Alignment, High Stakes, and the Inflation of Test Scores

CSE Report 655

Daniel Koretz
Harvard Graduate School of Education

June 2005

National Center for Research on Evaluation,
Standards, and Student Testing (CRESST)
Center for the Study of Evaluation (CSE)
Graduate School of Education & Information Studies
University of California, Los Angeles
GSE&IS Building, Box 951522
Los Angeles, CA 90095-1522
(310) 206-1532

Project 1.1 Comparative Analyses of Current Assessment and Accountability Systems, Strand 3: The Validation of Gains, Daniel Koretz, Project Director, CRESST/Harvard School of Education

The work reported herein was partially supported under the Educational Research and Development Centers Program, PR/Award Number R305B960002, as administered by the Institute for Education Sciences, U.S. Department of Education.

The findings and opinions expressed do not reflect the positions or policies of the National Center for Education Research, the Institute of Education Sciences or the U.S. Department of Education.

ALIGNMENT, HIGH STAKES, AND THE INFLATION OF TEST SCORES[1]

Daniel Koretz
CRESST/Harvard Graduate School of Education

Abstract

There are many reasons to align tests with curricular standards, but this alignment is not sufficient to protect against score inflation. This report explains the relationship between alignment and score inflation by clarifying what is meant by inappropriate test preparation. It provides a concrete, hypothetical example that illustrates a process by which scores become inflated and follows this with more complete discussion of the mechanisms of score inflation and their link to teachers' responses to high-stakes testing. Policymakers embarking on an effort to create a more effective system less prone to the drawbacks of simple test-based accountability cannot rely solely on alignment and should consider several additional steps: redesigning external tests in other ways to minimize inflation, setting attainable performance targets, relying on multiple measures, and reestablishing a role for professional judgment. Developing more effective alternatives will take us beyond what is well established and will require innovation, experimentation, and rigorous evaluation.

For several decades, some measurement experts have warned that high-stakes testing could lead to inappropriate forms of test preparation and score inflation, which we define as a gain in scores that substantially overstates the improvement in learning it implies (e.g., Koretz, 1988; Linn and Dunbar, 1990; Madaus, 1988; Shepard, 1988). This issue has been a concern in the public debate about education reform at least since the "Lake Wobegon" reports of the 1980s (Cannell, 1987; Koretz, 1988; Linn, Graue, and Sanders, 1990), which discussed the implausibly large

[1] Will also appear as Chapter 7 in: J. Herman and E. Haertel (Eds.), *Uses and misuses of data in accountability testing*. Yearbook of the National Society for the Study of Education, vol. 104, Part 1.

proportion of states and districts claiming to be above average in student achievement.

One common response to this problem has been to seek 'tests worth teaching to.' The search for such tests has led reformers in several directions over the years, but currently, many argue that tests well aligned with standards meet this criterion. If test are aligned with standards, the argument runs, they test material deemed important, and teaching to the test therefore teaches what is important. If students are being taught what is important, how can the resulting score gains be misleading?

No one can dispute that tests should measure important content, and for many (but not all) purposes, tests should be aligned with curricular goals. Thus in many cases, alignment is clearly better than the alternative, and nothing that follows here argues otherwise. Unfortunately, however, this does not imply that alignment is sufficient protection against score inflation. Inflation does not require that a test assess unimportant material, and focusing the test on important material—for example, through alignment—is not necessarily sufficient to prevent inflation.

The purpose of this report is to explain the relationship between alignment and score inflation. The first sections clarify what is meant by inappropriate test preparation and provide a concrete, hypothetical example that illustrates a process by which scores become inflated. These are followed by a more complete discussion of the mechanisms of score inflation and their link to teachers' responses to high-stakes testing. A final section discusses some implications.

Inappropriate Test Preparation and 'Tests Worth Teaching To'[2]

The problem of inappropriate test preparation has two related aspects. The first, already noted, is inflation of test scores. The second is undesirable pedagogy. This can take numerous forms, such as a boring drill and practice focused on test content or the elimination of important content not emphasized by the test. The two are obviously closely intertwined: undesirable forms of instruction are among the primary factors that cause inflation of scores.

These two aspects of inappropriate test preparation--undesirable pedagogy and score inflation—do not entirely overlap, however, and solutions to one of them will not necessarily solve the other. Instruction that creates meaningful gains in scores

[2] Note: The material in the section with the heading "Inappropriate Test Preparation and 'Tests Worth Teaching To'" is copyrighted by the author and is reprinted here with his permission.

could be undesirable in other respects. An example might be a successful but very stressful drill that is so aversive that students develop an abiding dislike of the subject being tested or of formal schooling itself. Similarly, instructional changes that are desirable in other respects may nonetheless contribute to score inflation.

