Center for
K–12 Assessment
& Performance Management

An independent catalyst and resource for the improvement of measurement and data systems to enhance student achievement.

Exploratory Seminar:

Measurement Challenges Within the Race to the Top Agenda

December 2009

Implications of Current Policy for Educational Measurement

Daniel Koretz

Created by Educational Testing Service (ETS) to forward a larger social mission, the Center for K–12 Assessment & Performance Management has been given the directive to serve as an independent catalyst and resource for the improvement of measurement and data systems to enhance student achievement.

Some Implications of Current Policy for Educational Measurement

Daniel Koretz

Harvard Graduate School of Education, Cambridge, Massachusetts

This paper was presented by Daniel Koretz at the Exploratory Seminar: Measurement Challenges Within the Race to the Top Agenda, December 2009. Download copies of other papers presented at the seminar at http://www.k12center.org/publications.html.

A fundamental shift in the uses of large-scale assessment has been underway in the United States for 40 years. Low-pressure, primarily diagnostic uses of test scores have been largely supplanted by test-based monitoring and accountability. Over the decades during which test-based accountability (TBA) has developed, the pressure exerted on educators to improve scores has increased dramatically.

The field of measurement has not kept pace with this transformation of testing. Certainly, the change has spurred a number of important innovations in testing, some good and others bad. However, the field has not ad*a*pted some of its core practices to reflect the challenges posed by TBA. Fundamental changes are needed at all stages of the testing enterprise, beginning with test design and ending with validation.

The Policy Context

In the years following World War II, large-scale testing was generally a low-stakes enterprise. Many states left it to districts to decide whether and how to test. For the most part, testing was seen as a diagnostic exercise, and scores had no serious consequences for most students and teachers. This state of affairs began to change in the 1960s, with the establishment of the National Assessment of Educational Progress (NAEP) to monitor the achievement of the nation's youth and the imposition of test-based evaluation requirements for programs funded under Title I of Elementary and Secondary Education Act (see Koretz & Hamilton, 2006).

More substantial change began with the minimum-competency testing movement of the 1970s (Jaeger, 1982), which initiated on a large scale what Popham, Cruse, Rankin, Sandifer, and Williams (1985) later dubbed measurement-driven instruction: the use of testing to generate direct incentives to change behavior. Minimum competency testing did not last long as a national movement, but it marked the inception of TBA as a cornerstone of education policy. Since then, the nation has seen three or four waves of test-based reform, and the form of TBA has varied markedly both across jurisdictions and over time (Koretz & Hamilton, 2006). The general trend, however, has been an increase in pressure, particularly an increase in pressure on educators. No Child Left Behind (NCLB) and the Race to the Top (R2T) are just the most recent, and in many respects the most extreme, examples of this trend. At this

point, it is not an exaggeration to say that educational practice in the United States is dominated by a drive to raise scores on large-scale assessments.

Whether TBA has caused an improvement in student achievement in tested subjects is a matter of intense controversy, but that is a controversy that need not be resolved for present purposes. What is clear is that TBA has often generated substantial distortions of practice (e.g., Stecher, 2002) and inflation of test scores, that is, increases in scores larger than the actual improvements in the latent proficiencies that tests are intended to estimate. In some instances, this inflation has been severe, dwarfing real improvements in learning (e.g., Jacob, 2005, 2007; Klein, Hamilton, McCaffrey, & Stecher, 2000; Koretz & Barron, 1998; Koretz, Linn, Dunbar, & Shepard, 1991).

Because the documented distortions of practice and score inflation are common and often severe, it is essential that they be confronted squarely in the design of both testing programs and the accountability programs into which testing is embedded. In this paper, I will focus on the design of testing programs.

Incomplete Sampling in Educational Testing

The distortions of practice and score inflation revealed by research on the effects of TBA should have surprised no one. Within the field of measurement, they were predicted at least as early as a 1951 chapter by E. F. Lindquist. Even though testing was then relatively low-stakes, Lindquist warned that:

> 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, pp. 152-153)

The problems of which Lindquist warned are a specific manifestation of a much more general problem in performance accountability systems. A quarter century after Lindquist's chapter was published, Don Campbell, one of the seminal thinkers in the emerging discipline of program evaluation, wrote that:

> The more any quantitative social indicator is used for social decision making, the more subject it will be to corruption pressures and the more apt it will be to distort and corrupt the social processes it is intended to monitor. (Campbell, 1979, p. 87)

Decades of research have confirmed that this is a general phenomenon that appears in a wide variety of fields (e.g., Rothstein, 2008). The problem is so general that it is often called "Campbell's Law."

