The roots of score inflation: An examination of opportunities in two states' tests

Citation	Holcombe, R., Jennings, J., & Koretz, D. (2013). The roots of score inflation: An examination of opportunities in two states' tests. In G. Sunderman (Ed.), Charting reform, achieving equity in a diverse nation, 163-189. Greenwich, CT: Information Age Publishing.
Citable link	http://nrs.harvard.edu/urn-3:HUL.InstRepos:10880587
Terms of Use	This article was downloaded from Harvard University's DASH repository, and is made available under the terms and conditions applicable to Other Posted Material, as set forth at http://nrs.harvard.edu/urn-3:HUL.InstRepos:dash.current.terms-of-use#LAA

THE ROOTS OF SCORE INFLATION:
AN EXAMINATION OF OPPORTUNITIES IN TWO STATES' TESTS

To appear in G. Sunderman (Ed.), *Charting Reform, Achieving Equity in a Diverse Nation*.
Greenwich, CT: Information Age Publishing.

July 18, 2012

Rebecca Holcombe, Ed.M., M.B.A
Director of Teacher Education
Dartmouth College
Hanover, NH
Rebecca.Holcombe@dartmouth.edu
(603) 643-1502

Jennifer L. Jennings, Ph.D.
Assistant Professor
Department of Sociology
295 Lafayette Street, 4th Floor
New York University
New York, New York 10012
jj73@nyu.edu
(212) 992-7465

Daniel Koretz, Ph.D.
Henry Lee Shattuck Professor of Education
Harvard Graduate School of Education
415 Gutman Library 6 Appian Way
Cambridge, MA 02138
daniel_koretz@harvard.edu
(617) 384-8090

Acknowledgements: The research reported here was supported by the Institute of Education Sciences, U.S. Department of Education, through Grant R305AII0420, and by the Spencer Foundation, through Grants 201100075 and 201200071, to the President and Fellows of Harvard College. The opinions expressed are those of the authors and do not represent views of the Institute, the U.S. Department of Education, or the Spencer Foundation.

Since the 1970s, policymakers have relied on *test-based accountability* (TBA) as a primary tool for improving student achievement and reducing racial and socioeconomic achievement gaps. These policies have produced striking gains in scores on some accountability tests and, in some cases, seeming evidence of narrowing achievement gaps. As a result, support for test-based accountability has been widespread. Most policymakers are confident that score gains signify commensurate increases in achievement and will translate into improvements in children's long-term life chances, particularly for poor and minority children.

However, more than two decades of research indicates that TBA policies have not been the unqualified success that gains on high-stakes tests might suggest and have led to a variety of undesirable side effects. A recent National Research Council review concluded that the effects of TBA programs have ranged from zero to small (National Research Council, 2011). In addition, studies have identified distortions of educational practice, such as reducing instructional time allocated to material not emphasized on state tests (e.g., Koretz, Barron, Mitchell, & Stecher, 1996; Koretz, Mitchell, Barron, & Keith, 1996; Stecher, 2002; Stecher et al. 2000) in order to focus instruction on material that is predictably emphasized on the test. These practices can lead to *score inflation*- that is, gains in scores on the tests used for accountability that are markedly larger than the actual gains in student learning they are intended to signal (Koretz and Hamilton, 2006). Numerous studies have found that score gains on high-stakes tests often do not generalize to lower-stakes assessments used as audit tests, such as the National Assessment of Educational Progress (NAEP) (e.g., Center on Education Policy, 2008; Fuller et al., 2006; Jacob, 2007; Klein et al., 2000; Koretz & Barron, 1998; Koretz, Linn, Dunbar & Shepard, 1991; Lee, 2007). Although score inflation is highly variable across states (Ho, 2007), it is often severe. Some studies also have found that TBA programs can generate illusory improvements in equity (Klein et al., 2000).

Score inflation has at least four important consequences. First, when scores are inflated, students have learned less than their scores suggest. Second, because schools serving poor and minority students face the most pressure to quickly increase test scores and greater barriers to doing so, inflation may affect them more severely, with negative implications for educational equity. Third, because inflation varies across schools, *relative* improvements become difficult to evaluate, and researchers and policymakers will misidentify effective and ineffective schools and teachers. Finally, accurate data on how students are performing is necessary for policymakers to identify and implement interventions that improve students' long-term outcomes. In the absence of such data, we may pursue interventions that are not effective or fail to implement policies that could improve student achievement.

To date, few studies have attempted to understand the sources of variation in score inflation across testing programs. In particular, research has not identified the specific characteristics of tests that facilitate or impede score inflation and inappropriate test preparation, that is, test preparation that inflates scores. Without this information, it is impossible to improve existing assessments to lessen these problems.