It is important to be clear what is meant by 'tests worth teaching to.' Some people use the term to refer to tests that encourage meaningful improvements in student performance, while others use it to refer to those that encourage desirable changes in instruction, such as an increase in the amount of writing across the curriculum or the inclusion of larger, more complex problems in mathematics instruction. Yet others seem to use the term to refer to both of these at once, incorrectly assuming that if you accomplish one of these problems, you necessarily accomplish the second.

In this report, our focus is on validity and score inflation. We do not systematically discuss the research that has shown both positive and negative effects of test-based accountability on instruction (see, for example, Stecher, 2002). We discuss the incentives to change instruction created by test-based accountability primarily because of their link to validity and inflation.

Tests as Samples of Performance: A Concrete Example

A hypothetical, concrete example can illustrate the principles that underlie score inflation, as well as many other important issues in measurement. Assume that you confront the following challenge. You produce a journal, and you plan to hire several people newly graduated from college to work on it. You find that you have a large number of applicants and need a procedure for selecting from among them. Assume you have decided that among the factors you will consider in making your choice is the strength of the applicants' vocabularies. It is not essential for this example that vocabulary be the basis for selection because the same principles apply to other cognitive skills and knowledge. However, vocabulary provides a particularly clear and uncontroversial illustration.

The sheer size of the applicants' vocabularies would be an impediment to evaluating them. Several studies suggest that the typical graduate of four-year colleges has a vocabulary of approximately 17,000 root words (Biemiller, 2001). Therefore, the only practical option is to test the students on a small sample of words that they might know. This is precisely how vocabulary tests are constructed.

One can obtain a serviceable estimate of the relative strength of the applicants' vocabularies using a small sample of words. Let's assume 40 words for this example.

To evaluate individuals' vocabularies on the basis of 40 words, it is essential to select them carefully. Suppose that you were given three lists from which to choose words for the test. Figure 1 shows the first three words from each of these three lists. On each list, the words not shown are roughly similar in difficulty to those shown. It is obvious that one would learn nothing useful from lists A and B. List A comprises highly unusual, specialized words that few if any of the applicants are likely to know. Therefore, all of the applicants would do extremely poorly on a test comprising them, and one would learn essentially nothing about the relative strength of their vocabularies. Conversely, List B is made up of simple, extremely easy words that all college graduates would know, so they too would provide no useful information. So one would choose List C: words that are middling in difficulty, such that some students would know each word and others would not. Only the use of such words would allow one to differentiate between the applicants with stronger and weaker vocabularies. (For certain other purposes, one might select test content without regard to difficulty, but that is beyond the scope of this report.)

A	B	C
siliculose	bath	feckless
vilipend	travel	disparage
epimysium	carpet	miniscule

Figure 1. Three words from each of three hypothetical word lists

Taken this far, the example illustrates a fundamental principle that could be called the *sampling principle of testing.* In most achievement testing, one is interested in reaching conclusions about students' proficiency in a broad *domain* of achievement. In this case, the domain is vocabulary; in a more common case, it might be reading or eighth-grade mathematics. In most cases, one cannot measure these proficiencies exhaustively, or even close to exhaustively, because the domains of interest are so large. Instead, one creates a small sample of a given domain and measures students' proficiency in that small sample. One must then *generalize* from

performance in the small sample to the mostly unmeasured performance in the larger domain about which one draws conclusions.

Thus, any useful conclusion based on scores requires that one draw an inference about proficiency in the domain from proficiency in the small sample. The quality of that inference—the degree to which the inference is supported by performance on the test—is what is meant by *validity*. This is why experts in measurement say that validity is not an attribute of an inference, nor an attribute of the test itself. Even a test that provides very good support for one given inference may provide inadequate support for another, so it is misleading to talk about a 'valid test.'

Returning to the hypothetical example, suppose that someone intercepted each of your applicants on the way to your testing session and taught them all of the words on your vocabulary test. Let's examine what would happen to the validity of several of the most important inferences one might draw from scores.

In the problem as given, the primary inference is one about *relative* performance: identifying which applicants have relatively strong vocabularies. Clearly, this form of teaching to the test would render the scores useless for supporting this inference. As a result of this test preparation, all of the students would receive perfect or nearly perfect scores. The applicants would become indistinguishable in terms of this particular basis for selection.

The same problem usually arises when the user of scores is not interested in ranking but instead wants to draw inferences about the *absolute* level of performance or about the size of a gain in performance. This is particularly important at present because of the emphasis on evaluating whether students meet performance standards. In the case of our example, such inferences might be of the form "all applicants showed very strong vocabularies" or "applicants showed large gains in their vocabularies." Note that these conclusions do not rely on comparisons with the performance of other students.

The teaching to the test in this example would undermine the validity of these absolute inferences as well. It is true that many of the students would have learned additional words, and let's assume that they will remember them. We tested 40 words, and we chose words that were moderate in difficulty, so let's say that the typical applicant knew 20 of them and learned the other 20. So this intervention—teaching to the test in a very direct way—would have increased their vocabularies

by roughly a tenth of one percent. This is trivial. By the same token, it would be misleading to conclude that all students had very strong vocabularies. If one constructed another test from a similar sample of words, one would find that many of the students would do much less well on it, because in fact their vocabularies would have remained the same as before.

