Koretz, Daniel

Test-Based Educational Accountability. Research Evidence and Implications

Zeitschrift für Pädagogik 54 (2008) 6, S. 777-790

urn:nbn:de:0111-opus-43768

ZEITSCHRIFT FÜR PÄDAGOGIK
Jahrgang 54 – Heft 6
November/Dezember 2008

Inhaltsverzeichnis

Thementeil: Systeme der Rechenschaftslegung und Schulentwicklung

Allgemeiner Teil

Besprechungen

Dokumentation

Daniel Koretz

Test-Based Educational Accountability

Research Evidence and Implications

Abstract: *In recent years, many nations, including Germany, have begun to use students' scores on achievement tests to monitor the performance of schools and educational systems. Such systems have been in place for a considerable time in a few nations, notably the United States, and numerous studies of their effects have been conducted. While these studies are limited, they are sufficient to reveal serious problems that should be confronted as new systems are put in place in other nations. Research in the U.S. has shown two related types of problems in test-based accountability (TBA) systems. Studies have revealed a mix of positive and undesirable effects on teaching and other aspects of educational practice. Research has also shown that increases in scores can become seriously inflated. That is, scores can increase by a larger amount – in some cases, a far larger amount – than actual improvements in student learning warrant. This paper summarizes studies of score inflation, describes several mechanisms that produce it, and notes implications for evaluation, testing, and the design of accountability systems.*

1. Score Inflation

In 1975, Donald Campbell wrote what has come to be called Campbell's Law: "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 1975). This corruption of indicators has been found in many different fields, including health care, social services, and environmental regulation. (For a description of many examples, see Rothstein 2008.)

In education, Campbell's Law can take many forms because a variety of measures can be used in accountability systems. For example, some systems in the United States have used dropout rates or attendance rates in addition to test scores. Most of these indicators are subject to the corruption about which Campbell warned. Inflation of test scores, however, is a particularly significant instance of Campbell's Law because of the centrality of scores in current accountability systems and their importance for the public's perception of the success or failure of the educational system.

It is useful consider two types of score inflation that that have different causes. The first category of inflation is created by changing the group of tested students – for example, by excluding students from lower-scoring groups or by reclassifying students in a way that helps to raise aggregate scores (see, for example, Figlio/Getzler 2002; Jacob 2005). These forms of gaming can substantially distort aggregate scores, such as school averages or the percentages of students reaching performance standards. However, they need not bias the scores of individual students, and they are therefore not of psychometric interest. The second category of inflation arises from actions that bias not only

aggregate scores, but also the scores of individual students. The balance of this paper considers only the latter type of inflation, which appears to be the primary cause of the most serious instances of inflation in the U.S. research literature.

The principle underlying score inflation of the latter type is that most achievement tests are only small samples from much larger domains of achievement. For example, responses to the small sample of items on a mathematics test are used as a basis for inferences about mastery of the mathematics learned over one or many years of education. In this way, a test is analogous to a political poll, in which the responses of a relatively small number of people are used to estimate the preferences of a far larger group of voters. In a poll, one samples individuals, while in a test, one samples behaviors from an individual's larger repertoire, but the fundamental principal is the same: using the sample to estimate the larger set from which it is drawn. For this reason, scores are only meaningful to the extent that they justify an inference about the larger whole. That is, the validity of inferences depends on the extent to which performance on the tested sample generalizes to the much bigger, largely untested domain of achievement.

This principle leads to the design of most studies of score inflation and some studies of educators' responses to test-based accountability. If teachers and students respond to the pressure to raise scores by focusing too much on the tested sample rather than the domain as a whole, performance on tested content will increase more than unmeasured performance on untested content, and scores will rise faster than real mastery of the domain. If this is *not* occurring and observed increases in scores do generalize to the domain, they must also generalize to other tests of the same domain – that is, to other, similar samples. However, if students and teachers are focusing too much on the specific sample tested, they will generate test-specific gains that do not generalize to the domain and that are not reflected in other tests of the same constructs. Therefore, most studies of score inflation in the U.S. have compared gains on a test used for accountability (usually called the *high-stakes* test) to gains on another test of the same domain (often called the *audit* test). In the absence of inflation, the trends in scores should be similar, although not identical. Similarly, some studies of teachers' responses to test-based accountability have investigated behaviors that entail excessive focus on the content or details of the high-stakes test.