A primary cause of Campbell's Law is *incomplete measurement of desired outcomes*. Economic theory about the functioning of performance accountability systems (e.g., Baker, 2002) focuses on the distortions that arise because of performance measures are usually incomplete and may weight given

actions differently than would an ideal measure. People responding to this system—*agents* in the terminology of economics, educators, and students in the specific case of TBA—have an incentive to focus on the measured portion of the goals at the expense of others and to allocate efforts according to the weights assigned by the performance measure rather than the importance of actions to the ultimate goals of the system.

Those in the field of measurement might rightly ask, "What is new about this?" The notion that an achievement test is just a sample from a larger domain—in other words, that it is necessarily incomplete—is a core principle of measurement and has been the starting point for generations of work in the field. However, a careful look at the stages of this sampling and their effects under low-stakes and high-stakes conditions reveals critically important issues to which the field has given insufficient attention.

To clarify this, it is necessary to consider all stages of sampling that take us from the goals of education to an operational achievement test. A conventional view of the creation of an achievement test is represented in Figure 1. The first row of Figure 1 represents all of the goals of education, which are not limited to achievement as typically defined. For example, we expect schools to encourage persistence, to teach students to work in groups, to encourage creativity, and so on. Standardized achievement tests, as conventionally understood, exclude these nonachievement outcomes and are perforce limited to the second row of Figure 1. Not all testable domains of achievement are in fact tested, of course, and that brings us to the third row of Figure 1.

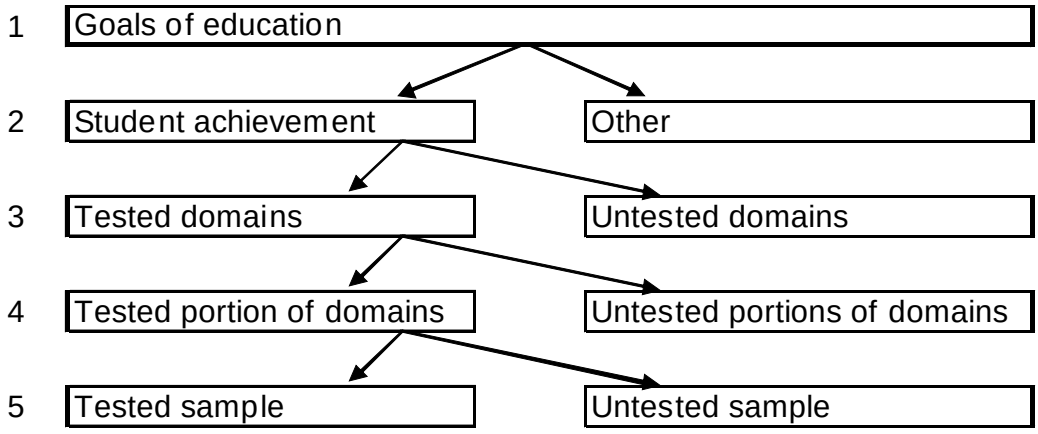

Figure 1. Stages of sampling in the creation of an achievement test.

Even once the domains to be tested have been selected, one has to decide which aspects of those domains will be measured. This stage of sampling is represented by the fourth row of Figure 1. For example, does eighth-grade mathematicsinclude algebra? If so, which aspects of algebra? Some portions of a given domain are easier to test than others, and those tend to be the ones chosen for testing. And, finally, from this defined subset of the selected domain, one has to sample actual tasks for purposes of creating a test. In current TBA systems, educators are held accountable for the sample represented by

the bottom row of Figure 1, not the broader goals represented by the first row, or even the intermediate levels in between.

Consequences of Incomplete Sampling

Under low-stakes conditions, there are several consequences of these several stages of sampling. The first, which was stressed by Lindquist, is a consequence in large part of the top levels of sampling: test scores alone cannot suffice to evaluate schools or teachers. This is generally ignored in today's policy community.

The second consequence under low-stakes conditions is a consequence of the sampling from row 3 in Figure 1 on: different tests, even if built to similar standards of quality, often provide results that are often modestly different and occasionally dramatically different. For example, Trends in

International Mathematics and Science Study (TIMSS) and Program for International Student Assessment (PISA) rank countries somewhat differently, and in recent years, the performance of U.S. fourth grade students increased more than three times as much on NAEP as on TIMSS. This too is widely ignored by policymakers and the media.