There is one case in which teaching the examinees the specific 40 words from the test would not cause inflation—that is, when the inference based on scores refer to the 40 tested words, not to a larger domain from which they are drawn. That is, this form of test preparation is not problematic when the sampling principle of testing does not apply. One can find examples of this (for example, one could test knowledge of the rules of English punctuation exhaustively) but they are generally not important in current achievement testing. Put differently, the important inferences based on large-scale assessment scores almost all depend on the sampling principle; that is, on generalizing from the tested sample to the domain from which items are sampled.

Thus, this intervention would have created score inflation. All students would score perfectly or nearly perfectly on the test, but this sharp increase would reflect at best a trivial improvement in vocabulary. In addition, it would have eliminated the gap between high- and low-performers, but this seemingly dramatic accomplishment would be entirely illusory. To put this in terms of the sampling principle of testing, this intervention would have made the tested sample of words unrepresentative of the domain of words about which inferences are to be drawn. By doing so, it has made score-based inferences, both inferences about levels of proficiency and the types of gains, invalid.

Note that this score inflation did not require that the tested words be unimportant. Score inflation arises when performance on the tested sample increases substantially more than proficiency in the domain about which inferences are drawn, even if the tested sample comprises important content. It is for this reason that alignment is insufficient to guard against inflation. We will return to this after showing some real examples of score inflation.

Real Examples of Score Inflation

Although the hypothetical example above is contrived, it accurately represents some of the ways in which test-based accountability can lead to misleadingly large

gains in scores. Only a handful of empirical studies have evaluated the validity of score gains under high stakes conditions, but they have usually found severe inflation.

The first study to evaluate score inflation empirically (Koretz, Linn, Dunbar, and Shepard, 1991) looked at a district-testing program in the 1980s that used commercial, off-the-shelf, multiple-choice achievement tests. By the standards of the day, the program was moderately high-stakes, although by the standards of 2004, it was actually quite low-stakes. The system entailed pressure and publicity but no concrete sanctions or rewards for test scores.

Through 1986, the district used one of the major tests of this type, indicated by the diamond for 1986 and labeled "first district test" in Figure 2. During the years before 1986, not shown in Figure 2, scores on this test had risen substantially. By 1986, the average student mathematics score of third graders in the district had reached a grade equivalent (GE) of 4.3. GEs show performance in terms of the point in schooling at which a score is typical, measured in academic years and months (with 10 months per academic year). These students were tested in the 7th month of third grade, so if they had achieved at the typical level for students at their point in schooling, they would have attained an average GE of 3.7. Instead, they achieved an average score equivalent to that reached by the typical (median) student in the third month of fourth grade nationwide. In other words, their scores made them appear to be half an academic year above average—a good showing, given that the district enrolled many low-income and minority students.

In 1987, the district switched to a competing test, and scores dropped by half an academic year. Now their average GE was 3.7, exactly typical of the nation as a whole. Over the next three years, the district's average score on the new test rapidly climbed, reaching in 1990 the same level observed on the old test in 1986. This trend is shown by the squares in Figure 2 and is labeled "second district test." This 'sawtooth' pattern—a large drop in scores when a new test is introduced, followed by rapid gains to a level similar to that before the change—is common and well documented (e.g., Linn, 2000; Linn, Graue, and Sanders, 1990).