Few detailed studies of score inflation have been carried out, in part because they are politically controversial. These few studies indicate that score inflation is common and that it can be very large. Two clear examples can be found in evaluations of the high-stakes testing program implemented in Kentucky in 1992. In this program, all schools were given numerical targets for increases in scores in numerous subjects for each two-year period. Increases substantially greater than the target earned a school cash rewards (which could not be used to augment teachers' salaries), while gains sufficiently below the target could bring sanctions. The legislature required that the new tests (called KIRIS) reflect the framework used to construct the National Assessment of Educational Progress (NAEP), which is a low-stakes, sample-based assessment. Therefore, NAEP, which was administered to representative samples of Kentucky students, was an appropriate audit for KIRIS.

Over the first two years of the KIRIS program, the average fourth-grade reading score on KIRIS increased by 0.76 standard deviation. This was a suspiciously large increase; U.S. data from tests not subject to score inflation suggest that large-scale changes rarely average more than 0.04 standard deviation per year (see, for example, Koretz 1986). During the same period, the performance of the state's fourth grade students on the NAEP reading assessment declined by a trivial 0.03 standard deviation (Hambleton et al. 1995). Over the first four years of the program, eighth-grade students' mean score on the KIRIS mathematics test increased by 0.52 standard deviation. During the same period, the state's mean score on the NAEP mathematics test increased only about one-fourth as much, 0.13 standard deviation. The increase on NAEP in the state was similar to that in the U.S. as a whole (Koretz/Barron 1998). Several other studies have found disparities similar to these (Jacob 2005, 2007; Klein et al. 2000; Koretz et al. 1991).

Research has shown not only that score inflation can be very large, but also that it is highly variable from school to school. Moreover, we still know little about factors that predict which schools' scores are most inflated. Indeed, we lack good tools for identifying variations in inflation because we rarely have a reasonable audit test that is administered regularly in all schools.

The result is that the many of the most important inferences based on scores can badly biased. In the absence of confirmatory evidence, neither the large aggregate increases in scores that often accompany test-based accountability nor relative differences in gains among schools can be trusted.

2. A Model of Test Construction

The primary theoretical response to this problem has been the work of Koretz and colleagues (Koretz/McCaffrey/Hamilton 2001; see also Koretz/Hamilton 2006),who suggest a new model of test construction to better explain score inflation and to guide investigations of educators' responses to test-based accountability.

The model begins by considering all elements of performance that are given substantial emphasis either by the test or in the inferences users base on test scores. These performance elements may be finer-grained than individual test items. Many are substantive, that is, related to the intended inference. However, some performance elements are non-substantive, unrelated to the intended inference. The authors of tests must make decisions about the format of test items, the type of representations used (for example, whether an algebraic problem is presented graphically or pictorially), the rubric used to score the item, and so on. Although some of these choices will be related to the intended inferences, many will not be. Elements that are unrelated to the inference may nonetheless have an appreciable impact on students' performance on the test, particularly if training focuses on them.

Each element that is tested has an *effective test weight*, which is the influence performance on that element has on the test score. This weight, which can reflect both content emphasis and technical factors such as item discrimination, is the partial derivative

of scores with respect to the element. That is, where π_i is a performance element and θ is the score, the test weight of the element is:

$$\lambda_i = \frac{\partial \theta}{\partial \pi_i}.$$

Koretz et al. (2001) refer to the construct or domain about which inferences are drawn as the *target of inference*. The target, like the test, comprises a set of performance elements, which are given *inference weights* that reflect the importance of the elements to the inference based on scores. However, the test and target differ in important ways. The target, unlike the test, is often only vaguely and incompletely defined. For example, even though the test specifications for a mathematics test are generally quite specific, the inferences users base on scores are often simple and vague. The target is larger and generally gives substantial weight to elements that are not tested. Elements included in the test may have test weights that differ from their inference weights.