This chapter is the first attempt in the literature to systematically investigate the opportunities for score inflation within current tests. We evaluate released items from the mathematics tests of two states, Massachusetts and New York, in the years 2006-2008, to identify predictable recurrences that provide opportunities for narrowed instruction that might inflate scores. In this paper, we focus primarily on the 8th-grade tests in both states and the 10th-grade test in Massachusetts, but we also draw selectively from other grades for purposes of illustration. We selected Massachusetts and New York because they provide a strong contrast. The New York and Massachusetts content standards are very different; New York's are much narrower than those of Massachusetts, sometimes specifying a single fact or skill. At least through 2005, NAEP trends did not suggest inflation of Massachusetts Comprehensive Assessment System (MCAS) math scores in grades 4 or 8 (Ho, 2007), although no audit studies have been conducted with the higher-stakes tenth-grade MCAS test that is our focus. In contrast, New York State's eighth-grade math scores increased far more rapidly than its NAEP scores did after 2010. Both the Massachusetts MCAS tests and the state's reforms are generally well-regarded in the policy community. For example, Diane Ravitch asserted that "Massachusetts… has earned its high marks the old-fashioned way, by improving its curriculum, testing incoming teachers, and sponsoring assessments far superior to those in most other states" (Ravitch, 2010). On the other hand, New York's testing program has been the focus of considerable controversy in the press and policy community. Unlike policymakers in many states, then-Commissioner David Steiner acknowledged in 2010 the likelihood of inflation in the state's test scores, and the New York State Department of Education has since that time pursued a policy of making its tests broader and less predictable in order to combat inflation.

Previous Research

The research literature on responses to high-stakes testing falls into three categories: 1) studies of the manipulation of the tested population, 2) studies of score inflation, and 3) investigations of changes in instructional strategies—including test preparation—and other aspects of practice.

The first category of studies, primarily in economics and sociology, examines strategies to remove students from the tested population in order to raise aggregate scores. For example, one study found that when schools are permitted to exclude special education students from testing, they assign more students to special education (Figlio & Getlzer, 2006). Other studies have found increased exclusion of low-scoring students (e.g., Jacob, 2005; Jennings & Beveridge, 2009). Although such strategies can create a modest bias in aggregate scores, they do not inflate the scores of individual students. Moreover, they are unrelated to the content of the specific tests. Therefore, they are not pertinent to the present study.

Studies in the second category (score inflation) follow a common logic. Achievement tests are small samples drawn from large domains such as “eighth grade mathematics” that are the target of users’ inferences. The inference about mastery of the large domain based on scores is only valid to the extent that one can generalize from the small, tested sample to mastery of the largely untested domain. If score gains do signify increased mastery of the larger domain, reasonably similar gains should appear in other samples from that domain—that is, in scores on other tests intended to support similar inferences about achievement. Therefore, most studies of score inflation have investigated the extent to which gains in scores on a high-stakes test generalize to trends on a lower-stakes audit test.

The first empirical investigation of inflation was conducted by Koretz, Linn, Dunbar, and Shepard (1991) in a district with a testing policy that was high-stakes by the standards of the day, but much lower-stakes than the norm today. The district administration pressured schools and teachers to improve scores, but there were no concrete sanctions or rewards for scores. When the district replaced one commercial, multiple-choice, basic skills achievement test battery with a quite similar competitor, math scores at the end of third grade dropped by half an academic year. Four years later, scores on the new test had reached the level that had been reached on the old test in its final year of use. This “sawtooth” pattern accompanying the adoption of new tests had been widely recognized in the psychometric community, but no prior studies examined whether the skills assessed by the earlier test remained stable or atrophied after the introduction of the new test. Koretz et al. administered the test that the district had abandoned four years earlier to a random sample of classrooms. Scores on that test had dropped by half an academic year while scores on the new test had risen by a like amount. Another random sample of classrooms was administered a parallel form of the new test (that is, a form designed to be equivalent) to evaluate whether performance on the researchers’ tests were depressed by motivational factors. No motivational bias was found.

Since then, a substantial number of studies have followed the same approach, but using audit tests that were already in place rather than administering an alternative test experimentally. The most commonly used audit test has been NAEP. NAEP represents a degree of national consensus about what students should know, and scores are unlikely to be inflated because teachers have no incentive to prepare students specifically for it.

Because of a widespread belief in the policy community that the multiple choice format was responsible for inflation, Koretz & Barron (1998) evaluated score inflation in

Kentucky's KIRIS high-stakes assessment program, which deemphasized the multiple-choice format in some years and avoided it entirely in others. They found that gains on the KIRIS tests in mathematics in grades 4 and 8 were three to four times as large as the state's gain on NAEP. Hambleton et al. (1995) examined performance gains in fourth grade reading in the same KIRIS assessment and found that scores on the state assessment increased by .76 of a standard deviation in the space of only two years—an extraordinarily rapid increase by historical standards (see Koretz, 1986)—while scores on the NAEP reading assessment did not increase at all.