Finally, there are consequences of the final two stages of sampling. One is possible inadequacies in the sampling of knowledge and skills at these last two stages—for example, construct underrepresentation or sampling that produces construct-irrelevant variance. The second is imprecision, or measurement error. These final stages of sampling also underlie the difficulties that arise in horizontal and vertical linking and scaling.

For the most part, notwithstanding Lindquist's warning, the field of measurement has largely ignored the top levels of sampling in Figure 1. In part, this reflects the fact that test designers are usually brought into the educational process only after others have decided what should be measured. However, this lack of attention to the top levels of sampling may also stem from the uses to which standardized achievement tests were once put—or, at least, the purposes to which some of their designers said they should be put. Standardized tests were originally intended to serve as incomplete measures of a subset of educational goals. For example, the following advice is included in a manual for educators using the Iowa Tests of Basic Skills:

> The primary purpose of using a standardized achievement battery is to provide information that can be used to improve instruction . . . .Though standardized achievement scores cannot and should not replace teacher observations and classroom assessment information, they can provide unique supplementary information . . . .No assessment method or instrument can supply the full range of information required to evaluate the entire school program, or even the complete academic curriculum . . . . Standardized test scores alone should not be used [to evaluate the entire school program] because achievement batteries are not designed to cover the full range of objectives that make up the school curriculum. (Hoover et al., 1993, pp. 7, 9)

For many purposes, the field of measurement has also often ignored the sampling that takes us from row 3 to row 4 of Figure 1. Some of the sampling between rows 3 and 4 is taken as a given because it is defined in part by the standards or curriculum framework of the client. Thus, the evaluation of content-based validity evidence could focus on both the final stages of the sampling in Figure 1, but it sometimes focuses only on the final stage. For example, in evaluating content-based evidence of validity of seven of the Massachusetts Comprehensive Assessment System (MCAS) tests, Hambleton, Smith, and Zhao (2006) noted:

> According to the *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 1999), three requirements about test content apply to programs such as the MCAS: (1) the content of the tests must be consistent or in alignment with the content specifications for the tests, (2) the tests must show content diversity over time, if the tests from a given year don't cover all of the curricula, and (3) the test items themselves must assess the learning standards to which they are referenced or linked. (Hambleton et al., p. 1)

This is a focus on the final stage of sampling in Figure 1. Evaluation of reliability (e.g., test-retest reliability of parallel forms) also generally focuses on the sampling required to obtain row 5.

Thus, it is only a modest oversimplification to say that, traditionally, the design of achievement tests, the evaluation of tests, and the validation of inferences based on scores have all focused on the final two stages of sampling in Figure 1, especially the final stage. The quality of a test and the validity of inferences based on it are evaluated primarily in terms of three factors: the test's representativeness of the constructs selected (by others) for standardized measurement; its precision; and the adequacy with which the resulting performances are linked, scaled, and reported. Moreover, quality and validity are evaluated in terms of the initial representativeness of the tested sample, often using data obtained before any operational administration of the test. This last issue, as I will show shortly, is critically important.

While this incomplete focus was reasonable and productive under low-stakes conditions, it is entirely inadequate under the high-stakes conditions imposed by today's TBA systems. The concerns of traditional psychometrics still pertain, but they are no longer sufficient basis for evaluating quality validity. There are two reasons: tests are now used to evaluate schools (and, soon, teachers), and one now confronts Campbell's Law. Far more than in Lindquist's day, educators have an incentive to focus on the small and unrepresentative sample that is actually tested. In practice, this means a much greater risk of narrowed and otherwise degraded instruction, inappropriate test preparation, and inflated scores.

Therefore, under high-stakes conditions, it is no longer defensible to consider only the quality of the initial representation of the domains selected for measurement. Validation needs to focus on representation of the constructs after any unwanted behavioral responses to testing. Similarly, test design must focus on methods for minimizing the undesired behavioral responses that undermine validity after operational use as begun.

Three Forms of Campbell's Law in Test-Based Accountability

The distortions described by Lindquist and Campbell take three different forms in the case of test-based accountability systems. The three have different implications for test design and validation.