The test and its target of inference can be viewed as two vectors of weighted performance elements, as in Figure 1. Elements *1* through *j* in the top block of Figure 1 are included in both the test and the target. Therefore, their test weights λ_i and their corresponding inference weights ω_i are non-zero, but they may not be proportional. Apart from possible differences in weighting, this block of performance items is not problematic.

Test	Target
$\lambda_1\pi_1$	$\omega_1\pi_1$
$\lambda_2\pi_2$	$\omega_2\pi_2$
$\vdots$	$\vdots$
$\lambda_j\pi_j$	$\omega_j\pi_j$
$0\cdot\pi_{j+1}$	$\omega_{j+1}\pi_{j+1}$
$0\cdot\pi_{j+2}$	$\omega_{j+2}\pi_{j+2}$
$\vdots$	$\vdots$
$0\cdot\pi_k$	$\omega_k\pi_k$
$\lambda_{k+1}\pi_{k+1}$	$0\cdot\pi_{k+1}$
$\lambda_{k+2}\pi_{k+2}$	$0\cdot\pi_{k+2}$
$\vdots$	$\vdots$
$\lambda_n\pi_n$	$0\cdot\pi_n$

Fig. 1: A test and a target of inference as weighted vectors of performance elements

The second block, subscripts *j+1* to *k*, comprises performance items that are important for the inference but are omitted from the test and therefore have test weights of zero.

When the inference is about a broad domain of achievement, this block of untested elements is typically very large.

The final block of elements, beginning with *k+1*, comprises elements that are included in the test but are not important to the inference and therefore have inference weights of zero. These elements reflect decisions by the test authors that are not substantively important. This focus on substantively unimportant performance elements is an important difference between the framework proposed by Koretz et al. (2001) and the otherwise somewhat similar model of validity subsequently suggested by Kane (2006). As the examples below indicate, these non-substantive performance elements may be an important source of score inflation.

3. Educator's Responses to Testing

Koretz and colleagues (Koretz/McCaffrey/Hamilton 2001; Koretz/Hamilton 2006) suggest grouping educators' responses to test-based accountability into seven categories. On the positive side, educators may spend more time teaching, work harder, or find more effective methods of teaching. Within limits, all of these responses are likely to produce meaningful gains in scores, that is, increases in scores that mirror real improvements in student learning. At the other extreme, both educators and students may simply cheat, which can produce only score inflation. (For an ever-growing account of cheating incidents in the U.S., see http://www.caveon.com/citn/index.php.) More interesting are the three remaining categories of responses to testing, labeled *reallocation, alignment*, and *coaching*, that can produce either score inflation or meaningful gains.

Reallocation refers to shifting instructional resources to better match the sampling of content by the test. In the terminology above, this is shifting instructional resources among substantive performance elements so that the emphasis in instruction better matches the effective test weights. Perhaps the most important of these resources, and the most often studied, is instructional time (see, for example, Stecher 2002), but other resources may be reallocated also, such as homework assignments. Other resources that are only indirectly under the control of educators may be reallocated as well, such as students' effort and parental pressure.

Reallocation of instructional resources, if effective, reallocates achievement. Whether it increases achievement – to be more precise, whether it increases the achievement that the test score is designed to estimate – depends on the nature of the reallocation.

Reallocation may occur between subjects – for example, time may be taken away from untested subjects and added to tested ones (e.g., Stecher/Barron 1999). This is particularly likely in the elementary grades in which individual teachers teach multiple subjects. Reallocation between subjects is an important issue for education policy, but it is usually not relevant to score inflation. For example, if schools in the U.S. take time away from subjects that are not tested for accountability (such as history in most states) to add time to subjects such as mathematics that are tested, this will presumably reduce

learning in history, but it does not undermine the validity of inferences about mathematics achievement based on test scores.