Klein et al.'s (2000) study of Texas' state test and NAEP trends found a disparity in trends between the state's TAAS test and NAEP that was similar in magnitude to that found in Kentucky by Koretz & Barron (1998). In addition, they found that the widely cited "Texas miracle" of a rapidly shrinking gap between minority and non-Hispanic white students was not reflected in NAEP data. Similarly, Jacob (2005) analyzed data from the Chicago Public Schools' Iowa Test of Basic Skills (ITBS), which at that time was high-stakes and used for student promotion decisions as well as school accountability, and the low-stakes Illinois Goals Assessment Program (IGAP). He found large gains on the high-stakes ITBS following the introduction of accountability, but no similar effects of the accountability system on the IGAP.

A number of other studies that compared trends across numerous states found that state test gains typically outpace national test gains, often by a large magnitude (Center on Education Policy, 2008; Fuller, 2006; Carnoy and Loeb, 2002; Hanushek and Raymond, 2004, 2005; Jacob, 2005, 2007; Grissmer and Flanagan, 1998, Rouse et al

2007). These disparities in trends are not uniform, however, and if NAEP is used as a comparison, some states appear to have escaped serious inflation (Ho, 2007).

The third category of studies examined a variety of behavioral responses to high-stakes testing that have the potential to inflate scores, including the use of explicit test preparation, other instructional strategies, changes in school management, and inappropriate testing practices. We are concerned here with test preparation and other instructional responses. Following the terminology introduced by Koretz and colleagues (Koretz et al., 2001; Koretz & Hamilton, 2006), we distinguish between three types of responses that can inflate scores: cheating, reallocation, and coaching. Cheating obviously inflates scores and is therefore not a focus of this study.

Reallocation refers to redistributing instructional resources, including instructional time, to better align with the content of the test. Reallocation is not necessarily undesirable. Indeed, one of the aims of TBA is to encourage teachers and students to align their activities with standards, that is, to focus more on material that has been deemed important and is therefore tested. However, reallocation (and alignment, which is a form of reallocation) can inflate scores, if it entails reducing the resources directed to portions of the domain that are important for the inferences based on scores but that are omitted from or given little emphasis in the test. Doing so makes performance on the tested sample unrepresentative of mastery of the domain as a whole (Koretz et al., 2001; Koretz, 2005).

The term *coaching* has been used in many ways, but we use it to mean focusing on incidental details of the test that are unimportant for the inference based on scores. These can be either aspects of format or unimportant details of content. We include under coaching test-preparation strategies such as process of elimination for multiple-choice items. Coaching inflates scores if it leads to higher performance than one would see if those unimportant details were changed. For example, a study by Shepard (1988) illustrated how severe a failure of generalizability across formats can be. Using data from a state test, Shepard found that when addition items were presented in a vertical format, 86 percent of students answered these addition questions correctly, but in horizontal format, only 46 percent of students did. In subtraction, the percentages were 78 and 30, respectively.

A number of studies have determined that teachers reallocate time within subjects in response to TBA, often at the expense of attention to important material (e.g., Koretz, Barron, Mitchell, & Stecher, 1996; Koretz, Mitchell, Barron, & Keith, 1996; Stecher, 2002). In a recent RAND survey of teachers in California, Pennsylvania, and Georgia, teachers reported that because there were so many standards, they identified “highly assessed standards” on which to focus their attention (Hamilton and Stecher, 2007). Earlier studies by Shepard and Dougherty (1991) and Romberg, Zarinia & Williams (1989) found evidence of increased focus on basic skills to the exclusion of more complex skills in response to the emphasis placed on basic skills in state tests.

Studies have found a variety of forms of coaching in response to high-stakes testing. Studies by Darling-Hammond and Wise (1985), Shepard and Dougherty (1991), Smith and Rottenberg (1991), McNeil (2000), and Pedulla et al (2003) all showed that teachers focused their instruction not only on the content of the test, but also on its format, by presenting material in formats as they will appear on the test and designing tasks to mirror the content of the tests.

While this research provides ample evidence of test preparation and other instructional responses that have the potential to inflate scores, the dividing line between inappropriate and appropriate test preparation remains unclear. Indeed, many state and local policies have blurred the distinction even more in recent years, e.g., by encouraging the use of old test forms as instructional materials and announcing “power standards” well in advance of testing dates. In the end, the distinction is empirical: test preparation is desirable if it produces valid gains in scores and undesirable if it results in inflated gains. Research has not yet tied variations in score inflation to specific forms of test preparation.

Opportunities for Inflation in the Construction of a Test

Tests used for accountability are necessarily only a small sample of a knowledge domain—for example, “eighth grade mathematics”—but must support inferences about students’ command of the entire domain. Moreover, the tested sample is almost always systematically incomplete, because some types of knowledge and skills are harder to assess than others. Nonetheless, under low-stakes conditions, this incomplete sampling often works reasonably well. Because of the small sample, estimates of performance are subject to substantial measurement error, but this adds only imprecision, not bias. Because test authors create different samples from the domain, the results from different tests of the same domain often differ, but these differences are usually modest.