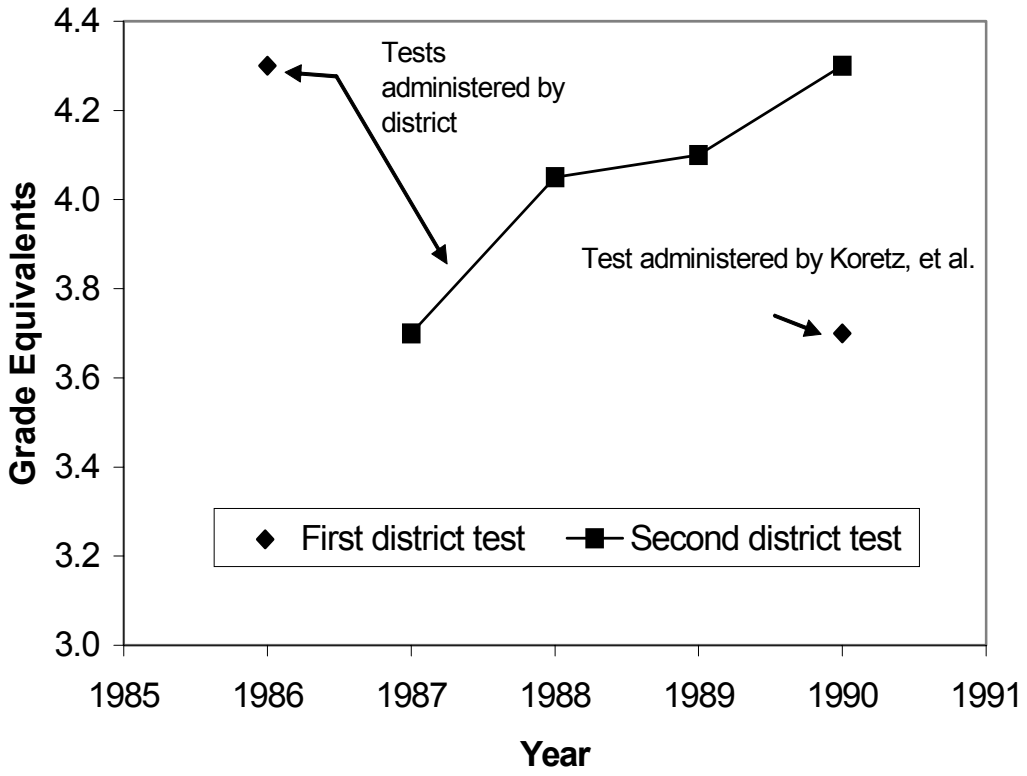

Figure 2. An example of score inflation on a moderate-stakes multiple-choice test.

In the Koretz et al. (1991) study, randomly selected classrooms were administered one of a variety of other tests in addition to the district's current test in 1990. The largest sample of classrooms was administered exactly the same test that had been used by the district through 1996. The scores of this set of classrooms (randomly selected and therefore representative of the entire district) are shown by the diamond labeled with "test administered by Koretz et al." in Figure 2. The performance of students on this test had dropped by half an academic year since the district had switched tests and was essentially identical to the performance students had shown on the second test when it was first administered in 1987.

Regardless of the test used, students scored half a year lower on a test that was unexpected than on a test for which teachers had time to prepare. It does not appear that the second test was more difficult, or that it contained new material and student performance improved as students mastered this additional material. Rather, what appears to have happened is that students and teachers *substituted* mastery of

material emphasized on the second test for mastery of material emphasized on the first test. Achievement was *transferred* among material sampled from the domain for the two tests. Unless one could argue that the material given more emphasis was much more important than that which was deemphasized, this represents score inflation.

Similar patterns have been shown by a number of other studies (e.g., Jacob, 2002; Klein, Hamilton, McCaffrey and Stecher, 2000; Koretz and Barron, 1998). Typically, gains on high-stakes tests have been 3 to 5 times as large as gains on other tests (such as the National Assessment of Educational Progress) with low (or lower) stakes, and in numerous cases, large gains on high-stakes tests have been accompanied by no gains whatsoever on lower-stakes tests. Moreover, the problem is not confined to commercial, off-the-shelf, multiple-choice tests. It has appeared as well with standards-based tests and with tests using no multiple choice items.

Score Inflation, Teacher Behavior, and Alignment[3]

Although the vocabulary example above illustrates a mechanism of score inflation, it oversimplifies the problem. A more detailed framework is needed to show the several ways in which scores can become inflated, to link these to teachers' responses to testing, and to clarify what must be done to avoid inflation, even when tests are aligned with standards.

To evaluate the validity of score gains obtained under high-stakes conditions, one needs to examine the specific sources of gains in scores and compare these to the improvements that users of scores infer from the gains. To the extent that improvements reflected in the test score signify commensurate improvements in the aspects of performance about which inferences are drawn, the inferences are valid. In the above example, the sources of gains were the specific words, fewer than 40 in total, that students learned as a result of the intervention. The intended inference, however, referenced 17,000 words, not 40, and the gains in scores would have justified an inference about improved vocabulary only if they signaled a broad increase.

To put this more formally, Koretz, McCaffrey, and Hamilton (2001) suggested thinking of both scores, and the inferences based on them, in terms of *performance*

[3] For a more detailed discussion of the framework described in this section, see Koretz, McCaffrey, and Hamilton (2001).

elements. This term refers to all the aspects of performance that underlie both performance on a test and inferences based on it. Some of these performance elements are substantive and are intended to represent the domain about which inferences are drawn. An example would be the specific words you decided to include on your vocabulary test. Others are not substantively important but may nonetheless have a substantial impact on performance. For example, the choice of item format or rubric may influence performance, even when those choices are not dictated by the test's intended inferences. Decisions of format may go far beyond the choice between multiple-choice and constructed response. For example, one might present an algebra problem verbally, algebraically, graphically, or even pictorially (see Figure 3). In some cases, the choice among these presentations may be substantive, in the sense of being tied to the intended inferences, but in many cases it will not be, and it may influence scores regardless.

