The Misunderstanding and Use of Data from Educational Tests

Daniel Koretz
Harvard Graduate School of Education

Jennifer L. Jennings
New York University/Harvard University

Prepared for The Process of Data Use
A meeting at the Spencer Foundation
Chicago, IL, February 11, 2010

Revised:
November 21, 2010
In the last decade, there has been an explosion of interest in “data-driven decision making” as a lever of educational improvement. Education sections of bookstores are now populated by “how-to” books like *Driven by Data: A Practical Guide to Improve Instruction* (Bambrick-Santoyo, 2010) and *Data-wise: A Step-by-Step Guide to Using Assessment Results to Improve Teaching and Learning* (Boudett, City, & Murnane, 2005). Data, in these publications, refers to the results of large-scale achievement tests, both formative and summative. Test-based accountability is currently the cornerstone of U.S. education policy, and the drive to improve test scores dominates practice in a great many American schools. When educators speak about “data-driven instruction,” they usually mean “test-score-driven instruction.” At present, no other use of data is nearly as important to educational policy and practice.

We find that research on the use of test score data is limited, and research investigating the understanding of tests and score data is meager. In this paper, we review the existing literature in this area, and describe a research agenda to advance this field of study. Because of the sparse research literature, we rely on experience and anecdote in parts of this paper, with the premise that these conclusions should be supplanted over time by findings from systematic research. We focus on three issues that are especially relevant to test-based data and about which research is currently sparse:

- How do the types of data made available for use affect policymakers’ and educators’ understanding of data? Educators and policymakers can only use data and research findings that are available to them, and there are substantial constraints placed on the availability of data from current testing programs and,
even more important, limitations on the testing-related research and evaluation that policymakers typically permit.

- What are the common errors made by policymakers and educators in interpreting test score data? Data and research findings that are misunderstood cannot be used appropriately to improve practice; we review common misunderstandings of test-based data to illustrate this point.

- How do high-stakes testing and the availability of test-based data affect administrator and teacher practice? Fusarelli (2008) observed that educators often use research in ways very different from policymakers, which is to be expected, given that testing has become a tool by which policymakers control educators.

In the final section of the paper, we suggest directions for research that would address these issues.

**How do the types of data made available for use affect policymakers’ and educators’ understanding of data?**

The test-based data available to educators are seriously limited. This stems from both the ways in which data from testing programs are reported and from the constraints that many education policymakers impose on testing-related research and evaluation.

Aggregate data from state testing programs are now readily available, but in forms that limit their use by educators, researchers, and policymakers other than those with direct control over test data. First, in most cases, the publicly available data are limited to school-level aggregates of total scores. Item-level and student-level data are rarely available, although some states, such as Massachusetts, have made these data public in limited forms. Second, in most instances, aggregate data are made public only
in the form of percents above performance standards (“percents above cuts,” or PACs), not as scale scores—that is, numerical scores that show the entire distribution of performance. PACs are a poor way to report performance, for several reasons. They obscure a great deal of information because changes in performance within one performance level—each of which on average will include one-fourth of the entire distribution of performance and will often include far more—are not measured. PACs exaggerate the importance of other information, because even very small changes in performance that push a student just over or under a cut score has the same impact on the measure as many changes that are many times as large. In addition, reporting in terms of PACs necessarily distorts relative trends across groups that start at different levels of performance (e.g. Koretz & Hamilton, 2006). For example, it will almost always provide misleading answers to questions such as “is the achievement gap between racial/ethnic groups shrinking?” Reporting in terms of PACs also greatly complicates comparison of trends across tests, which makes it difficult to evaluate the validity of the score gains that are the focus on both public attention and a great deal of research.

Finally, in most cases, the available data are limited to scores on the test used for accountability. Two decades of research have shown that these scores can be severely inflated by inappropriate test preparation (e.g., Jacob, 2005, 2007; Klein et al., 2000; Koretz & Barron, 1998; Koretz, Linn, Dunbar, & Shepard, 1991). Score inflation refers to increases in scores substantially greater than the real improvement in students’ skills in the tested area. Inflated scores do not provide a valid measure of students’ knowledge and skills in the larger knowledge domain that is measured. Worse for some purposes, this inflation can be highly variable because schools face different incentives to inflate scores.
As a result, the relative performance of schools is difficult to interpret in the presence of score inflation. At this point, we know very little about the factors that may predict higher levels of inflation—for example, characteristics of tests, accountability systems, students, or schools.