Using tests for accountability, however, poses a fundamental threat to the validity of inferences based on these tests. Even under low-stakes conditions, teachers may begin focusing on the specifics of the tested sample at the cost of material that is important but not emphasized on the test, thereby undermining the test’s representation of the larger domain (e.g., Lindquist, 1951). However, under low-stakes conditions, the incentives to focus on the tested specifics are usually mild. In contrast, under high-stakes conditions like those created by test-based accountability, educators face strong incentives to focus instruction on the specific content and format of tested material rather than the full domain of knowledge and skills represented in the state standards. *If these aspects of the test are predictable* and teachers focus on them, scores will become inflated, and inferences about mastery of the domain will be undermined.

Koretz et al. (Koretz, McCaffrey, & Hamilton, 2001; Koretz & Hamilton, 2006) provided a formal framework for evaluating the effects of this sampling on validity under high-stakes conditions. They used the term *performance element* to denote each aspect of performance that affects performance on a test or is relevant to the inference the test scores are intended to support. An element's *test weight* is the emphasis given to it by the test—specifically, its influence on scores—while the *inference weight* is the importance of that element to the inference based on scores. Tests will often omit elements that are important for the inference or assign them weights that are not proportional to the importance of these elements to the inference. The validity of an inference about improved learning is the degree to which improvement on the tested elements, as weighted by the particular test, accurately signals improvement in command of the entire domain, as weighted by the inference. Following this framework, we label standards that are emphasized by the test as *high-weight standards* and items assessing them as *high-weight items*.

What specific opportunities do tests provide for inappropriate narrowing that can inflate scores? Test-preparation materials often suggest a variety of strategies, ranging from focusing on the standards most often emphasized by the test to taking advantage of item format, for example, by using process of elimination. To examine the opportunities systematically, it is useful to think of the construction of a test as comprising several successive stages of narrowing down from the broad “target” (Kane, 2006)—the largely unmeasured construct about which we want to draw conclusions based on scores—to the specific items presented to students. Although the process of building a test need not always follow this sequence, the steps provide a helpful way to categorize the opportunities for inappropriate narrowing presented to teachers and students.

The first opportunity for inflation can arise in the specification of a state’s standards. This stage of sampling reduces a domain selected for testing, such as “eighth-grade mathematics,” to the subset of that domain included in a given state’s content standards. This is represented as the step from Box 1 to Box 2 in Figure 7.1. (In each row of Figure 7.1, the included or emphasized material is in the left-hand box, while the material omitted or de-emphasized at that stage is in the shaded, right-hand box.) States vary in terms of both the breadth of their standards as a set and in the breadth of individual standards. For example, the Massachusetts standards for eighth-grade math include all five of the content strands represented in NAEP, while the New York standards omit one of the NAEP content strands—data analysis and statistics. Many of New York’s standards are framed very narrowly, in some cases so narrowly that they come close to implying specific test questions. For example, geometry standard 8G4 requires that students, “Determine angle pair relationships when given two parallel lines cut by a transversal.” In contrast, most of the individual standards specified by Massachusetts delineate somewhat broader categories of knowledge and skills. For example, the Massachusetts 8th measurement standard 8.M.3 states that students will:

"Demonstrate an understanding of the concepts and apply formulas and procedures for determining measures, including those of area and perimeter/circumference of parallelograms, trapezoids, and circles. Given the formulas, determine the surface area and volume of rectangular prisms, cylinders, and spheres. Use technology as appropriate." (Note, however, that the second-to-last statement suggests a narrower operationalization of the standard.) Inflation that arises because educators no longer teach important content that is omitted from standards can only be identified by comparison to an external audit test that is judged to be a good measure of the relevant domain—in practice, most often the NAEP.

Figure 7.1 about here

Because of limits on testing time, a second opportunity for narrowing arises: a given test is unlikely to sample exhaustively from the state's standards. Moreover, even among standards that are sampled, some standards may be sampled much more heavily than others, giving performance on these heavily emphasized standards disproportionate impact on students' overall scores. The selection of standards for inclusion or emphasis is represented by Box 3 in Figure 7.1. If this stage of sampling is unpredictable over time, it is unlikely to produce score inflation. For example, the students of teachers who focus unduly on standards emphasized in previous tests are likely to do poorly on previously unemphasized material included in the current-year test. Variations in test weights will also not inflate scores if they correspond to variations in inference weights—that is, if only material more important for the inference is given more emphasis by the test.

However, if variations in emphasis are persistent, predictable, and unjustified by the inference, this stage of sampling provides an opportunity for reallocation that will inflate scores.

The third stage of sampling that provides opportunities for narrowed instruction is the uneven sampling of skills within standards (Box 4 in Figure 7.1). Some standards are written broadly, such that a single standard includes mastery of a cluster of skills and concepts. When a standard is written broadly, test designers sample from it in constructing a test, just as they sample among standards. If the within-standard sampling is predictable, that creates an incentive for teachers to focus on the emphasized aspects of the standard at the expense of those not emphasized or excluded. For example, if a given standard specifies that students learn rotation, transformation, and dilation of a polygon on a coordinate plane but only rotations are included in the test, teachers in high stakes contexts have an incentive to focus their instruction only on rotations. If they do, their students are likely to perform better on items representing that standard than they would on an alternative set of items representing the standard more fully.