Any test will assign weights—that is, relative importance—to these performance elements. For example, the more items on a test measure a given element, the more impact performance on that element will have on a student's test score. However, these weights may not be entirely intentional, and they may not entirely reflect the emphases in state standards, even when the test is well designed. One reason is that a given test item may require various knowledge and skills to solve. A clear example arose some years ago when the author and several colleagues were asked to review a pilot form of a state's new ninth-grade mathematics test. A sizable percentage of items required facility with coordinate geometry, even though this was not specifically mentioned in the state's standards. The test developers had not misread the standards. Coordinate geometry can be a good way to present topics in basic algebra that were emphasized by the state's standards. The developers had made use of this fact and, in so doing, had inadvertently given a very high weight to coordinate geometry. This sort of inadvertent emphasis can be a key to score inflation.

Use the balance scales below to answer the question below.

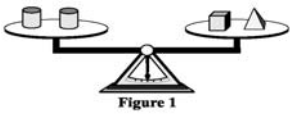

Figure 1

Figure 2

How many cylinders must be placed on the empty side of the second scale to make that scale balance?

A. 5

B. 2

C. 3

D. 4

Figure 3. An eighth-grade question from the Massachusetts MCAS test.

Similarly, users assign weights to performance elements in drawing inferences from test scores, even though these weights are typically not explicit. For example, in reading that tenth-grade scores in mathematics improved, many users will infer that this represents improvement in secondary-school material, such as algebra, but not in very basic elementary-school material, such as subtraction.

The validity of gains can be expressed in these terms. Users will infer from an increase in scores some weighted improvement in performance on a collection of elements, many of which will not be included in the test—just as most words were not included in your vocabulary test. To the extent that the increase in performance on tested elements warrants this inference, it is valid. But if the increases on tested elements do not signify an increase in performance on many elements given substantial weight by the users of scores—as was the case in the example above – then the inference about improvement is not warranted, and scores have become inflated.

Forms of Test Preparation

To understand how these considerations play out in actual practice and how they relate to alignment, we will consider a variety of ways in which teachers may prepare students for a high-stakes test. Note that in this discussion, we deliberately avoid some common ways of categorizing these responses. The term "test preparation" is often used pejoratively to refer to inappropriate forms of test preparation. However, in this discussion, the term has only its literal meaning: all of the methods (whether desirable or undesirable) that teachers use to prepare students for a test. In this discussion, we also avoid the common distinction between "teaching to the test" and "teaching the test," where the latter refers to teaching the exact items on the test and the former refers to desirable forms of test preparation. We find this usage more confusing than helpful because it mischaracterizes as a dichotomy a continuum of behaviors, many of which (other than teaching the actual items) have the potential to inflate scores.

A number of forms of test preparation may produce unambiguous, meaningful gains in performance. Teachers may work harder, for example, or they may find ways to teach more effectively. They may also teach more by providing remedial instruction outside of regular school hours. These are the sorts of responses that proponents of test-based accountability envision.

At the other extreme, teachers (or students) may cheat. For example, teachers may provide students with advance access to test items, provide inappropriate assistance during the testing session, or even change incorrect answers after the fact. Whatever the form and whatever the motivation, cheating by its very nature cannot produce meaningful gains in scores.

More interesting and more problematic is the gray area between these two extremes. The responses in this gray area might produce either meaningful gains, score inflation, or both. Therefore, these responses are the most difficult to address and warrant the most careful attention. Following Koretz, McCaffrey, and Hamilton (2001), we distinguish between three types of responses that fall into this gray zone: *reallocation, alignment,* and *coaching*.

Reallocation

Reallocation refers to shifts in instructional resources among the elements of performance. Research has shown that when scores on a test are important to

teachers, many of them will reallocate their instructional time to focus more on the material emphasized by the test (e.g., Koretz, Barron, Mitchell, & Stecher, 1996; Shepard & Dougherty, 1991). The resources that are reallocated are not necessarily limited to instructional time. They include all of the resources that parents, students, teachers, and administrators can allocate among elements of performance. Many observers believe that reallocation is among the most important factors causing the sawtooth pattern show in Figure 2.

Reallocation *transfers* achievement among elements of performance. These transfers can have a variety of effects on scores and on the validity of inferences about gains, depending on both the characteristics of the test and the nature of the performance elements that receive both increased and decreased emphasis in instruction. Clearly, if the test leads educators to stress material that is not important for the main intended inferences (as might happen if the test were poorly aligned), an increase in scores is likely to represent score inflation. However, reallocation can lead to inflation even if the test and the resulting instruction focus on important material. For example, suppose that teachers increase emphasis on tested elements but do not change their emphasis on other elements that are important for inferences but are given little or no weight by the test. In such a case, real achievement would increase, but far less than scores would, because the increase on the tested elements would overstate the increase across the whole domain. This would be analogous to our vocabulary example. More likely, given the time constraints confronting teachers, is that an increase in emphasis on the tested elements would lead to a decrease in emphasis on other, untested elements. If some of these untested elements are important for the intended inferences, then the increase in scores could mask either no change or an actual decrease in mastery of the domain the test is supposed to represent. This sort of reallocation could account for the results of the experiment shown in Figure 2 above.

