tests. In G. Sunderman (Ed.), *Charting reform, achieving equity in a diverse nation*, 163-189. Greenwich, CT: Information Age Publishing. http://dash.harvard.edu/handle/1/10880587.

Holmstrom, B., & P. Milgrom. (1991). Multitask principal-agent analyses: Incentive contracts, asset ownership, and job design. *Journal of Law, Economics and Organization*, 7, 24–52.

Hoover, H. D., Dunbar, S. B., Frisbie, D. A., Oberly, K. R., Bray, G. B., Naylor, R. J., Lewis, J. C., Ordman, V. L., & Qualls, A. L. (2003). *The Iowa Tests Interpretive Guide for Teachers and Counselors, Forms A and B, Levels 9-14.* Chicago: Riverside.

Jacob, B. A. (2005). Accountability, incentives and behavior: The impact of high-stakes testing in the Chicago public schools. *Journal of Public Economics, 89*(5-6), 761-796.

Jacob, B. A. (2007). Test-based accountability and student achievement: An investigation of differential performance on NAEP and state assessments (NBER Working Paper No. 12817). Cambridge, MA: National Bureau of Economic Research. Retrieved March 22, 2011, from http://www.nber.org/papers/w12817

Jennings, J. L., & Bearak, J. M. (2014). "Teaching to the test" in the NCLB era: How test predictability affects our understanding of student performance. *Educational Researcher,43*(8), 381-389.

Kane, M. T. (2006). Validation. In R. L. Brennan (Ed.), *Educational measurement (4th ed.*, pp. 17–64). Westport, CT: American Council on Education/Praeger.

Klein, S. P., Hamilton, L. S., McCaffrey, D. F., & Stecher, B. M. (2000). *What do test scores in Texas tell us? Education Policy Analysis Archives, 8*(41). Retrieved July 16, 2013, from http://epaa.asu.edu/ojs/article/view/440/563

Kolen, M. J., and Brennan, R. L. (2004). *Test equating, scaling, and linking, second edition*. New York: Springer-Verlag. *Trends in Educational Achievement*. Washington, D.C.: Congressional Budget Office, April.

Koretz, D. (2007). Using aggregate-level linkages for estimation and validation: comments on Thissen & Braun & Qian. In N. J. Dorans, M. Pommerich, & P. W. Holland (Eds.), *Linking and aligning scores and scales* (339-353). New York: Springer-Verlag.

Koretz, D., and Barron, S. I. (1998). *The Validity of Gains on the Kentucky Instructional Results Information System (KIRIS).* MR-1014-EDU, Santa Monica: RAND.

Koretz, D., and Beguin, A. (2010). Self-monitoring assessments for educational accountability systems. *Measurement: Interdisciplinary Research and Perspectives, 8*(2-3: special issue), 92-109.

Koretz, D., and Hamilton, L. S. (2006). Testing for accountability in K-12. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., 531-578). Westport, CT: American Council on Education/Praeger.

Koretz, D., Jennings, J. L., Ng, H. L., Yu, C., Braslow, B., and Langi, M. (2014). *Auditing for score inflation using self-monitoring assessments: Findings from three pilot studies.* A working paper of the Education Accountability Project at the Harvard Graduate School of Education. Last accessed on January 2, 2015 at http://projects.iq.harvard.edu/files/eap/files/combined_audit_paper_submission_11_25_14_wp.pdf.

Koretz, D., Linn, R. L., Dunbar, S. B., and Shepard, L. A. (1991). The Effects of High-Stakes Testing: Preliminary Evidence About Generalization Across Tests. In R.L. Linn (chair), *The Effects of High Stakes Testing,* symposium presented at the annual meetings of the American Educational Research Association and the National Council on Measurement in Education, Chicago, April. http://dash.harvard.edu/handle/1/10880553.

Koretz, D., McCaffrey, D., and Hamilton, L. (2001). *Toward a Framework for Validating Gains Under High-Stakes Conditions*. CSE Technical Report 551. Los Angeles: National Center for Research on Evaluation, Standards, and Student Testing.