Reallocation may also occur within subjects, as teachers shift resources from material that receives little or no emphasis on the test to content in the same subject that is likely to be emphasized by the test. This form of reallocation, which has been found in numerous studies in the U.S. (for example, Koretz et al. 1996), may create either meaningful gains in scores or inflation.

Whether within-subject reallocation causes score inflation depends on both the content that receives more emphasis and that which receives less. The validity of an inference about gains depends on the extent to which improvements in performance on the tested vector of performance elements supports an inference about improvements on the entire vector relevant to the inference, including the many performance elements that are untested. Therefore, inflation arises if reallocation increases performance on tested performance elements substantially more than it improves mastery of untested elements that are important for the inference. If teachers take away time from untested performance elements that are important for the inference and students learn less about them as a result, that deterioration will not appear in scores. Similarly, if teachers focus on elements have substantially larger test weights than inference weights, gains will be exaggerated.

Note that even when it causes score inflation, reallocation need not bias estimates of performance on specific elements. Rather, the bias arises from aggregating the tested performance elements into a composite score. When inflation occurs, changes in that composite do not represent proportionate changes on the target.

In the United States, a core element of accountability policies is "alignment." States accepting funds under the federal No Child Left Behind Act – currently, all states – are required to establish "content standards" that describe what students should know and be able to do, and "performance standards" that indicate the level of proficiency students are expected to show with respect to the content standards. States must then create tests that are aligned with these content standards, and the use of these tests is intended to induce instruction that is similarly aligned with standards.

Advocates of current accountability programs often insist that this alignment protects against score inflation. Their argument is that if tests are aligned with standards, they are measuring important content, and if teachers focus instruction on the test, they are therefore necessarily focusing instruction on important material. They then argue that focusing on important material cannot produce inflation. This argument is incorrect. Clearly, some degree of alignment is good; one would not want tests that encouraged a focus on unimportant material. However, alignment is just a special case of reallocation, and the same principle applies: inflation depends not only on the material that gains additional emphasis, but also on the material that loses emphasis. Because tests are small samples of behavior, they typically constitute only a sample from the standards. Therefore, teachers can align their instruction with the test while still deemphasizing material that is important for the inference about students' mastery of the stan-

dards. In other words, focusing on important material, while desirable, is not sufficient to prevent score inflation.

The term *coaching* is used loosely and inconsistently in writing about testing. Here we follow the specific usage suggested by Koretz et al. (2001), who used the term to describe various forms of test preparation that focus on details of the test. Some of these details might be called substantive but unimportant: they are in some way related to the target, but they are not important for the inference. For example, suppose that a mathematics assessment is supposed to measure students' understanding of basic principles of plane geometry. In designing this test, one would need to make many decisions about how plane geometry should be represented. Some of these decisions may reflect the target of inference, but many will not. For example, should irregular polygons be presented, or only regular ones? What is the maximum number of sides in the figures that will be presented? If estimation is included in the inference, will it be extended to the plane geometry items, or will these require only calculation of areas, and perimeters? If triangles are presented, which attributes of triangles are tested? Such decisions are unavoidable, but some are not relevant to the inference based on test scores.

These decisions are often repeated from one instance of testing to the next. There can be many reasons for such repetition of details, but some recurrences are not important for the inference. For example, some repetition occurs simply because developers have limited resources or because they do not understand the unintended consequences of the repetition.

Under low-stakes conditions, repetition of details may not be problematic, and traditional psychometric theory does not pay it much attention. However, repetition becomes very important under high-stakes conditions because it offers opportunities for coaching. For example, in one study, an American secondary school teacher asked one of my students, "Why would I teach irregular polygons?" She was not questioning the importance of irregular polygons. She asked the question rhetorically because her state's test virtually never includes them. Examination of American test-preparation materials shows many examples of focusing on recurrent details of this sort. Koretz et al. (2001) referred to this as *substantive coaching*.