When reallocation inflates scores, it does so by making the score created from the tested elements unrepresentative of the domain about which inferences are drawn. However, it does not bias a student's performance on the individual elements. Their improved performance on those particular elements is real, but just like the improved performance of students on the words included on our hypothetical vocabulary test, this does not indicate a similar improvement of mastery of the entire domain.

Alignment

Content and performance standards comprise material—performance elements, in the terminology used here—that someone (not necessarily the ultimate user of scores) has decided are important. Setting the standards indicates that this material warrants high weights in inferences which users base on test scores. Alignment gives this same material high weights in the test as well.

Alignment between tests and standards affects scores when teachers in turn align their instruction with the test. This alignment of instruction with the test is simply a special case of reallocation, and the conditions under which it will cause meaningful gains or score inflation are the same as in any other case of reallocation. That is, the issue is not merely the importance of elements that receive greater instructional emphasis as a result of alignment. It is also essential to consider the material that receives either constant or decreased emphasis. No matter how well aligned, most tests can cover only a sample of the material implied by the standards and important for inferences based on scores. Therefore, alignment of instruction with the test is likely to produce incomplete alignment of instruction with the standards, even if the test is aligned with the standards. If performance on the elements omitted from or deemphasized by the test stagnates or deteriorates while performance on the emphasized ones improves, scores will become inflated. That is, scores will increase more than actual mastery of the content standards.

Coaching

The term "coaching" is used in a variety of different ways in writings about test preparation. Here is it is used to refer to two specific, related types of test preparation, called *substantive* and *non-substantive* coaching.

Substantive coaching is an emphasis on the narrow, substantive aspects of a test that capitalizes on a particular style or emphasis of test items. The aspects of the test may have been emphasized either intentionally or unintentionally by the test designers. For example, in one study of the author's, a teacher noted that the state's test always used regular polygons in test items and suggested that teachers should focus solely on those and ignore irregular polygons. The intended inferences, however, were about polygons, not specifically regular polygons.

Substantive coaching shades into both reallocation and cheating. For example, several years ago, an article in *The Washington Post* reported on the following

example of test preparation provided by the district office in Montgomery County, Maryland, a wealthy and high-achieving district outside of Washington:

> The question on the review sheet for Montgomery County's algebra exam [provided by district officials] 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." (Strauss, 2001, p. A09)

One might reasonably argue whether this is substantive coaching, as defined here, or simple cheating, but in either case, any resulting increase in scores was almost certainly inflated.

Non-substantive coaching refers to the same process when focused on non-substantive aspects of a test, such as characteristics of distractors (incorrect answers to multiple-choice items), substantively unimportant aspects of scoring rubrics, and so on. Teaching test-taking tricks (process of elimination, plug-in, etc.) can also be seen as non-substantive coaching. In some cases—for example, when first introducing young children to the op-scan answer sheets used with multiple-choice tests—a modest amount of certain types of non-substantive coaching can increase scores and improve validity by removing irrelevant barriers to performance. In most cases, however, it either wastes time or inflates scores.

Coaching differs from reallocation and alignment in the mechanism of score inflation. Recall that reallocation and alignment inflate scores by making the tested elements unrepresentative of the domain as a whole, without biasing estimates of performance on individual elements. In contrast, coaching does bias performance on individual elements.

Conclusion

Despite its benefits, alignment is not a guarantee of validity under high-stakes conditions. Even with superb alignment, the unavoidable incompleteness of tests makes them vulnerable to the inflationary effects of reallocation, of which alignment is a special case. Moreover, alignment offers no protection against the corrupting effects of coaching.

These facts are neither an argument against alignment nor justification for throwing one's hands up in despair. Rather, they indicate that regardless of alignment, policymakers designing test-based educational accountability systems face two fundamental challenges:

- Evaluating the validity of observed score gains, and
- Tuning the system to create the right mix of incentives, thereby minimizing score inflation.

Evaluating the Validity of Score Gains

Research to date makes clear that score gains achieved under high-stakes conditions should not be accepted at face value. The same is true of apparent improvements in historic achievement gaps between groups, such as racial/ethnic groups. The sawtooth pattern widely observed when new tests are introduced—even tests without very high stakes—casts doubt on the meaningfulness of the large initial gains that often accompany the implementation of new testing programs. Investigations of score gains under high-stakes conditions, although as yet few in number, consistently show large inflation, in some cases dwarfing the meaningful, generalizable improvements in student performance. This inflation creates an illusion of overall progress and can be misleading in other ways as well. For example, variations in the amount of inflation can incorrectly suggest that some programs or schools are more effective than others.