As a result of these limitations, it is impossible to answer some of the most pressing questions without research that goes beyond publicly available data. For example, data from state’s own testing programs cannot identify which states and localities have the most effective programs, and we cannot with any confidence identify which schools within a state or district are making the most progress. If policymakers and educators turn to research to answer these questions, the research that uses scores on high-stakes tests as the outcome will often be seriously misleading. We cannot obtain an accurate view of trends in important achievement gaps, and the frequent reports of changes in achievement gaps found in press releases and in the press are usually wrong. The sole exceptions to this bleak picture are survey-based assessments that are not subject to accountability pressures, such as the National Assessment of Educational Progress (NAEP), and the available data of this sort are limited in frequency, limited in subject and grade coverage, and too sparse to provide information below the level of states.

Unfortunately, it is often exceedingly difficult to obtain the permission and access needed to carry out testing-related research in the public education sector. This is particularly so if the research holds out the possibility of politically inconvenient findings, which virtually all evaluations in this area do. In our experience, very few state or district superintendents or commissioners consider it an obligation to provide the
public or the field with open and impartial research. Data are considered proprietary—a position that the restrictions imposed by the federal Family Educational Rights and Privacy Act (FERPA) have made easier to maintain publicly. Access is usually provided only for research which is not seen as unduly threatening to the leaders’ immediate political agendas. The fact that this last consideration is often openly discussed underscores the lack of a culture of public accountability.

This unwillingness to countenance honest but potentially threatening research garners very little discussion, but in this respect, education is an anomaly. In many areas of public policy, such as drug safety or vehicle safety, there is an expectation that the public is owed honest and impartial evaluation and research. For example, imagine what would have happed if the CEO of Merck had responded to reports of side-effects side-effects from Vioxx by saying that allowing access to data was “not our priority at present,” which is a not infrequent response to data requests made to districts or states. In public education, there is no expectation that the public has a right to honest evaluation, and data are seen as the policymakers’ proprietary sandbox, to which they can grant access when it happens to serve their political needs.

**What are the common errors made by policymakers and educators in interpreting test score data?**

Systematic research exploring educators’ understanding of both the principles of testing and appropriate interpretation of test-based data is meager. A number of studies have investigated teachers’ views of the appropriateness of various forms of test preparation, and all found that a sizeable proportion of teachers considered acceptable various forms of test preparation that experts consider inappropriate and likely to bias
scores (Johns & Davis, 1991; Kher-Durlabhji & Lacina-Gifford, 1992; Mehrens & Kaminski, 1989; Moore, 1993; Urdan & Parris, 1994). These findings imply a lack of understanding of some of the core principles of testing, such as the sampling principle explained below, but these studies did not directly measure this understanding. Moreover, these studies are old, which makes their applicability to the current situation problematic. The pressure to raise scores has grown enormously over the intervening years, and the pre- and inservice training of teachers has changed as well. For example, some of the test-preparation approaches authors of the older studies listed as inappropriate, such as using old test items in instruction (Urdan & Parris, 1994) and using commercial test preparation materials (Moore, 1993) are widely and openly encouraged now by states and districts, and some of the other forms of preparation noted as inappropriate (e.g., giving test preparation work on weekends; Moore, 1993) seem positively quaint by today’s standards.

Although current, systematic information is lacking, our experience is that that the level of understanding of test data among both educators and education policymakers is in many cases abysmally low. It seems that the current policy context, in which testing is central and the pressure to raise scores is intense, has led to improvements in some respects but deterioration in others. Here we list some of the most essential issues. These include concepts that are the traditional concern of psychometrics as well as some additional issues that have arisen because of the pressure to raise scores from test-based accountability.

**Tests as samples of performance.** This is perhaps the most fundamental principle of achievement testing. The domains of achievement testing about which
policymakers are concerned, and for which educators are held accountable, are large. Tests are necessarily small samples drawn from them, and scores are only useful as an estimate of mastery of these larger domains. This principle underlies a wide variety of issues that arise in measurement, some widely acknowledged and others not. These include measurement error and reliability, the usually modest but occasionally sizable differences in results among tests not used for accountability, and the problem of score inflation.

It is helpful to represent the sampling entailed in building a test as a series of steps that successively narrow what is tested. One begins by defining a domain that is to be assessed, such as “eighth-grade mathematics” or “cumulative mastery of mathematics through grade 8” (Box 1 in Figure 1). The delineation of these domains is, of course, a matter of policy and judgment. Much of the extant research on the validity of gains under high-stakes conditions takes the NAEP as a working operationalization of the domain, at least in mathematics, because the NAEP frameworks and assessments represent a degree of national consensus about what students should know and be able to do. However, a given state’s standards (Box 2) may be narrower than the NAEP framework and assessment, and in that case, the sample is narrowed to exclude material not specified by the state.