The first form taken by Campbell's Law in TBA systems is that the higher the stakes, the greater the incentive is to reallocate resources so as to de-emphasize or ignore the goals of education not selected for including in the measurement and accountability system—that is, the goals not sampled in the first two stages of Figure 1—in order to shift those resources to goals that are counted by the system. Examples include the innumerable anecdotes about schools shifting time away from untested subjects to increase time allocated to those tested and counted in the accountability system. My colleagues and I, with the sampling at row 3 of Figure 1 in mind, labeled this process between-subjects reallocation (Koretz & Hamilton, 2006; Koretz, McCaffrey, & Hamilton, 2001). However, that term only subsumes part of the issue, because this process can also entail reallocating resources between measurable achievement and other goals.

Regardless of the terminology, this form of reallocation is a serious concern for policy, but it is not an issue for measurement. If time is taken away from teaching history, and that time is devoted to high-quality, effective teaching in mathematics, the resulting gains in mathematics test scores should be valid. Whether this is a desirable outcome, however, is a serious matter for policy. One of the key issues in designing accountability systems that are more effective and less vulnerable to Campbell's Law is finding ways to make other goals of education count in order to lessen undesirable aspects of this reallocation. However, in this paper, I focus on the measurement issues, not the design of the accountability systems into which tests are embedded, and therefore I will not address this form of reallocation further.

The second form taken by Campbell's Law is reallocation within subjects, as teachers and students attempt to focus their attention on the aspects of tested subjects that will be measured and rewarded. In our discussion of these responses, my colleagues and I have used within-subject reallocation to refer specifically to reallocation among elements of performance that are substantively important, that is, potentially relevant to the inference based on scores. Think of the construct about which test-based inferences are made as comprising a bundle of elements of performance, all of which matter to the users of test scores. In Figure 1, these would be the elements of performance found in rows 3 or 4, depending on the inference. The test samples only a modest share of these. If this sampling is predictable—and, to a substantial degree, it is—then teachers and students have a stronger incentive to focus on the tested sample, at the expense of untested elements of the domain. This form of reallocation is common (e.g., Stecher, 2002).

Reallocation within subjects, unlike reallocation between subjects, should be a core concern of the measurement community because it can lead to inflation of test scores. Even if the test initially provides a reasonable representation of the tested domain, it will not provide a reasonable representation after substantial reallocation (Koretz, 2008; Koretz & Hamilton, 2006).

Anyone familiar with schools knows that reallocation within subjects is now commonplace. Indeed, state and district education agencies often encourage it, either implicitly, for example, by providing previous test forms as a guide for study, or explicitly, by informing educators which portions of the domain are given substantial weight by the test. This latter approach often goes by the disarming name of "power standards." A nice example of power standards is a presentation file provided by the Quincy, Massachusetts, education agency to its high school mathematics teachers (Quincy Public Schools, 2004). The first page of the presentation shows three texts used in Quincy's high school. When teachers choose their text, they are given a table of contents. Picking a chapter brings one to a page such that represents that chapter of the algebra text. Each section of the chapter is given a row, and in that row is listed all of the items testing the content of that section that appeared on the state's Grade 10 mathematics test over a period of four years. The item numbers are live links that take one to the actual items.

This sort of aid is just one of many that facilitate within-subject reallocation that can inflate scores. It makes it easier for teachers to capitalize on predictable recurrences and omissions of content to narrow their instructional focus to the tested sample rather than the domain from which the test is drawn.

Proponents of standards-based testing often brush off this concern about within-subject reallocation, arguing that if a test is aligned with standards, there is nothing wrong with aligning instruction with it. Alignment, however, is simply reallocation, with the added constraint that the material gaining emphasis must be consistent with standards. Alignment can still produce inflation if the material de-emphasized is important for the inference (Koretz, 2005).

The third form of Campbell's Law in TBA systems is what my colleagues and I have called coaching. The term coaching is used in many different ways, often as a generic term for test preparation, but we have used it to refer to a focus on fine details of the test. The difference between coaching and reallocation may seem subtle, but it has important implications for test design and validation.

Performance on a test is determined by more than the selection of substantively important elements of content or skill. First, performance is also influenced by choices of content that are too fine-grained to be of any substantive importance. For example, a test framework may specify that students should be familiar with the properties of simple polygons. But what kind of polygons, and what forms should that familiarity take? These decisions are represented by the final row in Figure 1, and they are often of no particular importance to the inference. If they are repeated over time, however, educators can focus on them at the cost of other aspects of that portion of the domain. One can find predictions of recurrences of these minor content details in many test-preparation materials. We refer to the focus on these unimportant details as substantive coaching (Koretz & Hamilton, 2006).