In response to this uncertainty, policymakers should institute regular monitoring and evaluation of the validity of score gains achieved in high-stakes systems and should resist the temptation to take score increases at face value. This can be done by means of a combination of redesigned tests (for example, by deliberately adding items that are sufficiently novel to thwart coaching) and occasional larger-scale evaluations. This monitoring and evaluation is the only way to provide the public, policymakers, and educators with trustworthy information about the condition of education. While this would be a fundamental and somewhat burdensome change in practice, it would merely bring education into line with practice in numerous other areas of public policy. In many other areas—for example, the evaluation of the safety and efficacy of drugs—it is widely taken as a given that the public is owed a rigorous evaluation of policies and activities that have

potentially serious effects on its well-being. Given the power of high-stakes testing, the students, parents, and educators who are subject to it deserve the same.

Generating the Best Incentives

In many quarters, enthusiasm for test-based accountability appears to rest on a very simple model of incentives. If the system measures what is valuable (hence the importance of alignment) and rewards and punishes educators, and sometimes students, for their degree of success in producing it, students and teachers will be motivated to produce more.

It now seems clear that this model is too simple and that merely holding teachers accountable for increases in scores runs the risk of creating the wrong mix of incentives (Koretz, 2003). The existence of score inflation is one indication of this. In addition, a number of studies have documented a variety of undesirable responses to test-based accountability. (For a brief review, see Stecher, 2002.) Thus, even in a well-aligned system, policymakers still face the challenge of designing educational accountability systems that create the right mix of incentives—incentives that will maximize real gains in student performance, minimize score inflation, and generate other desirable changes in educational practice.

This is a challenge in part because of a shortage of relevant experience and research. The nation has been so confident in simple test-based accountability—more specifically, so certain that it would work as desired if we could only find a 'test worth teaching to'—that more complex, and potentially more successful, models have not been widely tried or evaluated. Thus, policymakers embarking on an effort to create a more effective system less prone to the drawbacks of simple test-based accountability must face uncertainty about how well alternatives will function in practice, and should be prepared for a period of evaluation and mid-course correction.

With that caveat, however, several steps seem potentially helpful:

Evaluating Gains

In addition to its other benefits, evaluating score gains may be one of the most practical ways to improve the incentives created by test-based accountability. By identifying particularly severe score inflation, evaluation of gains would lessen the incentives to engage in the forms of test preparation most likely to produce it.

Moreover, it might generate a more productive debate among educators and policymakers about the appropriate ways to respond to accountability.

Redesigning External Tests

Currently, the design of the tests used in accountability systems is guided by both traditional psychometric concerns, such as reliability and freedom from apparent bias, as well as the desire for alignment. The risk of undesired responses to testing might be lessened if the factors that facilitate both coaching and undesirable reallocation were also explicitly considered in designing tests. For example, developers should be alert to unnecessary recurrent patterns in content or presentation, to inadvertent overweighting of performance elements, and to omissions of elements with substantial importance to users' inferences. If teachers and test-preparation companies can find these patterns, so can test developers and the policymakers who hire them. The goal of these additional steps in design would be to change the way teachers evaluate and respond to the test. For example, rather than seizing on something that had recurred as a basis for narrowing the curriculum, teachers might wonder whether something omitted in the past might replace something else that had appeared for several years. The financial and practical costs of these changes may be appreciable, particularly in an era when the capacity of the testing industry is badly stretched. In addition, this approach poses technical issues in maintaining comparability over time. Nonetheless, this may prove to be an essential step in combating unwanted narrowing of the curriculum and score inflation.

Setting Attainable Performance Targets

Currently, the establishment of performance targets is arbitrary. States set their "Proficient" standard by whatever means they choose, and the rate of increase required of schools is then set formulaically by law. Rarely is there any consideration of research or historical evidence about the magnitude or rate of improvement that is reasonable to expect. Moreover, the rate of improvement can vary over time only within limits established by law. In some cases, this results in targets that are simply unrealistic.

Research has yet to clarify how variations in the performance targets set for schools affect the incentives faced by teachers and the resulting validity of score gains. One argument is that there is no harm in setting targets that are too high; the

reasoning is that in striving for these unreachable targets, teachers will affect smaller but nonetheless important improvements. The counterargument is that because teachers can take shortcuts and create large gains in scores without improving student performance, excessively high targets will only increase the incentives to do so. In terms of research, the jury is still out, but the accumulating evidence on teachers' responses to test-based accountability and the validity of score gains on high-stakes tests suggests that there may be serious risks in setting targets too high.

Relying on Multiple Measures

It is axiomatic in the field of measurement, if often ignored in practice, that important decisions should not be based on a single test score. The traditional reason for avoiding reliance on a single measure is the risk of incorrect decisions stemming from measurement error and the unavoidable incompleteness of any test. In the present context, however, there is an additional reason: relying on a single measure can exacerbate undesirable incentives (Koretz, 2003) and hence exacerbate score inflation.