Koretz, D., Mitchell, K., Barron, S., and Keith, S. (1996). *The Perceived Effects of the Maryland School Performance Assessment Program.* CSE Technical Report No. 409. Los Angeles. CA: National Center for Research on Evaluation, Standards, and Student Testing.

Koretz, D., Yu, C., & Braslow, D. (2013). *Auditing for Score Inflation using Newly Tested Standards*. Working paper of the Educational Accountability Project. http://projects.iq.harvard.edu/files/eap/files/2011_op_audit_paper_11.14_wp.pdf

Lindquist, E. F. (1951). Preliminary considerations in objective test construction. In E. F. Lindquist (Ed.), *Educational measurement* (2nd ed., 119–158). Washington: American Council on Education.

Linn, R. L. (1997). Evaluating the validity of assessments: the consequences of test use. *Educational Measurement: Issues and Practice, 16* (2), 14–16.

Linn, R. L., Baker, E. L., & Dunbar, S. B. (1991). Complex, performance-based assessment: expectations and validation criteria. *Educational Measurement: Issues and Practice, 20* (8), 15–21.

Madaus, G. F. (1988). The distortion of teaching and testing: High-stakes testing and instruction. *Peabody Journal of Education*, 65(3), 29-46.

Massachusetts Department of Education (2000). *Massachusetts Mathematics Curriculum Framework*. Malden, Massachusetts: Author.

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

Michaelides, M. P., and Haertel, E. H. (2004). *Sampling of common items: An unrecognized source of error in test equating.* Los Angeles: Center for Research on Evaluation, Standards, and Student Testing (CRESST), CSE Report 636.

Mislevy, R. J., & and Haertel, G. D. (2006). Implications of Evidence-Centered Design for educational testing. *Educational Measurement: Issues and Practice, 25*(1), 6-20.

Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (2003). On the structure of educational assessments. *Measurement: Interdisciplinary Research and Perspectives*, *1*, 3–67

Morley, M.E., Bridgeman, B., and Lawless, R. R. (2004). *Transfer Between Variants of Quantitative Items*. Princeton , N.J.: Educational Testing Service (GRE Board Research Report No. 00-06R).

Neal, D. (2013). The consequences of using one assessment system to pursue two objectives. (NBER Working Paper No. 19214). Cambridge, MA: National Bureau of Economic Research.

Ng, H. L., & Koretz, D. (2013). Sensitivity of school-performance ratings to the test used. A working paper of the Education Accountability Project at the Harvard Graduate School of Education, http://projects.iq.harvard.edu/files/eap/files/houston_paper_wpdraft_032513_1.pdf.

Partnership for Assessment of Readiness for College and Careers (2012). *PARCC Model Content Frameworks, Mathematics, Grades 3-1, Version 3.0*. Author:

Pedulla, J. J., Abrams, L. M., Madaus, G. F., Russell, M. K., Ramos, M. A., & Miao, J. (2003). *Perceived effects of state-mandated testing programs on teaching and learning: Findings from a national survey of teachers*. Boston, MA: National Board on Educational Testing and Public Policy. Retrieved July 15, 2013, from http://www.bc.edu/research/nbetpp/statements/nbr2.pdf.

Porter, A., McMaken, J., Hwang, J., & Yang, R. (2011). Common Core Standards: The new U.S. intended curriculum. *Educational Researcher, 40*(3), 103-116.

Romberg, T. A., Zarinia, E.A., & Williams, S.R. (1989). *The influence of mandated testing on mathematics instruction: Grade 8 teachers' perceptions*. Madison: National Center for Research in Mathematical Science Education, University of Wisconsin–Madison.

Rothstein, R. (2008). *Holding accountability to account: How scholarship and experience in other fields inform exploration of performance incentives in education.* Nashville, TN: National Center on Performance Incentives, Vanderbilt University. Last accessed from http://s4.epi.org/files/2014/holding-accountability-to-account.pdf on September 25, 2014.