Second, performance can also be influenced by aspects of test items that are entirely unrelated to the inference, such as some aspects of item format or response demands. Test preparation materials can alert educators to recurrences of these attributes as well. We refer to a focus on such details as nonsubstantive coaching. We also include under the rubric of coaching the testing of test-taking tricks, such as process of elimination.

An example of a coaching strategy that capitalizes on an incidental attribute of items testing the Pythagorean theorem is shown in Figure 2.

> Whenever you have a right triangle—a triangle with a 90-degree angle—you can use the Pythagorean theorem….The sum of the squares of the legs of the triangle (the sides next to the right angle) will equal the square of the hypotenuse (the side opposite the right angle)….
>
> Two of the most common ratios that fit the Pythagorean theorem are 3:4:5 and 5:12:13. Since these are ratios, any multiples of these numbers will also work, such as 6:8:10, and 30:40:50.

Figure 2. Coaching based on an incidental characteristic of test items.

This example is from The Princeton Review's materials for the Grade 10 MCAS test (Rubinstein, 2002, p. 56). The first paragraph, which is followed in the original by a diagram for illustration, might be characterized as weak but acceptable instruction. It is entirely free of any explanation that would help a student gain an intuitive understanding of the theorem, but if students memorized the rule, they would have gained a generalizable skill that would appear in real-world activities, as well as this particular test. The second paragraph, however, which provides students with common ratios (called elsewhere in the text popular Pythagorean ratios), is coaching, capitalizing on a substantively unimportant attribute of these items. These particular ratios are not especially common in the real world. They are "popular" with item-writers, however, because few students know how to calculate noninteger square roots, and therefore, using leg lengths that require a noninteger solution would conflate ability to calculate roots with knowledge of the Pythagorean theorem. If next year's item required a noninteger solution and allowed the use of calculators, the gains obtained by teaching popular Pythagorean ratios would vanish.

A more extreme example of coaching is taken from materials provided by the Montgomery County, Maryland, school district to help its teachers prepare students for the County's own algebra end-of-course exam (Figure 3). Items such as this are commonly used to determine whether students have learned to translate simple relationships presented verbally into algebraic form. With this test-preparation material, however, one can teach a mechanical solution to the problem that requires none of the skill that the item is intended to measure.

> The question on the review sheet for…[the] exam…reads in part:
>
> > The average amount that each band member must raise is a function of the number of band members, b, with the rule f(b)=12000/b.
>
> The question on the actual test reads in part:
>
> > The average amount each cheerleader must pay is a function of the number of cheerleaders, n, with the rule f(n)=420/n.

Figure 3. An example of coaching.

Test Design and Opportunities for Reallocation and Coaching

As the previous discussion and examples make clear, undesirable reallocation and coaching depend on the predictability of both content and nonsubstantive aspects of tests. This predictability varies among tests, but it is generally substantial. One finds predictability in the standards selected for testing or emphasis (e.g., Figure 2), in the content and skills sampled from within a given standard, and in the nonsubstantive aspects of item style and task demands. In some cases, items are near-clones of previously used (and publicly released) items. For example, Figure 4 shows one of numerous cases of near-clone items in the New York State Grade 7 mathematics test.

The changes between the two years were trivial: for example, deletion of the girl's name, changing the object to be measured from a watermelon to a serving of cheese, and changing a single distracter. With replication that is this extreme, there is no need to teach the concept of mass. All a student needs to know is that if the stem asks about measuring mass, whatever that may be, she should pick the answer with the word scale. In fact, with a bit of a gamble, one could coach students who speak no English to answer this question: if you see the letters mass in the stem, hunt for the letters scale in an answer. While the use of near-clones of this sort may be extreme, predictability of content and item style is not. Perusal of many test-preparation materials shows a substantial focus on these recurrences. It would be useful to consider a less extreme example than the one in Figure 4. Figure 5 shows an eighth-grade

From 2009 (Standard 7M9)

Which tool would be the most appropriate for Natasha to use when finding the mass of a watermelon?

- scale
- inch ruler
- meter stick
- measuring cup

From 2008 (Standard 7M9)

Which tool is most appropriate for measuring the mass of a serving of cheese?