The final stage of sampling, represented by Box 5 in Figure 7.1, is the choice of representations used to test each standard. We use the term *representation* to refer to both unimportant details of content and what we term *item style*. Item style is broader than item format, in the usual sense. For example, it includes the type of visual representation in the item, if any, the magnitude and complexity of the numbers used, and so on. For example, consider a hypothetical standard stating that students should understand the concept of slope in the context of simple linear equations. Problems involving slope could be presented verbally, graphically, or algebraically, or they could require translation among those representations. If the problems are presented graphically, they could be presented only with positive slopes in the first quadrant, or with a mix of positive, negative, and zero slopes in all four quadrants. In some cases, these choices are similar or identical from one year to the next. In extreme cases, the chosen representations may be so similar over time that new items are virtual clones of earlier ones.

The use of predictable representations facilitates coaching. If tests employ similar or identical representations in successive years, then using prior tests to practice renders these items familiar and may make them easier than other items with different item styles that were measuring the same underlying mathematical content. In addition, the use of predictable representations may make it easier to develop forms of test preparation that allow students to correctly answer items without gaining the knowledge or skills the item is intended to tap. As Shepard (1988) observed, policy makers should be interested in whether children can add and subtract, and not in whether they can add and subtract when items are presented in a vertical format.

Data and Methods

We examined the 2006 through 2008 eighth-grade mathematics tests in Massachusetts and New York and the tenth-grade mathematics test in Massachusetts. Both Massachusetts and New York released item maps that link each test question to a state standard and a strand of mathematics (e.g., number sense and operations, measurement, statistics and probability, algebra, and geometry). Using these item maps, we constructed item-level datasets that organized all released items sequentially within individual standards. Released items in NY include all items used on the 8th grade math test. Released items in MA at that time includes all common items—that is, items taken by all students and used to generate students' scores. This database facilitates comparison of related items to identify recurrences of content or presentation that would afford opportunities for narrowed instruction and test preparation that might induce score inflation.

We examined recurrences largely following the schematic in Figure 7.1. We calculated the fraction of the strands and standards tested in each year and state, as well as the fraction of each area ever tested over 2006-2008. We examined the relative weighting of standards on the state exams both within and across years. To address differences in the narrowness of individual standards, we mapped the eighth-grade standards in New York onto those in Massachusetts and evaluated the breadth of the New York tests in terms of Massachusetts' broader standards. We isolated clusters of extremely similar items on each test in the years 2006-2008, and mapped backwards to the standards to identify cases where similar items were used to sample nominally different standards.

Because some standards, particularly in Massachusetts, represent broader bundles of skills and knowledge, we explored how well the tests evenly sample skills represented within a given standard. For example, a standard might require students to master tax, percent increase/decrease, simple interest, and sale prices, but consistently ask questions about interest rather than sale prices.

We examined the sequence of items within each standard in all of the tests to identify recurrences of representations. In theory, these recurrences can entail either aspects of item style or unimportant details of content—that is, details that are not warranted by the inference based on scores. However, in practice, the distinction between these two aspects of representation is often unclear, and we found that we could not reliably code it. Therefore, we do not make this distinction in the presentation of results.

Results

We present results here in the order presented in Figure 7.1.

Elements from the domain included in the standards (Box 2 in Figure 7.1)

Narrowing from the domain to the standards occurs in the selection of content for inclusion in the standards, the wording of the standards, and the operationalization of the standards in the test. To quantify the degree of narrowing between the larger domain in an area such as “eighth grade mathematics” and the domain described in a given state’s standards, we would need a formal definition of what constitutes the larger domain, and in the U.S., agreement about the domain is, at best, incomplete. Nonetheless, we can illustrate the process by comparing state standards to the framework of standards of the National Assessment of Educational Progress, which is intended to reflect a degree of national consensus about what students should know and be able to do. We can also compare the standards of the two states to compare the breadth, depth and complexity of skills and understandings they describe.[1] A comparison of eighth-grade standards from the two states provides striking examples of this stage of narrowing. In several respects, New York’s standards and tests were markedly narrower than those of Massachusetts, and the details of that comparison offer concrete illustrations of the ways in which this narrowing can occur.

We have already an obvious instance of narrowing: the New York State eighth-grade mathematics standards entirely omitted the data analysis, statistics, and probability strand included in the NAEP, although the state’s seventh-grade standards did include this strand. In contrast, the eighth-grade Massachusetts standards included five strands similar to those in the NAEP framework. Mapping from state standards to NAEP standards is inherently ambiguous because the two sets of standards divide content differently, and often there is a limited degree of overlap between two seemingly different standards. Nonetheless, it is clear that the New York standards were also somewhat narrower than the NAEP standards in the four strands they have in common. For example, the eighth-grade NAEP standards “Estimate square or cube roots of numbers less than 1,000 between two whole numbers” and “Visualize or describe the cross section of a solid” were not present in the New York eighth-grade standards.