Nonsubstantive coaching is similar but focuses on focuses on non-substantive performance elements – elements that are unrelated to the target – which recur from one instance of testing to another. Item format, when it is not substantively important for the inference, can provide opportunities for non-substantive coaching. A common example is advice to capitalize on the multiple-choice format by using the process of elimination. For example, one book published by the Princeton Review, a major U.S. vendor of test-preparation materials, advises students that "Sometimes the best way to solve a problem is to figure out what the…wrong answers are and eliminate them…. It's often easier to identify the *wrong* answers than to find the *correct* one" (Rubinstein 2002, p. 15). As this example indicates, many test-taking tricks fall under the rubric of coaching in Koretz et al.'s classification.

Coaching can be focused on other aspects of items, such as the choice of graphical or pictorial representation. It may also make use of regularities in scoring rubrics, what

Stecher and Mitchell (1995) dubbed "teaching to the rubric" – as described by one teacher, "What's in the rubrics gets done, and what isn't doesn't." Stecher and Mitchell noted that this "May cause teachers to neglect important…skills not addressed by the rubrics and neglect tasks not well aligned to [them]" (1995, p. ix).

Coaching can inflate scores because it focuses attention on tested details at the cost of other content, representations, and task demands. For example, consider the method of process of elimination. This capitalizes on the use of the multiple-choice format, but that choice of format is generally irrelevant to the inference. Parents and employers would expect that students learn mathematical skills that they can apply in the real world, which rarely provides mathematics problems in that format. Any gains produced by this strategy will evaporate if the format is changed to constructed-response.

Although in practice, the distinction between reallocation and coaching is not always entirely clear, the two approaches differ in terms of the mechanism of score inflation. As noted, reallocation can leave estimates of performance for individual elements unbiased even while inflating total scores. In contrast, coaching can inflate estimates of performance on individual elements. For example, suppose that one substantive performance element is 'factoring quadratic equations,' and students are taught to rely on process of elimination to solve items testing that element. To the extent that the coaching is effective, it will improve students' performance on these items more than their ability to factor quadratic equations warrants.

The following examples illustrate the distinction between substantive and non-substantive coaching. Both are taken from the Princeton Review's test preparation materials for the 10th grade Massachusetts MCAS mathematics assessment. Consider the test item show in Figure 2, which is taken from an MCAS test. The Princeton Review materials note that items of this sort, as well as two other types of items involving triangles, appear frequently in MCAS assessments. They labeled these types of items "special triangle rules," of which this is the first. They noted: "One triangle rule that is often tested on the MCAS exam is the *third side* rule. The rule is: The sum of every to sides of a triangle must be greater than the third side" (Rubinstein 2002, p. 52). They then explain the reason why this is true.

This is an example of substantive coaching, taking advantage of an unimportant substantive detail that recurs from one instance of testing to the next. There is nothing wrong with the item if it is used infrequently. It is just one performance element sampled from among the many elements in the area of plane geometry that one might consider reasonable to test at this point in students' schooling. However, it is not so important an element that it should be included much of the time in the small set of elements sampled – in the case of this test, a total of 42 items that count toward a students' score. To put this in terms of the framework above, this very fine-grained performance element – knowledge of the third-side rule – is probably to small to have its own inference weight. Rather, it is most likely just one of many small aspects of performance that contribute to a larger whole that does have a substantial weight, perhaps "knowledge of the properties of basic plane figures." If the third-side rule were important enough to have its own inference weight, than this type of test preparation would properly be consid-

ered reallocation rather than coaching. In either case, however, preparing students for this particular detail that happens to recur will create an increase in scores that will not generalize to other, similar tests that happen to sample content differently.

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

Fig. 2: An item from a 10th grade MCAS mathematics assessment

For an example of non-substantive coaching, consider another example from the same test-preparation book. The book includes a section devoted to the Pythagorean theorem. It begins:

Whenever you have a right triangle – a triangle with a 90-degree angle – you can use the Pythagorean theorem. The theorem says that 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) (Rubinstein 2002, p. 56).