- ruler
- thermometer
- measuring cup
- weighing scale

Figure 4. Near-clone items from New York State's Grade 7 math test, 2009 and 2008.

plane geometry item from the MCAS assessment. It is—at least in my opinion—a perfectly reasonable item. It is problematic, however, because it is predictable. The Princeton Review's test preparation materials for the MCAS include the following: "One triangle rule that is often tested on the MCAS exam is the third side rule. The rule is: The sum of every two sides of a triangle must be greater than the third side" (Rubinstein, 2000, p. 52). The area of plane geometry, and the Massachusetts standards more specifically, could be represented by a wide variety of content. These test-preparation materials alert teachers and students that this particular bit of content is repeatedly sampled, helping them reallocate effort. Predictable recurrence has converted a reasonable item into an opportunity for score inflation.

Under sufficiently low-stakes conditions with a high degree of test security, a degree of predictability is tolerable. A test of a large domain—particularly, a test that cannot be matrix-sampled because of a need for comparable individual scores—is likely to be a very small sample from the domain and therefore will necessarily be incomplete. It may still be seen as a reasonable initial representation of the domain. As long as teachers lack both the ability and incentive to focus on the specific content sampled—that is, if test security is strong and stakes are low—this predictability may be acceptable. Even under those conditions, educators' responses to testing may generate bias. However, this bias may be small enough that the scores remain a serviceable basis for the intended inferences.

Indeed, under such conditions, a degree of predictability offers several important technical advantages. For example, it permits a tighter link among forms, and it makes it closer to a matter of indifference to individuals which form they are administered. It is also cheaper and faster than designing and pretesting more novel items. For this reason, test developers often strive to create a high degree of similarity across forms. The advantages of item similarity led to the development of model-based, sometimes computerized systems for generating new items. The rationale for this approach is that "an item model can be thought of as a means of generating close variants with the intention that the isomorphs will be psychometrically and otherwise exchangeable and equivalent" (Morley, Bridgeman, & Lawless, 2004, p. 1). Unfortunately, even under the circumstances investigated in that study, which entailed a high degree of test security, performance generalized better across isomorphs than among less similar items designed to tap similar mathematical content (Morley et al.).

Eva has four sets of straws. The measurements of the straws are given below. Which set of straws could *not* be used to form a triangle?

A. Set 1: 4 cm, 4 cm, 7 cm
B. Set 2: 2 cm, 3 cm, 8 cm
C. Set 3: 3 cm, 4 cm, 5 cm
D. Set 4: 5 cm, 12 cm, 13 cm

Figure 5. An example of a predictable test item from the Massachusetts Comprehensive Assessment System (MCAS), Grade 8.

Thus, under low-stakes conditions, predicable recurrences and omissions both confer advantages and exact costs. The net benefit or cost is unclear and presumably varies with the nature of the test, the uses to which it is put, and the nature of the predicable sampling. However, one might argue that the net cost, if the costs outweigh the benefits, is often small.

Under high-stakes conditions, these calculations are entirely different. Educators and students have strong incentives to behave in ways that can undermine the ability of the tested sample to represent the domain. Predictable recurrences and omissions, either substantive or nonsubstantive, provide opportunities to create this bias, which research has shown can be very large.

What Should Be Done?

What can be done to lessen the problems described above? At least four different steps can be taken.

An essential step is to broaden the focus of educational accountability systems to focus on more than test scores, that is, to make the goals that are removed in the first levels of sampling in Figure 1 count in the accountability system. However, the focus of this paper is the design of testing programs, not the design of the accountability systems into which they are embedded, so I will not elaborate on this step.

One essential step for the measurement community is to adapt principles of test design to reflect the risks posed by high-stakes testing. This will require limiting unnecessary, predictable recurrences and omissions. While some have suggested richer, less predictable sampling of content and skills, this would be necessary but insufficient. It will also be necessary to limit predictability of nonsubstantive performance elements in order to lessen coaching. Decreasing predictability will improve the incentives created for educators and students. It will also make inappropriate coaching and reallocation more difficult and thereby lessen the risk of degraded instruction and score inflation. However, this change in design, while straightforward in principle, may prove difficult in practice. It will obviously increase test-development costs. More problematic, it may pose technical problems, such as greater linking error. Research is needed to ascertain the most effective and cost-effective ways to introduce greater unpredictability into tests.

At the other end of the measurement process, it is also critically important that the field expand the process of evaluating tests and validating inferences based on scores. Traditional types of validation remain essential, but they are insufficient because they cannot evaluate score inflation, which under high-stakes conditions is one of the most important threats to validity. Much of the validity evidence traditionally used is collected before reallocation, coaching, and score inflation can occur. Moreover, traditional validity evidence is cross-sectional, so inadequate to evaluate gains. Cross-sectional correlations between scores and other variables can remain high even in the presence of score inflation because they reflect deviations from means. This is not merely a theoretical possibility, as Koretz and Barron's (1998) evaluation of the Kentucky Instructional Results Information System (KIRIS) testing program illustrated.