A simple count of standards would suggest that New York’s standards are broader than those of Massachusetts. The Massachusetts Mathematics Curriculum Framework (Massachusetts Department of Education, 2000) lists 39 eighth-grade content standards spanning five strands. New York State’s Mathematics Core Curriculum (New York State Education Department, 2005) lists 48 eighth-grade content standards, even though this document excludes the data analysis, statistics, and probability strand.

More detailed examination, however, shows that the reverse is true: New York samples more narrowly from the domain. One reason is the wording of New York’s standards, many of which were very narrowly worded, some so much so that they encompassed only one or a few pieces of information. For example, standard 8.G.1 was simply “Identify pairs of vertical angles as congruent.” Moreover, this standard was just a subset of another, slightly less narrow standard, 8.G.6: “Calculate the missing angle measurements when given two intersecting lines and an angle" (New York State Education Department, 2005, p. 86).

In contrast, the Massachusetts standards were typically much broader. For example, Massachusetts standard 8.G.6 was "Predict the results of transformations on unmarked or coordinate planes and draw the transformed figure, e.g., predict how tessellations transform under translations, reflections, and rotations" (Massachusetts Department of Education, 2000, p. 64). This single Massachusetts standard encompasses fully six separate New York standards that are grouped under the heading "Students will apply coordinate geometry to analyze problem solving situations:"

- 8.G.7 "Describe and identify transformations in the plane, using proper function notation (rotations, reflections, translations, and dilations)"
- 8.G.8 "Draw the image of a figure under rotations of 90 and 180 degrees"
- 8.G.9 "Draw the image of a figure under a reflection over a given line"
- 8.G.10 "Draw the image of a figure under a translation"
- 8.G.11 "Draw the image of a figure under a dilation"
- 8.G.12 "Identify the properties preserved and not preserved under a reflection, rotation, translation, and dilation" (New York State Education Department, 2005, p. 86).

While not all cases were this extreme, we found that in general, the Massachusetts standards were broader, so the small number of standards in Massachusetts actually sampled more broadly from the domain than did the more numerous New York standards.

This difference in the breadth of standards was reflected in the allocation of items and raw score points on the tests in the two states. As detailed further below, we found that 40 to 50 percent of the raw score points on the New York eighth-grade tests in any given year mapped to only three Massachusetts eighth-grade standards.

The wording of standards may narrow content in subtle ways. For example, the Massachusetts standards include calculation of the volume of rectangular prisms but specify that the student will be given the formula:

> 10.M.2 *Given the formula*, find the lateral area, surface area, and volume of prisms, pyramids, spheres, cylinders, and cones, e.g., find the volume of a sphere with a specified surface area (Massachusetts Department of Education, 2000, p. 75, emphasis added].

Including the italicized clause fundamentally narrows what is being measured. The student does not need to know the mathematical relationship between volume and dimensions to solve items addressing this standard; she needs only the elementary-school skills of "plugging" numbers into an equation and multiplying. Figure 7.2 shows a tenth-grade MCAS item assessing this standard, along with an item assessing similar content from Singapore's Primary School Leaving Examination, which students take at the end of sixth grade. Because the formula is given in the tenth-grade MCAS (it is not shown in Figure 7.2 because it is not contiguous in the test booklet), one could discard any mention of volume or the particular shape and simply ask students to solve the equation. By way of contrast, sixth-graders in Singapore are not given the formula and are expected to know the mathematical relationship between dimensions and volume.

Insert Figure 7.2 about here

A simple listing of standards may understate the narrowing that can occur at this stage. The tests also create operational definitions of the standards, which may further narrow sampling from the domain. For example, items may predictably omit aspects of a standard that may be more difficult. The items in Figure 7.2 provide an example of this. The Massachusetts item (and the corresponding items in the New York tests, which are found in grade 6) require either calculating the total volume of a single prism or calculating one dimension given a full prism. In contrast, the Singapore item in Figure 7.2 requires comparison of two prisms and calculation of a fractional volume.

Although narrowing within standards in the writing of test items is discussed below, it is important to note here that the content of items may further narrow the span of the standards when items that purportedly measure two standards are so similar that they are effectively sampling the same content. For example, Figure 7.3 shows very similar items used to assess tenth-grade Massachusetts standards D.2 and P.2. These are not only different standards; they are also drawn from two different strands (Data Analysis and Interpretation, and Patterns and Algebra, respectively). Similar items assessing nominally different standards are found in the New York tests as well.