Figure 1 about here

Further narrowing arises because state tests often include predictable portions of the state standards, either omitting or giving very little emphasis to some of the standards
(Box 3 in Figure 1). As an example, we tabulated the proportion of the New York State mathematics standards assessed in the state’s tests in three grades. Until this year, the mathematics test was administered in March, and the state publicized which standards were “pre-March” and thus eligible for testing. Over a four-year span, the tests assessed almost 90 percent of the pre-March standards in grade 8 but only about 70 percent in grades 6 and 7 (Figure 2), and the tested proportion of the total set of grade-level standards was lower. Moreover, some of the standards that were tested were given very limited representation on the test, often only a single item. Similarly, when standards are written broadly, tests may sample content predictably within certain of the tested standards (Box 4).

Figure 2 about here

The result of the sampling in Boxes 2 through 4 can be elimination or de-emphasis on a substantial portion of the domain. Additional narrowing can arise because of non-substantive aspects of the test that can influence performance—for example, the formats used, the ways in which material is presented, and the rubrics used for scoring (Box 5 in Figure 1). In some cases, this last stage of narrowing is so extreme that some items are virtual clones of items used earlier—as in the extreme example from New York State seventh-grade math tests, is the pair of item clones shown in Figure 3.

Figure 3 about here
It is not an overstatement to say that educators and policymakers cannot sensibly use test data if they do not understand this sampling principle, but both the principle and its implications appear to be widely misunderstood.

**Types of tests and their advantages and disadvantages.** Much of the debate about choices among tests is rhetorical, fueled by a lack of understanding of types of tests and their strengths and weaknesses for specific uses. For example, it is not uncommon to hear advocates of standardized performance assessments strongly criticize the putative weaknesses of standardized tests, not recognizing that the tests for which they advocate are standardized, and the technical difficulties inherent in complex performance assessments are poorly understood. These issues will likely become even more important in the near future because of the widespread enthusiasm in the policy community for the incorporation of more complex performances into the large-scale assessments used for accountability.

**Imprecision and reliability.** Years ago, one of us served on a committee empaneled by the Kentucky legislature to evaluate the state’s KIRIS assessment program—in many ways, an influential precursor of NCLB. In a meeting with some legislators, an argument arose between panel members and staff from the Kentucky Department of Education about the Department’s estimates of “error”—by which statisticians do not mean “mistakes,” but rather imprecision—in schools’ aggregate scores. One legislator listened in silence for a bit, then turned to the Commissioner in evident anger. He was not concerned with whose estimate of error was more reasonable, which hinged on an esoteric argument about statistics and psychometrics. What angered him was that error existed—as he put it, that some schools had been rewarded because of
chance—and that he and other legislators had not heard about this previously. Of course, there is necessarily imprecision in test-based evaluations of schools, and even more in test-based evaluations of teachers.

This may be one bright spot, an instance in which the understanding of educators and policymakers has improved in recent years. Some states, e.g., Massachusetts, routinely include estimates of measurement error in reports of students’ scores. The option of excluding subgroups from adequate yearly progress calculations under No Child Left Behind if the groups are too small to provide reliable estimates has led to routine calculation of confidence intervals by education agency staffs. Statutes routinely make reference to “valid and reliable” assessments. Some news media have begun discussing the problem of imprecision in evaluations of teachers and schools.

Still, there is a difference between having a vague understanding of imprecision and being able to understand both the sources and magnitude of imprecision well enough to make sensible use of that information, and major misunderstandings still appear to be common. An example is the following quote from Jim Liebman, then Chief Accountability Officer of the New York City schools, defending the City’s school Progress Reports from criticisms about their unreliability:

The Progress Reports don’t use the outcomes of a sample of students; they average the actual outcomes of all students…The Progress Report is less like…polls…and more like an election…The results of the election are an estimate of people’s preferences – and an imperfect estimate at that; if the election were conducted five days in a row, the outcomes might differ by a few percentage points each day...Of course, we can’t wait for a “perfect”
vote. We move forward over the next four years by virtue of, an imperfect estimate based on a single high-stakes test that we arbitrarily set for the first Tuesday in November each Leap Year. Actual outcomes on actual single occasions have real consequences (Liebman, 2008).

This statement shows two fundamental misunderstandings of imprecision. First, it fails to distinguish sampling error (arising from the sampling of students to be tested) from measurement error (arising from imprecision in the test used to assess them).

Arithmetically more important, it confuses inferences about schools with conclusions about the particular students in who happen to be in a given school at one time. When the inference is about schools—as it is here—the important source of sampling error is variations in the students attending from one year to the next. The result of these two mistakes is that Liebman underestimates the degree of error in the school report cards.