So far, this is simply a description of the theorem, which could be part of a good lesson about it. However, the text then goes on to note:

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 (Rubinstein 2002, p. 56).

It is this latter point that is emphasized; it is repeated in a prominent box in the margin of the text, under the phrase "popular Pythagorean ratios."

Where are these ratios "most common," and with whom are these ratios "popular?" They are not particularly common in the real world. In real life out of school, the ratios can be anything at all, as long as they conform to the rule that $c^2=a^2+b^2$. They are nonetheless popular among test authors for a non-substantive reason: calculating non-integer square roots is difficult, and many students who understand the Pythagorean theorem would nonetheless answer an item about it incorrectly if they were required to calculate a non-integer solution. If test authors intend to measure knowledge of the theorem while avoiding this bias, they have only two choices: use these simple ratios, or make the item one on which students can use a calculator. If the authors choose the first option, then they have inadvertently created this opportunity for non-substantive coaching.

4. Accountability and the Effects of Sampling

The sampling required for constructing tests of broad achievement domains has been a central concern in psychometrics for many decades. However, the traditional consideration of sampling tacitly assumes "low-stakes" conditions, that is, little incentive for teachers or students to focus on the tested sample rather than the target of inference. Because of behavioral responses to testing, sampling takes on a different, and greater, importance when tests are used for accountability.

Under low-stakes conditions, the adequacy and representativeness of the sample of tested performance elements is an essential precondition for valid inference. If the tested sample meets this condition, if there are no efforts to prepare examinees for the specific tested sample, and if the domain is unidimensional, then the consequence of sampling of elements is merely unreliability. If one drew different samples to construct parallel test forms, the performance of examinees would fluctuate from one instance to the next, but that inconsistency would be simple measurement error, not bias. In the classical test theory model, an observed score is thus simply a true score plus random error arising from the sampling of content or random variations in examinee behavior over time, $X = \tau + \varepsilon$. Modern test theory, both generalizability theory and item response theory, require more elaborate specifications of error but retain the notion that inconsistencies arising from the sampling of performance elements are merely measurement error.

If the test and target are modestly multidimensional, as most achievement tests are, sampling has an additional effect: one tested sample may favor one group over another because of differential matches to the groups' curricula. For example, the rankings of nations in the TIMSS mathematics assessment could be modified by changing the weights assigned to the five tested content areas because different nations emphasize different material within the domain of mathematics. However, these additional affects of sampling are usually minor because when accountability is not an issue, performance is usually very highly correlated across the subdomains of the target.

In contrast, when teachers or students are held accountable for scores, the effects of sampling can be much more serious and can include bias – score inflation – as well as measurement error. The reason is that accountability creates incentives to focus on the tested sample rather than the domain as a whole. If educators and students respond to these incentives in the ways describe above, mastery of the tested sample becomes over time less representative of mastery of the domain from which the sample is drawn. That is, the correlation between performance on tested and untested elements weakens, and scores become inflated.

5. Implications

The problem of score inflation described here is not specific to educational accountability or to the American context. It is a specific instance of the problem of Campbell's Law

that has been found in many other fields, such as health care (e.g., Bevan/Hood 2006; Dranove/Kessler/McClellan/Satterthwaite 2003) and social services (e.g., Heckman/Heinrich/Smith 2002). Economists have warned that holding people accountable using incomplete measures of performance yields a variety of distortions in behavior (e.g., Baker 2002; Smith 1995). The empirical research and theoretical work in the U.S. described here has begun to clarify the forms this problem takes in the case of test-based educational accountability and the specific mechanisms that underlie it.

Therefore, the problem of score inflation can be expected to arise in other nations as well, and it has important implications for program evaluation, testing, and the design of accountability systems in education.