Insert Figure 7.3 about here

Selection of a subset of the standards (Box 3 in Figure 7.1)

Tests may create incentives for teachers to narrow instruction by omitting certain standards, providing the opportunity for teachers to exclude or deemphasize content without negatively affecting students' test scores. Table 7.1 shows that over the years 2006-2008, an average of 42 percent of New York eighth-grade standards, 67 percent of Massachusetts eighth-grade standards, and 60 percent of Massachusetts tenth-grade standards were tested.[2] That only a fraction of the state curriculum is tested in any given year would not enable score inflation so long as different standards are tested each year and teachers cannot predict which standards will be tested. For example, if the tests were based on a random sample of the standards each year, there would be no incentive for teachers to focus on a limited fraction of the curriculum.

Therefore, to assess whether sampling of standards was predictably incomplete, we examined the percentage of standards tested over the three-year period of the study, 2006-2008. Over this three-year period, 58 percent of relevant standards were sampled on the New York eighth-grade test, 83 percent were sampled on the Massachusetts tenth-grade test, and 90 percent were sampled by the Massachusetts eighth-grade test. All three tests created opportunities for reallocation by omitting content, although to varying degrees. However, these simple counts understate opportunities for inflation because they do not consider the relative emphasis the tests assign to the tested standards. If variations in emphasis—that is, test weights—are predictable, scores may become inflated if teachers emphasize highly sampled content, while decreasing emphasis on standards that are never (or only infrequently) sampled. Therefore, it is important to examine test weights in addition to omissions.

In both states, a small number of standards accounted for the majority of test points on the state tests. Figure 7.4 compares the distribution of standard weights on the three exams. The *x*-axis reports the number of standards ever actually assessed in the exams, sorted from the most to the least highly assessed. The *y*-axis indicates the percentage of points each standard is worth. The area below the step functions, therefore, corresponds to the percentage of test points covered by the number of standards at any point along the *x*-axis. The dotted drop lines indicate how many standards students must master to earn 50% of test points. On the New York eighth-grade test, about 50 percent of possible test points corresponded to only 10 of 48 standards. Similarly, on the Massachusetts eighth-grade math test, 8 of 29 standards account for a little more than 50 percent of test points, and on the tenth-grade test, 5 of thirty standards corresponded to about 50 percent of test points. The patterns are particularly stark for the Massachusetts tenth-grade mathematics tests, where the single standard with the highest test weight consistently contributed between 15 and 17 percent of test points per year, while the second highest weight standard consistently contributed between 7 and 17 percent of test points. To the extent that these weights represent consistent patterns of emphasis, teachers could profitably identify and teach to the cluster of standards that drives the highest number of test points, sacrificing instruction focusing on the others.

However, the extent of narrowing depends not just on the number of standards tested, but also on their breadth. As noted, the New York standards are generally narrower than those of Massachusetts, and a comparison between them should take this into account. To illustrate this point, Figure 7.5 represents the distribution of standards tested over 2006-2008 once the New York eighth-grade standards have been mapped onto those in Massachusetts. Here, we see that the uneven sampling of standards is much more dramatic on New York's eighth-grade test than on Massachusetts's; just 4 Massachusetts standards make up 50 percent of test points on the New York eighth-grade tests.

Coaching to high frequency items requires that teachers or students recognize the predicable patterns in the test. This is made easier when states release items that are similar to those that will be used in the future. In the period we studied, teachers in both states could download entire released forms that contain all the items used to calculate students' scores.[3] Moreover, in Massachusetts, teachers could download all the items in past years associated with a given standard. This type of sorting makes it easier for administrators and teachers to isolate aspects of their curriculum that are actually tested, as well as how and with what frequency performance on each standard is measured.

Insert Figures 7.4 & 7.5 about here

Selecting from within standards (Box 4 in Figure 7.1)

The previous section showed that standards are often unevenly sampled on state tests. When standards are framed broadly, tests may further narrow the sample by consistently excluding some skills from within a standard. This is more of a potential concern in Massachusetts than in New York because of the wording of standards. Many of New York's standards are written very narrowly, causing a narrowing of focus at the level of Box 3 and leaving less room for further narrowing at the level of Box 4. In contrast, most Massachusetts standards are written to include a cluster of skills, concepts, and representations, which creates a risk of score inflation if tests do not adequately sample from the skill set implied by the standards.

To evaluate how large a problem this was in Massachusetts, we examined items testing standards that contributed an average of 3 percent or more of test points per year in the years 2006-2008. We chose a 3 percent threshold because this represents sufficient weight to have a substantial effect on students' overall scores. We compared the test items to the specific standards they sampled, and noted which particular aspects of the standard the items sampled. (We did find a small number of examples of similar omissions on the New York tests even though the narrow framing of most standards made within-standard sampling less likely.)

We found evidence of incomplete sampling within standards on both the eighth- and tenth-grade Massachusetts tests. Four of the seven standards high-weight standards were incompletely sampled on the eighth-grade test. On the tenth-grade test, three of the six high-weight standards were incompletely sampled.