These issues are again prominent today because of the current debate about using growth models to evaluate teachers. Teacher-level value-added estimates are extremely imprecise, and even school-level estimates are imprecise enough to seriously limit their uses. This imprecision is well documented in the scholarly literature (e.g., McCaffrey et al., 2009), but much of that literature is extremely technical and inaccessible even to many researchers. Policymakers are largely ignoring this problem, but it is not clear to what extent this represents a lack of understanding or a willful decision to ignore an inconvenient fact.

What performance standards are, and how they should be interpreted.

Standards-based reporting has been widespread for years. It is now required by the federal No Child Left Behind (NCLB) law, and it dominates the reporting of many large-
scale assessments, including most state testing programs. There has been a considerably amount of research exploring problems with standards-based reporting, but less on the use and interpretation of standards-based data by important stakeholders. Nonetheless, it became apparent early on that performance standards are problematic in ways that virtually guarantee that they will be misinterpreted. For example, the fact that standards vary—often dramatically—across variations in methods, item formats, item difficulty, and arcane aspects of procedure has been well established in the literature for more than 20 years (e.g., Jaeger, 1989; Linn, 2003). Thus, when a state or its contractor selects a method of standard-setting and a way to implement it, that choice—for which there is usually no clear substantive basis—can have a very large effect on the results, for example, the proportion of students deemed “Proficient” or “Failing.” This is not widely understood outside of the field and is almost always ignored by the press, policymakers, and the media. Standard setting by whatever method has been selected by a state is treated as if it uncovers a true, underlying categorization of competence. Similarly, a review of reporting by the press found that writers generally took the labels attached to standards as inherently meaningful and treated performance as discontinuous: some students are proficient, and others are not (Koretz & Deibert, 1996).

Policymakers and educators are even less well informed about the distortions that arise when standards-based ordinal scales are used to report trends. To be fair, one of the most important distortions—the inevitable misrepresentation of differences in trends between lower- and higher-achieving groups—is not simple to explain, as it arises from conflating the amount of progress with the density of the groups’ distributions of scores at the cut score. Still, one can illustrate concretely the severity of the problem without
explaining its mathematics. For example, Linn (2007) showed that if one uses NAEP data
to investigate changes in the achievement gaps among racial/ethnic groups, using
percents above Basic or mean scores suggest that the gaps have narrowed, while using
the percents above Proficient suggests that they have widened. The latter “finding” is
entirely misleading, a mathematical artifact of the location of the Proficient cut score in
the distributions of scores for whites, Hispanics, and blacks.

**What scales are, and how they should be interpreted.** If one sensibly avoids
relying on either raw scores or performance standards, one is left with traditional scales.
There are quite a number of them, and some are hard to explain. Still, the idea of a scale,
and an understanding of the properties of common ones, are essential for sensible use of
test scores. For example, the scales used for public reporting by many large-scale
assessments are arbitrary linear transformation of an initial scale that has a mean of
approximately zero and a standard deviation of roughly one. This scale is unacceptable
for public reporting; it would hardly do to say that a school’s average score is -0.3 or that
a student had shown marked improvement, from 0.1 to 0.5. Therefore, test designers
simply select the average score and standard deviation they want for the scale used for
reporting and transform the initial scale to obtain these values. This seeming arcane fact
has two very important, practical implications for data use. The first is that the apparent
size of a change in performance is meaningless. It is simply a function of the standard
deviceon and mean chosen by the vendor. Thus, for example, people routinely treat
numerically small gains on the NAEP scale as substantively small even when they are in
fact large, because of the large mean and small standard deviation make them *appear*
small. Second, percent change in scores on these scales is entirely meaningless: one can
make a given change appear large or small in percentage terms simply by changing the scale’s arbitrary mean. For example, when the scale for the SAT was last set, moving from the median to the 84th percentile rank required an scale score increase from 500 to 600, or “20 percent.” However, had the scale’s designers arbitrarily picked a mean of 600 instead, the exact same improvement would have been an increase from 600 to 700, or “17 percent.”

Given that performance standards are a very poor way to describe performance, simple changes in scale scores are arbitrary, and percent change is meaningless, sensible use of test scores requires some understanding of scales.

Inferences about teacher and school quality. It is important to distinguish between two different reasons why test scores by themselves are inadequate for evaluating schools and teachers. The first is that standardized tests cannot capture many of the important aspects of educational quality. This has been recognized in the psychometric community for generations (see, for example, Lindquist, 1951). The other reason is the stuff of basic statistics courses: test scores are powerfully influenced by factors that have nothing to do with the quality of teaching, and our data and analytical methods are not fully capable of disentangling these other influences from the effects of teachers and schools. Both of these reasons are routinely ignored by policymakers. It appears that the insufficiency of tests is increasingly ignored by educators as well. We have heard former teachers discuss this frequently, saying that new teachers in many schools are inculcated with the notion that raising scores in tested subjects is in itself the appropriate goal of instruction. However, we lack systematic data about this, and it also
remains unclear to what extent the decisions to ignore the limitations of test-based data reflect a lack of understanding rather than a willful decision to ignore what is understood.