Therefore, it is essential that the field make the evaluation of gains a routine part of validation. This will require putting into practice the widely ignored principle that validation should be an ongoing process.

This will be a burdensome change because it requires a second measure, generally a lower-stakes audit test, that can be used to evaluate the generalizability of score gains. In practice, the most common audit measure has been the NAEP, which has several advantages for this purpose: it represents a degree of national consensus about the content students should learn, it is a broad test because of its matrix-sampled design, and it is thought to be relatively free of inappropriate test preparation. However, NAEP has several serious limitations as an audit measure as well, including the small number of grade-by-subject combinations and the risk of motivational biases.

One approach to both the design and validation issues raised by TBA is self-auditing assessments (SAAs). SAAs would incorporate audit components into operational high-stakes assessments. The audit components would comprise items that intentionally violate the predictable patterns that facilitate inappropriate test preparation. SAAs would make it practical to audit score gains routinely, permit identification of variations in score inflation across schools, and avoid the drawbacks of separately administered audit tests. Koretz and Beguin (2009) presented a number of designs for SAAs and discussed technical issues that would arise in their use.

Finally, we need to scrutinize the entire range of activities that extend between design and validation to ascertain which are vulnerable to biases arising from high-stakes uses of tests. One activity that does appear vulnerable is linking across time. For example, a common linking approach in current TBA systems is item response theory (IRT)-based nonequivalent group with anchor test (NEAT) linking, in which items carried over from the previous year serve as the anchor. The necessary assumption underlying this approach is that the relationship between the latent trait and observed difficulties of the anchor items is constant up to a linear transformation, where the linear transformation is an artifact of the context in which the items are recalibrated. The linking process is intended to undo this linear transformation, and it is assumed that, after this process, any remaining change in performance on the anchor items represents a true change in the latent trait. However, if anchor items become easier for students at a fixed level of the latent trait—either by the direct effects of test preparation or by generalization of preparation aimed at similar items—this assumption becomes false, and score inflation is built into the resulting scale (Koretz & Barron, 1998).

To be fair, the field of measurement has devoted a great deal of effort to respond to the demands of TBA. I do not mean to disparage those efforts, but however valuable they may be for other reasons, they are not helpful for confronting the core problem of Campbell's Law. Consider, for example, the current emphasis on incorporating growth models into monitoring and accountability systems. This has generated a large amount of valuable statistical and psychometric work. However, it does nothing to alleviate the problem of Campbell's Law; it merely changes some of its characteristics. For example, the fairness of a growth-based evaluation of teachers hinges not only on the appropriateness or inappropriateness of that teacher's test preparation, but also the amount of earlier inappropriate test preparation by the teachers previously teaching the same students and the persistence of its effects over grades.

The current push for richer, more complex assessments of higher level skills, which in many ways echoes the performance-assessment movement of 20 years ago, also does not directly address the problem of Campbell's Law, even though many of its advocates have argued that it will. The core of Campbell's Law,

as one can see more easily if one looks at the forms it takes outside of educational testing, is not the level of the tasks included on an assessment. Rather, it is the severity of sampling, the predictability of the assessment, and the susceptibility of the tasks to inappropriate test preparation. A reliance on complex tasks worsens the severity of sampling by decreasing the number of tasks one can administer per unit time. The smaller number of tasks, as well as the fact that many will be memorable, also can exacerbate the second problem, that is, predictability. The complexity of the tasks involved may make it more difficult to avoid gratuitous recurrences while maintaining acceptable levels of comparability over time (which was one reason Kentucky ended the use of complex performance tasks in its KIRIS assessment during the 1990s). The complexity of the tasks also is likely to introduce greater construct-irrelevant variance, which adds opportunities for inappropriate test preparation. And we should keep in mind that evaluations of Kentucky's KIRIS testing program, which was one of the most performance-oriented state assessments during the 1990s, showed widespread and in some cases extreme score inflation (Hambleton et al., 1995; Koretz & Barron, 1998). None of this is necessarily sufficient reason to avoid a greater use of more complex assessments, but there is no basis for assuming that this will substantially reduce the problem of Campbell's Law.