The implications for evaluation are obvious and challenging: scores on the tests used for accountability, taken by themselves, cannot be considered a dependable outcome variable for evaluating teachers, schools, educational systems, or specific educational programs. Test-based accountability systems are often considered self-evaluating: if scores on the test used for accountability increase, the system is assumed to be working. Given that scores on these tests may rise dramatically even when actual student learning increases either far less or not at all, this assumption is clearly unwarranted. Moreover, it is not only overall performance that may be misestimated. Because inflation is highly variable and remains largely unpredicted, even estimates of *relative* performance can be badly biased, which precludes identifying with confidence programs that warrant rewards, sanctions, or imitation.

Therefore, to be confident that student learning has improved under a test-based accountability system, one needs additional evidence to confirm or disconfirm performance on the test used for accountability. The most straightforward evidence is obtained from an audit test, but it is often the case that none is available, and in such cases, supplementary testing will be required. Other concurrent indicators of performance may be useful, as may indicators of later performance, such as performance in postsecondary education. One of the challenges facing those designing accountability systems in Europe will be deciding which additional measures can be employed for this purpose.

The use of testing systems for accountability also poses challenges for the design of tests and the operation of testing systems. Although accountability represents a fundamental change in the uses of large-scale achievement testing, the field of measurement has changed relatively little in response. The challenges posed by accountability affect the entire testing enterprise, from design to validation. Because accountability creates incentives to focus instruction on the particulars of the tested sample, the predictability of both substantive and non-substantive performance elements has become a serious problem both for measurement and for the incentives testing creates for educators and students. Despite the technical and financial difficulties this will entail, researchers need to explore the feasibility and impact of reducing this predictability in the design of tests. The validation of the results of testing systems must also change. Currently, validation focuses primarily on analysis of the initial representativeness of the sampled material and on cross-sectional analyses of scores. While still essential, neither of these types of

evidence can evaluate the validity of gains over time, which is one of the most important inferences based on scores in test-based accountability systems.

Score inflation has additional, important implications for the design of educational accountability systems. Economists working on incentive systems have often pointed out that the presence of distortions is not in itself reason to avoid implementing an accountability system. Some degree of distortion is inevitable when a complex role, such as that of a teacher, is reduced to a set of performance indicators, and the presence of distortion does not in itself indicate that the accountability system has failed. Accountability may increase overall performance even if there is considerable distortion.

Nonetheless, the evidence from the research conducted to date suggests that test-based accountability as it has been implemented in the U.S. – that is, simply holding educators accountable for scores on one or several tests, while giving little or no weight to other indicators of performance and avoiding human judgment altogether – is unlikely to be sufficient. The severity of the score inflation that can arise from this form of accountability results in untrustworthy and often severely distorted views of improvement, and the variability of inflation precludes identifying with confidence relatively effective or ineffective schools.

Yet while the research is sufficient to indicate that this simple approach is problematic, it is does not yet provide clear guidance for the design of better accountability programs. Both economic and psychometric theory and research on accountability in other areas suggest alternatives that may be more productive, but we have not yet conducted the rigorous research needed to test their effects in educational accountability systems. Therefore, there is a pressing need for experimentation with new approaches to educational accountability.

Several areas appear particularly important, in addition to experimentation with new designs of the tests themselves. One potentially important area is using multiple objective measures to lessen the incentive to focus inappropriately on the content of the test. It is axiomatic in educational measurement that one should rely on multiple measures, but we do not yet have empirical evidence of the effects on educators' behavior and student learning. We need to explore the effects of adding measures of variables other than student performance. We need to explore the effects of various ways of using performance data, such as the choice of summary statistics and performance targets. We need to explore the practicality and impact of "dynamic" accountability systems that are modified over time in response to undesired effects on behavior. And perhaps most important, we need to investigate the effects of various approaches to combining objective data from tests with subjective judgments by headmasters, inspectors, or others. This is a large and ambitious agenda, but the severity of the problems evidenced by the simple systems now in place indicates its importance. The implementation of new accountability systems in Europe provides an invaluable but transient opportunity to undertake this agenda.