The nature of and sources of student variability. The reforms of the past 30 years have been motivated in substantial part by a laudable desire to confront the glaring inequities in the U.S. school system. This goal is reflected not only in the policies themselves, but in the rhetoric that has accompanied them, e.g., “all children can learn to high levels” and “no child left behind.” This raises two unavoidable questions. How much of the variability in performance arises from these inequities? What policies would lessen the variability stemming from inequity while realistically accommodating variations in performance that do not arise from inequities and will persist?

The contribution of inequities to the overall variability of performance is far less than most people believe. For example, entirely eliminating the large mean differences between racial/ethnic groups in the U.S. would reduce the variability of performance only a very small amount. To illustrate this, we tabulated two nationally representative databases, the 1990 NAEP and the National Education Longitudinal Study of 1988 (NELS 88) to determine how much the standard deviation of student performance in eighth-grade reading and mathematics would shrink if every identified racial/ethnic group had the same distribution of scores as non-Hispanic whites. The reductions ranged from roughly half of one percent to 9 percent (Table 1). The mean differences in performance, as large as they are, are swamped by the very large variance of scores within groups.

International comparisons provide additional evidence that the variability of performance is not primarily a function of policy and actionable social inequities. If the unarguably severe social and educational inequities in U.S. society accounted for a large
share of the variability of our students’ performance, one might expect our variability to be much larger than that found in countries with more equitable educational systems. However, international comparisons routinely show that the variability of student performance is not anomalously large in the U.S. and that it is, in broad brush, quite similar in some high-performing countries, such as Japan and Korea. (The variability is distributed differently, however: in eighth grade mathematics, much more of the variation lies within classrooms in Japan and Korea than in the U.S., Germany, Australia, France, or Hong Kong; Koretz, McCaffrey, & Sullivan, 2001).

These facts are rarely discussed, and they are routinely ignored in the design of education policy. This information suggests that even if educational equity is markedly improved, we are stuck with a very large variability in student performance and that most students will reach proficiency only if the “proficient” standard is not set very high. For example, Linn (2000) showed that if the NAEP Proficient standard were used for accountability in mathematics, roughly one third of the eighth-grade students in the highest-scoring countries would fail to reach it. Nonetheless, our NCLB regulations hold that within a few years, even students with mild cognitive disabilities should reach the Proficient standard, and policymakers often point to the NAEP standards as indicating the level of rigor that states should impose with their own standards. The implications of this disparity between expectations and data are substantial. Unrealistic expectations about changes in the variability of student performance can lead to administering many students tests that are too easy or difficult and assess the wrong content, demoralizing some students and encouraging teachers to game the system.
Validity under high-stakes conditions. Research on score inflation is not abundant, largely for the reason discussed above: policymakers for the most part feel no obligation to allow the relevant research, which is not in their self-interest even when it is in the interests of students in schools. However, at this time, the evidence is both abundant enough and sufficiently often discussed that that the existence of the general issue of score inflation appears to be increasingly widely recognized by the media, policymakers, and educators.

Nonetheless, the problem still appears poorly understood. For example, it seems that few policymakers or educators recognize how severe score inflation has been in the cases where it has been documented. Policymakers often fail to recognize that score inflation appears to be highly variable and that it therefore can vitiate estimates of relative performance (hence, comparisons of schools and teachers) as well as overall estimates for a state or district. Both educators and policymakers seem predictably reluctant to consider that the problem might affect their own schools. Perhaps most important in terms of educational practice, most policymakers and educators appear to have a very weak understanding of the mechanisms that could lead to inflation, often falling back on some vague notion of inappropriate test preparation. A reasonable understanding of inappropriate test preparation is hindered by the notion of alignment. Many people argue, incorrectly, that if a test is “aligned with standards,” focusing preparation on the tested sample is fine. For example, in an online interview, Joel Klein, the Chancellor of the New York City Schools, asserted that “if you test the right skills, you want to make sure the kids learn them” (Klein, 2008). This confuses two issues: the quality of the tested material, and its completeness as a sample of the domain about which conclusions are
drawn. That is, many people mistakenly believe that inflation arises from bad (unaligned or undemanding) test content, whereas it actually arises from the incompleteness and predictability of the tested sample (Koretz, 2005).

The issue of score inflation is both poorly understood and widely ignored in the research community as well. To some extent, this reflects a lack of understanding. For example, we often hear researchers arguing that alignment should eliminate the problem of inflation or that inflation is consistent enough that it can be ignored for some analytical purposes. Still, we suspect that some of the failure to acknowledge the problem stems from the costliness of doing so: the risk of inflation makes most of the available test data potentially worthless for the research people want to conduct—for example, evaluations of educational interventions or investigations of the characteristics of “effective” schools or teachers. Inflation is not always present or severe, of course, but because auditing of score gains is so uncommon, we rarely know which data are meaningful and which not.