As long as TBA continues to be a primary use of large-scale achievement tests, the field must make accountability, the incentives it creates, and the effects of these incentives on validity core concerns of the measurement enterprise. Currently, research on accountability-related topics, such as score inflation and effects on educational practice, is slowly growing but remains largely divorced from the core activities of the measurement field. Given current policy, this is no longer tenable.

References

Baker, G. (2002). Distortion and risk in optimal incentive contracts*. Journal of Human Resources, 37,* 728-751.

Campbell, D. T. (1979). Assessing the impact of planned social change. *Evaluation and Program Planning, 2,* 67-90*.*

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

Hambleton, R. K., Smith, Z. S., & Zhao, Y. (2006). *Curriculum-test alignment study for new MCAS tests in 2006: Grade 3 Reading, Grades 5, 6, and 8 English Language Arts, and Grades 3, 5, and 7 Mathematics.* Amherst: University of Massachusetts, Center for Educational Assessment.

Hoover, H.D., Hieronymus, A. N., Frisbie, D. A., Dunbar, S. B., Oberly, K. R., Cantor, N. K., Bray, G. B., Lewis, J. C., & Qualls-Payne, A. L. (1993). *Iowa Tests of Basic Skills interpretive guide for teachers and counselors, Forms K & L, Levels 9-14.* Chicago, IL: 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,* 761-796.

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

Jaeger, R. M. (1982). The final hurdle: Minimum competency achievement testing. In G. R. Austin & H. Garber (Eds.), *The rise and fall of national test scores* (pp. 223–246). New York, NY: Academic Press.

Klein, S. P., Hamilton, L. S., McCaffrey, D. F., & Stecher, B. M. (2000). *What do test scores in Texas tell us?* (Issue Paper No. IP-202). Santa Monica, CA: RAND.

Koretz, D. (2005). Alignment, high stakes, and the inflation of test scores. In J. Herman & E. Haertel (Eds.), *Uses and misuses of data in accountability testing. Yearbook of the National Society for the Study of Education, Vol. 104, Part 2* (pp. 99-118). Malden, MA: Blackwell Publishing.

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

Koretz, D., & Barron, S. I. (1998). *The validity of gains on the Kentucky Instructional Results Information System (KIRIS).* Santa Monica, CA: RAND.

Koretz, D., & Beguin, A. (2009). *Self-auditing assessments for educational accountability systems. Working paper of the International Project for the Study of Educational Accountability Systems.* Retrieved from http://ipea.hmdc.harvard.edu/working-paper-self-auditing-assessments-educational-accountability-systems-koretz-beguin.

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

Koretz, D., Linn, R. L., Dunbar, S. B., & Shepard, L. A. (1991, April). 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, IL.

Koretz, D., McCaffrey, D., & Hamilton, L. (2001). *Toward a framework for validating gains under high-stakes conditions* (CSE Technical Rep. No. 551). Los Angeles, CA: University of California, Center for the Study of Evaluation.

Lindquist, E. F. (1951). Preliminary considerations in objective test construction. In E. F. Linquist (Ed.), *Educational measurement* (pp. 119-158). Washington, DC: American Council on Education.

Morley, M. E., Bridgeman, B., & Lawless, R. R. (2004). *Transfer between variants of quantitative items* (GRE Board Research Rep. No. GREB-00-06R). Princeton, NJ: ETS.

Popham, W. J., Cruse, K. L., Rankin, S. C., Sandifer, P. D., & Williams, P. L. (1985). Measurement-driven instruction: It's on the road. *Phi Delta Kappan, 66*(9), 628-634.

Quincy Public Schools. (2004). *MCAS math concordance* [Powerpoint presentation]. Retrieved from http://www.quincypublicschools.com/powerpoint/MCAS/Math%20Concordance%20-%20Show.pps

Rothstein, R. (2008). *Holding accountability to account: How scholarship and experience in other fields inform exploration of performance incentives in education.* Nashville, TN: Vanderbilt University, National Center on Performance Incentives.

Rubinstein, J. (2002). *The Princeton Review: Cracking the MCAS grade 10 mathematics*. New York, NY: Random House.

Stecher, B. (2002). Consequences of large-scale, high-stakes testing on school and classroom practice. In L. Hamilton, B. M. Stecher, & S. P. Klein (Eds.), *Making sense of test-based accountability* (pp. 79-100). Santa Monica, CA: RAND.

Strauss, V. (2001, July 10). Review tests go too far, critics say. *The Washington Post*, p. A09.