References

Baker, G. (2002): Distortion and risk in optimal incentive contracts. In: Journal of Human Resources 38, number 4, pp. 728–751.

Bevan, G./Hood, C. (2006): What's measured is what matters: targets and gaming in the English public health care system. In: Public Administration 84, pp. 517–538.

Campbell, D.T. (1975): Assessing the impact of planned social change. In: Lyons, G.M. (Ed.): Social Research and Public Policies: The Dartmouth/OECD Conference, pp. 3–45. Reprinted in Evaluation and Program Planning (1979) 2, pp. 67–90.

Dranove, D./Kessler, D./McClellan, M./Satterthwaite, M. (2003): Is more information better? The effects of "report cards" on health care providers. In: The Journal of Political Economy 111, number 3, pp. 555–588.

Figlio, D./Getzler, L.S. (2002): Accountability, ability and disability: Gaming the system. (Working Paper No. 9307). Cambridge, MA: National Bureau of Economic Research. (http://www.nber.org/papers/w9307).

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

Heckman, J.J./Heinrich, C./Smith, J. (2002): The performance of performance standards. In: Journal of Human Resources 38, number 4, pp. 778–881.

Jacob, B. (2005): Accountability, incentives and behavior: the impact of high-stakes testing in the Chicago Public Schools. In: Journal of Public Economics 89, pp. 761–796.

Jacob, B. (2007): Test-based accountability and student achievement: An investigation of differential performance on NAEP and state assessments. (Working paper No. 12817). Cambridge, MA: National Bureau of Economic Research.

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

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

Koretz, D. (1986): Trends in Educational Achievement. Washington, D.C.: Congressional budget Office, April.

Koretz, D./Barron, S.I. (1998): The Validity of Gains on the Kentucky Instructional Results Information System (KIRIS). MR-1014-EDU. Santa Monica, CA: 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./Hamilton, L.S. (2006): Testing for accountability in K-12. In: Brennan, R. L. (Ed.): Educational measurement, 4th ed. American Council on Education/ Praeger. Westport, CT: pp. 531–578.

Koretz, D./Linn, R.L./Dunbar, S.B./Shepard, L.A. (1991): The effects of high-stakes testing: Preliminary evidence about generalization across tests. In: Linn, R.L. (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 in Chicago, IL. April.

Koretz, D./McCaffrey, D./Hamilton, L. (2001): Toward a Framework for Validating Gains Under High-Stakes Conditions. (CSE Technical Report 551). Los Angeles, CA: Center for the Study of Evaluation, University of California.

Rothstein, R. (2008): Holding Accountability to Account: How Scholarship and Experience in Other Fields Inform Exploration of Performance Incentives in Education. Nashville, TN: Na-

tional Center on Performance Incentives. (http://www.performanceincentives.org/data/files/directory/ConferencePapersNews/Rothstein.pdf).

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

Smith, P. (1995): On the unintended consequences of publishing performance data in the public sector. In: International Journal of Public Administration 18, number 2 and 3, pp. 277–310.

Stecher, B. (2002): Consequences of large-scale, high-stakes testing on school and classroom practice. In: Hamilton, L. et al (Eds.): Test-based Accountability: A Guide for Practitioners and Policymakers. Santa Monica, CA: RAND, pp. 79–100.

Stecher, B./Barron, S.I. (1999): Quadrennial Milepost Accountability Testing in Kentucky. (CSE Technical Report 505). Los Angeles, CA: National Center for Research on Evaluation, Standards, and Student Testing.

Stecher, B.M./Mitchell, K.J. (1995): Portfolio Driven Reform: Vermont Teachers' Understanding of Mathematical Problem Solving (CSE Technical Report 400). Los Angeles, CA: University of California Center for Research on Evaluation, Standards, and Student Testing.

Author's Address:
Prof. Daniel Koretz, Harvard Graduate School of Education, 415 Gutman Library, 6 Appian Way, Cambridge, MA 02138