**Appropriate and inappropriate test preparation.** Directly related to the previous is the extent to which educators and policymakers understand the differences between appropriate and inappropriate test preparation.

The dividing line between appropriate and inappropriate preparation is not always entirely clear in a specific case, but the general principle is unambiguous. Preparation that leads to improved mastery of the domain is appropriate. This will necessarily produce score gains that show at least a modicum of generalization to performance in other contexts, including but not limited to other tests. Preparation that generates gains largely or entirely specific to the given test is inappropriate because it generates inflation rather than meaningful gains.
We have piloted a method of addressing this question but have not yet conducted a systematic study. We provided educators with a set of real test-preparation activities that ranged from appropriate (although generally boring and unintuitive, and hence not good instruction) to extreme coaching that borders on simple cheating. We had them work in small groups to evaluate each activity in terms of the principle noted above. This falls short of a test of their understanding, because we gave them the principle and only asked them to apply it. The responses were sobering. Many—in one group of 125 administrators, a substantial majority—defended almost everything. For example, one activity instructed students to use process of elimination to “solve” an algebra problem rather than do the mathematics. When most of the administrators said that this activity was fine, I pointed out that one could make the apparent gains it produces vanish simply by substituting a constructed response item, and I added that if the students encountered a similar problem in the real world, it was unlikely to appear in multiple-choice format. One administrator then defended the activity by saying that it was “an information retrieval strategy” and that her district emphasizes such strategies. Even the examples that were borderline cheating—presenting test preparation activities that were virtual clones of test items that would appear on the test—garnered strong support from some. In very few cases did they defend the inappropriate activities by arguing that they would produce generalizable gains. Part of the explanation may lie in the pervasive culture of test preparation, in which teachers are routinely encouraged to use inappropriate methods. Indeed, the most inappropriate example I provided them, which many of them defended, was provided by a district office to its own teachers, to prepare students for the district’s own algebra test. However, dissonance reduction may also play a role: I had asked them
to acknowledge that techniques similar to those upon which they routinely rely are in fact inappropriate.

**Uses of test data by educators**

The field has produced a substantial body of research on educators’ instructional responses to test-based information in the context of high-stakes testing programs. Much of this research has focused on educators’ responses to the characteristics of the tests or the test-based accountability system rather than test scores. For a short review that is now dated but still provides a good indication of the types of work that have been done, see Stecher (2002). A smaller number of studies examine teachers’ responses to scores. For example, several studies have explored the “bubble kids” phenomenon: teachers focusing undue attention on students thought to be near the cut score that matters in today’s accountability systems, at the expense of other students (Booher-Jennings, 2005; Hamilton et al., 2007; Neal & Schanzenbach, 2007).

It is useful to distinguish between two overlapping categories of responses to testing. The first, within-subject reallocation, refers to shifting instructional resources among elements of knowledge and skill that are relevant to the inference based on scores (Koretz & Hamilton, 2006; Koretz, McCaffrey, & Hamilton, 2001; Stecher, 2002). Reallocation has been found in a number of studies. For example, in surveys, many teachers have reported that they shift time away from material not covered or emphasized by the specific test used for accountability in order to allocate more to tested material (Stecher, 2002).

The second category of response is coaching. This term is used in many inconsistent ways by others, but we use it to refer to a focus on fine details of a specific
test (Koretz et al. 2001). These details may be substantive, that is, apparently related to the subject area, but when they are substantive, they are either incidental or are too fine-grained to be of importance to the inference. An example would be noting which specific polygons happen to be used in a test and focusing on those at the expense of others. (In the words of one Boston teacher responding to a survey instrument we piloted years ago: “why would I teach irregular polygons?” The implication was not that they are unimportant, but merely that they don’t show up in the state’s assessments). The details may also be nonsubstantive, for example, particular formats, forms of presentation, or response demands. We include test-taking tricks, such as process of elimination (which capitalizes on the incidental use of the multiple-choice format) under the rubric of coaching. Whereas reallocation refers to shifting resources among units of content—skills and knowledge that are relevant to the inference—coaching entails “shifting to the measured instantiations of a concept [or other material relevant to the inference] rather than the unmeasured instantiations of the same concept” (Stecher, 2007).

Research on coaching is very limited. Test preparation materials (e.g., Rubinstein, 2000) often place a substantial emphasis on coaching, and anecdotal reports from teachers and others about reliance on these techniques are extremely common. However, the systematic research on educators’ responses to testing stop short of what is needed for an understanding of coaching. For example, surveys indicate that many teachers use old test forms and old test items in instruction—which many states actively encourage—but they do not reveal how teachers use them.