Federal and state policies are beginning to reflect this axiom in calling for the use of multiple measures, but to date efforts to do so are generally limited, e.g., adding one or a few measures of dropout and retention rates to a system that places primary emphasis on test scores. Again reflecting a lack of experience, the field can offer only limited guidance about how best to make more ambitious and effective use of multiple measures. However, this is an area of potentially great promise and innovative efforts, coupled with rigorous evaluation, are warranted.

Reestablishing a Role for Professional Judgment

The test-based accountability systems of today are designed to be judgment-free. The systems produce a set of numbers by which anyone, including people with no knowledge of the nature or context of a given school, supposedly can judge a school to be sufficiently effective or not. Some observers attribute this design to a distrust of the educational establishment among some policymakers, but even absent such distrust, many observers would prefer to base a system on objective measures of performance than on more easily distorted and often more expensive subjective measures.

Economists, however, have long recognized that the choice between objective and subjective measures is a complex one. Derek Neal, in a recent paper that considers how accountability systems might be designed to respond to the problem of inflated test scores, noted that "straightforward incentive systems based on objective standards are often problematic because objective performance standards are often easy to game" (Neal, 2002, p. 36). The inflation of scores discussed here is just one example of this gaming. Neal argued that for this reason, professionals in fields other than education rarely face incentives based on "simple formulae tied to an objective performance standard" (Neal, 2002, p. 37). But he also noted a difficulty in relying instead on subjective measures in public employment: in the public sector, the managers responsible for the evaluations lack any financial stake in the evaluation and therefore may feel freer to bias the evaluation for inappropriate reasons.

Despite this tension, policymakers may find in the end that they have little choice but to add measures based on expert judgment back into the mix for evaluating schools. These measures may be needed not only to avoid gaming the system but also to focus attention on the many critically important aspects of educational quality that cannot be captured by standardized tests, in order both to provide a better appraisal of schools and to give teachers a better mix of incentives to improve practice. Here again, however, policymakers must be prepared for a difficult period of experimentation, evaluation, and mid-course correction.

In sum, the design of an effective test-based accountability system that minimizes score inflation while maximizing beneficial changes in instruction and increases in student achievement remains a difficult challenge. Extant research is sufficient to suggest that the current, very simple approach is unlikely to meet its proponents' goals, but developing more effective alternatives will take us beyond what is well established and will require innovation, experimentation, and rigorous evaluation. Alignment, while important for many purposes, does not solve this problem for us.

References

Biemiller, A. (2001). Teaching vocabulary: Early, direct, and sequential. *American Educator, 25*(1).

Cannell, J. J. (1987). *Nationally normed elementary achievement testing in America's public schools: How all fifty states are above the national average*. Daniels, W. V.: Friends for Education.

Jacob, B. (2002). *Accountability, incentives and behavior: the impact of high-stakes testing in the Chicago public schools*. Cambridge, MA: National Bureau of Economic Research, working paper W8968 (May).

Klein, S. P., Hamilton, L. S., McCaffrey, D. F., and Stecher, B. M. (2000). *What do test scores in Texas tell us?* Santa Monica, CA: RAND (Issue Paper IP-202; http://www.rand.org/publications/IP/IP202/).

Koretz, D.M. (1988). Arriving in Lake Wobegon: Are standardized tests exaggerating achievement and distorting instruction? *American Educator*, Summer, *12*(2): 8-15, 46-52.

Koretz, D. (2003). Using multiple measures to address perverse incentives and score inflation. *Educational Measurement: Issues and Practice 22*(2), 18-26.

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

Koretz, D., Barron, S., Mitchell, K., & Stecher, B. (1996). The perceived effects of the Kentucky Instructional Results Information System (KIRIS) (MR-792-PCT/FF). Santa Monica, CA: RAND.

Koretz, D., McCaffrey, D., and Hamilton, L. (2001). *Toward a framework for validating gains under high-stakes conditions*. CSE Technical Report 551. Los Angeles: Center for the Study of Evaluation, University of California.

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, IL.

Linn, R. L. (2000). Assessments and accountability. *Educational Researcher, 29* (2), 4-16.

Linn, R. L, and Dunbar, S. B. (1990). The Nation's report card goes home: Good news and bad about trends in achievement. *Phi Delta Kappan, 72* (2), October, 127-133.

Linn, R. L., Graue, M. E., and Sanders, N. M. (1990). Comparing state and district test results to national norms: The validity of the claims that `Everyone is above average." *Educational Measurement: Issues and Practice, 9*(3): 5-14.

Madaus, G. (1988). The influence of testing on the curriculum. In L. Tanner (Ed.), *Critical Issues in Curriculum*. Chicago: University of Chicago Press, 83-121.

Neal, D. (2002). How would vouchers change the market for education? *Journal of Economic Perspectives, 16*(4), 25-44.

Shepard, L. A. (1988, April). *The harm of measurement-driven instruction.* Paper presented at the annual meeting of the American Educational Research Association, Washington, D.C.

Shepard, L. A., & Dougherty, K. C. (1991, April). Effects of high-stakes testing on instruction. 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.

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

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