Rubinstein, J. (2000). *Cracking the MCAS grade 10 math*. New York: Princeton Review Publishing.

Shepard, L. A. (1988). The harm of measurement-driven instruction. In *Annual meeting of the American Educational Research Association*. Washington, DC.

Severson, K. (2011). A scandal of cheating, and a fall from grace. *The New York Times*, September 7, p. A16. Last retrieved on June 5, 2013 from http://www.nytimes.com/2011/09/08/us/08hall.html?pagewanted=all&_r=0.

Shepard, L. A. (1997). The centrality of test use and consequences for test validity. *Educational Measurement: Issues and Practice, 16* (2), 5–8, 13, 24.

Shepard, L.A., & Dougherty, K.C. (1991, April). *Effects of high-stakes testing on instruction*. Paper presented at the annual meeting of the American Educational Research Association and National Council on Measurement in Education, Chicago.

Singapore Examinations and Assessment Board (2009). *PSLE Examination Questions 2005-2009*. Singapore: Author.

Smarter Balanced Assessment Consortium (2012). *General Item Specifications, Draft 1*. Author, April 16. Last accessed November 22, 2013 from http://www.smarterbalanced.org/wordpress/wp-content/uploads/2012/05/TaskItemSpecifications/ItemSpecifications/GeneralItemSpecifications.pdf.

Stecher, B. (2002). Consequences of Large-Scale High-Stakes Testing on School and Classroom Practice. In L. Hamilton, B. M. Stecher, & S. Klein (Eds.), *Making Sense of Test-Based Accountability in Education* (pp. 79-100). Santa Monica, CA: RAND Corporation.

Stecher, B.M., & Mitchell, K.J. (1995). *Portfolio driven reform: Vermont teachers' understanding of mathematical problem solving* (CSE Tech. Rep. 400). Los Angeles: Center for Research on Evaluation, Standards, and Student Testing.

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.

COMPUTERS SOLD

Year	Computers Sold
2002	3.2×10^5
2003	8.4×10^3
2004	5.9×10^5
2005	1.2×10^4

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. 10th-grade MCAS (March retest), 2010: formula provided

33 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. 6th-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.

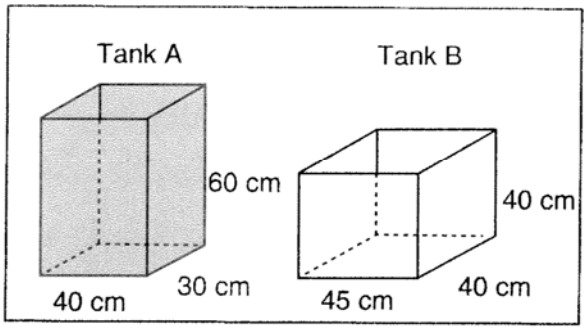

Figure 1

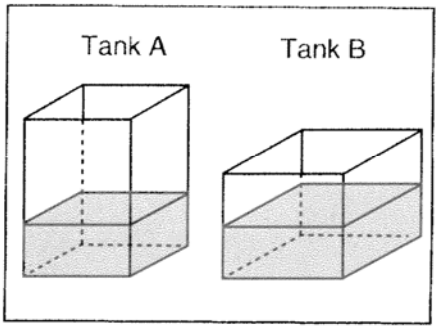

Figure 2

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, 10th grade Massachusetts MCAS and Singapore Primary School Leaving Exam (6th 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).

Tables

Table 1.

Comparison of the current status of the Neal and Koretz & Beguin approaches

	Neal	Koretz & Beguin
Clear how items should differ	Not yet	Not yet
Components[1] can be placed on one scale	No	Depends on design
Increase in testing time	Roughly doubled	Smaller
Risk of motivational differences between components	High	Low
Impact on educators' behavior	Not yet known	Not yet known
School-level estimates of score inflation possible	No	Yes
Reliability of estimates from audit component	Not yet clear	Low

[1] Components are the two tests in the Neal design and the audit and nonaudit components in the Koretz & Beguin design.