A hypothetical example to make this concrete: suppose that a teacher looks at test results for her class and finds evidence that her students do relatively poorly on problems
that entail proportional reasoning. In response, she looks for more effective methods of teaching these concepts, and over time, her students’ scores rise because they have learned more about proportional reasoning. This is a perfectly appropriate way in which to use item-level data, and if done well, it will produce generalizable gains. In contrast, suppose another teacher finds that her students tend to do poorly on items about the Pythagorean theorem. She looks to the Princeton Review for help and finds this suggestion: “Popular Pythagorean ratios include the 3:4:5 (and its multiples) and the 5:12:13 (and its multiples)” (Rubinstein, 2002, p. 56, emphasis added). She has her students memorize these ratios rather than struggle with teaching them the Pythagorean theorem. These ratios, however, are “popular” among test authors (to provide integer solutions), not in the real world, so any gains from this coaching will not generalize well to other contexts. That is, this coaching will contribute to score inflation. Similarly, consider the near-clone items showing Figure 3. If one expects this style of item to be repeated, one can prepare students for it simply by telling them that it a question asks about measuring mass, look for an answer with the word “scale.”

**Suggestions for a research agenda**

The discussion above suggests three broad areas for research.

**What do educators and policymakers understand about testing and test-based data?** As the description above suggests, this is a very broad topic. To make it tractable, we suggest viewing it as comprising three different but overlapping areas.

The first might be called low-stakes testing issues. These are the nuts and bolts of testing that have been important since the onset of standardized testing. This area includes, for example, the sampling principle of testing, imprecision and reliability,
validity and bias, and so on. A research program addressing these issues would be relatively uncontroversial, at least compared to some of the others here, and therefore it might be one of the easier to implement.

The second area is high-stakes testing issues, the issues that have become critically important because of test-based accountability. This area includes the validity of gains—score inflation—and the distinctions between appropriate and inappropriate test preparation.

The third area is the reporting and use of test-based performance data. In the current context, particularly important issues are the understanding of standards, the understanding of growth models, and notions about the using scores to infer the effectiveness of schools and teachers.

Although research of this sort is important, it will be difficult to carry out. One obstacle is that it is very difficult to design a survey instrument addressing this topic that does not seem like a test, and there is therefore a high risk that sampled individuals will refuse to participate or obtain answers from other sources before answering if a mail survey is used. One alternative would be computer-assisted telephone interviews, but these are difficult and expensive to conduct with teachers because they do not have their own phones at school. Another alternative is face-to-face interviews, but are far more burdensome and expensive.

What opportunities do testing programs provide for inappropriate test preparation and score inflation? This may seem out of place, since it is not a question of either understanding or the use of test-based data. However, answering this question is a prerequisite for addressing the next one adequately.
The root of inappropriate test preparation and score inflation is the incompleteness of tests as measures of the valued outcomes of education. Some portions of that incompleteness are outside the sphere of testing altogether and must be addressed by other means, such as teachers’ observations of students’ work or inspectors’ observations of teachers. However, within the domains that tests are intended to assess, the design of tests will determine what opportunities for inflation are present.

What makes inappropriate test preparation and score inflation feasible is predictable recurrences in and omissions from the tested material. These predictable patterns can include all of the levels of sampling shown in Figure 1 and thus can include not only substantive choices of content, but also substantively unimportant details of the test, such as forms of presentation and response demands. Under low-stakes conditions, teachers have relatively little incentive to focus on these predictable patterns, but even then, they often do to some degree (Lindquist, 1951). Under high-stakes conditions, they have strong incentives to focus on this predictable sampling, and their doing so undermines the ability of tested performance to represent mastery of the domain.

We noted above that vendors of test-preparation materials focus on these predictable recurrences. Increasingly, so do districts, schools, and even the state departments of education that are putatively using tests to monitor achievement. Some of these efforts are often given the disarming label of “power standards.” A nice example appears in Figure 4, which is extracted from a PowerPoint file provided by the Quincy, Massachusetts education agency to its secondary-school mathematics teachers.¹ The first page of the PowerPoint presentation (the link to which is provided below Figure 4) shows

¹ We are indebted to one of our students, Rebecca Holcombe, for this example.
three texts used in Quincy’s mathematics classes. When a teacher chooses her text, she is given a table of contents. Picking a chapter brings one to a page such as Figure 4, which represents Chapter 7 of the algebra text. Each section of the chapter is given a row, and in that row are listed all of the items testing the content of that section that appeared on the state’s 10th grade mathematics test over a period of four years. In the original, the item numbers are live links that take one to the actual items. Blank rows represent untested content. This makes it trivially easy to eliminate instructional time from all portions of the text that are never or rarely tested.

However, in many cases, the predictable patterns in tests that provide opportunities for score inflation are not as glaring as these two examples suggest. Test-preparation firms examine tests to identify these opportunities, but researchers rarely have. We are currently completing a systematic examination of several years of tests from Massachusetts and New York and will soon have a detailed comparison of the opportunities for score inflation provided by the two tests. To our knowledge, this paper will be the first academic paper of its kind.

Additional work of this sort is important for three reasons. First, in itself, it will provide a useful basis for discussions of improved test designs better suited to the demands imposed by accountability. Second, it can be paired with test data to determine empirically which opportunities for inflation are actually and effectively used. Third, it is
essential for addressing adequately the next question, how educators use test-based information.

**How is test-based information used by educators?** This area includes both how educators use information about the tests themselves—as in the preceding discussion—and how they use student performance data.

The types of research done to date on this topic, while useful, are insufficient. We have to expand research on reallocation to take into account the growing focus on ‘power standards’ and our growing understanding of the opportunities for reallocation created by systematic predictability of content sampling. We need to design ways of measuring coaching, which has been almost entirely unstudied.

Finally, we need to begin studying variations in educators’ used of test-based information. Much of the extant literature either uses very small-scale case studies to explore responses to testing in one locale or makes use of surveys to obtain information about aggregate responses in a state or district. We have few systematic studies of variations in educators’ responses—for example, variations that might be associated with the performance level of schools, the characteristics of students in the schools, the characteristics of teachers, the nature of the tests employed, or the characteristics of the accountability systems into which tests are embedded. This sort of comparative information is essential for the design of more effective programs, and it is essential for addressing issues of equity.

**Variations in score inflation.** Ultimately, our concern is the impact of educators’ understanding and use of test data on student learning. However, at this point, we have very little comparative information about the validity of gains, just as we have almost no
information about variations in educators’ use of test data. The comparative information that is beginning to emerge suggests that score inflation may be distributed in systematic ways that have important implications for policy, which underscores the importance of this portion of the research agenda. In particular, evidence is beginning to emerge that score inflation is often more severe for low achieving and disadvantaged students, which can create an illusion of increased equity. For example, Figure 5 shows trends in the white-black difference in 8th grade reading in a large, high-poverty district, on both a high-stakes test and a lower-stakes audit test. The first high-stakes test was replaced in 2002. Two disturbing patterns are evident. The first is that since 2000, the achievement gap narrowed substantially on the high-stakes tests but not on the audit test. The second is that the white-black gap increased markedly when the first high-stakes test was replaced by another. Both of these patterns suggest that the scores of black students on the two high-stakes test were inflated more than those of whites, creating a misleading impression that the achievement gap was closing.

Figure 5 about here

This one example illustrates the importance exploring how variations in educators’ behavior and the validity of gains vary as a function of characteristics of students and schools, test design, and the design of accountability systems. This information is essential for the design of more effective and equitable accountability systems.
References


Northern Illinois University Curriculum and Instruction Reading Clinic, Literacy Research Report 4.


Figure 1. Opportunities for narrowed instruction and score inflation in the design of a test

1. Domain selected for testing (math, ELA, etc.)

2. Elements from domain included in Elements from domain omitted

3. Tested subset of standards Untested subset of standards

4. Tested material from within tested Untested material from within

5. Tested representations (substantive, Untested representations

NOTE: “tested” vs. “untested” also represents material give substantial or little emphasis.
Figure 2. Proportion of state mathematics standards assessed in New York State’s tests, by year and grade

Percent of Pre-March NY State Math Standards Tested

- 6th
- 7th
- 8th

2009
2006-09
Figure 3. Near-clone items from New York State’s 7th grade math test, 2008 and 2009

2009 #9 (Standard 7M9)

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

2008 #27 (Standard 7M9)

27 Which tool is most appropriate for measuring the mass of a serving of cheese?
   A  ruler
   B  thermometer
   C  measuring cup
   D  weighing scale
Figure 4. A guide for reallocation

Source: Quincy, MA, Public Schools,

Figure 5. Trends in white-black gap in reading in the Houston Independent School District, two accountability tests and an audit test
Table 1. Percent reduction in standard deviation of scores if racial/ethnic differences were eliminated

<table>
<thead>
<tr>
<th></th>
<th>NAEP 90</th>
<th>NELS 88</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mathematics</td>
<td>-9%</td>
<td>-5%</td>
</tr>
<tr>
<td>Reading</td>
<td>-0.4%</td>
<td>-4%</td>
</tr>
</tbody>
</table>