For example, eighth-grade standard 8.N.3 required students to "Use ratios and proportions in the solution of problems, in particular, problems solving unit rates, scale factors, and rate of change." Over this period, students were never asked to solve problems using the rate of change. In other cases, certain operations were excluded. Standard 8.N.12. requires students to "Select and use appropriate operations—addition, subtraction, multiplication, division, and positive integer exponents—to solve problems with rational numbers (including negatives)," but multiplication and positive integer exponents were never tested. In some cases, omission of a portion of a standard from items explicitly linked to that standard may not be problematic because of other items on the test. For example, the tenth-grade standard 10.D.1 required students to, “Select, create, and interpret as appropriate (e.g. scatterplot, table, stem-and-leaf plots, box-and-whisker plots, circle graph, line graph, and line plot) for a set of data….” The items linked to this standard omitted scatterplots, line graphs, and selection of appropriate representations, but these were sampled in the context of other items addressing other standards.

Selecting representations (Box 5 in Figure 7.1)

All of the previous steps entail narrowing of what is tested on state tests, relative to the domain about which inferences based on test scores are made. The final stage, the selection of representations, entails *how* that content is tested. As noted, we use the term representation broadly to refer not just to item format and the appearance of material in the test item, but also to the recurrent use of minor details of content that are incidental to the construct the item is intended to assess.

In extreme cases, representations are so similar that items are essentially clones of those used in earlier years. A particularly extreme example of clone items was found in the seventh-grade mathematics tests in New York. Two of the four items that appeared over a four-year period are shown in Figure 7.6. Note that the stems are essentially identical except for the substitution of “watermelon” for “cheese,” and the distracters are similar except for the doubling of rulers in the second year. One of the remaining two items was also a clone of these two. The fourth was similar but more complex, requiring students to recognize that the appropriate measure is grams rather than kilograms.

A more subtle form of recurrent presentation can be seen in the items testing New York standard 8.N.4, “Apply percents to: tax, percent increase/decrease, simple interest, sale price, commission, interest rates, gratuities." This was one of the two most frequently tested standards in the eighth-grade tests, tested with 12 items during the four years from 2006 through 2010. Of those 12 items, only one required students to calculate a rate or percentage. The other 11 provided a base quantity and a percentage and required that the student calculate either a change (e.g., a discount or a tip) or a final quantity (e.g., a total bill including the tip). The use of a consistent form narrows what students need to know and facilitates teaching mechanical solutions that do not require a solid understanding of rates or percentages.

Discussion

This study documents that the tests in New York and Massachusetts– and in particular, the New York eighth-grade and Massachusetts tenth-grade tests—do provide opportunities for narrow test preparation that could lead to inflated scores. We found multiple similarities in the types of opportunities afforded by both tests. Predictable sampling of standards and predictable representations of skills, were present on all tests, though their severity varied. We also identified test-specific opportunities that derived from the way the standards are framed in each state.

Substantial resources have been devoted to identifying these patterns—by state agencies, districts, and commercial firms—and anecdotal evidence suggests that some teachers are aware of them. For example, the Quincy, Massachusetts district website lists the test weights of each section of secondary mathematics textbooks used in the district and shows all of the relevant test items from a four year period (Quincy Public Schools, 2004). In an interview with the first author, a data coach in the Boston Public Schools provided pie charts illustrating the proportion of the total possible points on the test that could be attributed to each standard, which was used to decide what to emphasize in the classroom. As another data coach in the Boston Public Schools told us, “(The attitude is) just focus on the questions and content with the most potential to drive up scores. Why teach anything else?” In other cases, commercial test preparation resources alert teachers to recurrent patterns.

These predictable patterns in tests used for accountability create incentives to narrow instruction in ways that can deprive students of good instruction, disadvantage good teachers and inflate scores. When teachers, administrators and schools are evaluated based the gains their students demonstrate on these tests, they face strong disincentives to provide students with opportunities to master the full breadth and depth of math learning described in each state’s framework of standards. Thus, tests that provide substantial opportunities for score inflation increase the likelihood that performance evaluation systems based on test score gains will misidentify successful teachers and schools and perversely reward teachers who teach to the test, while penalizing teachers who attempt to instruct students in the entire domain outlined in state standards. As Shepard (1990) notes, schools or teachers that teach to the full framework of a curriculum and thus provide a broad, rich curriculum beyond the specifics of the tested sample might be at a disadvantage on a test used for accountability purposes. Because these teachers seek to provide students with a broader educational experience, they spend proportionally less time on the specific content that offers the greatest return in terms of test score gains.

This study provides only a first glimpse of the opportunities for inappropriate test preparation provided by current high-stakes tests. Ample research, however, shows that the results—inappropriate instruction and score inflation—are widespread and often severe. Improving test-based accountability will require more widespread evaluation of these opportunities, monitoring educators' responses, and redesigning of tests and the accountability systems in which these tests are embedded to reduce incentives to narrow instruction,

We close by pointing towards important areas for future research on equity, which is the focus of this book. What remains incompletely researched is whether historically disadvantaged groups are more likely to receive test-specific instruction, and to what extent score inflation affects the measurement of between-group inequality in outcomes. To the extent that poor and minority students are concentrated in the lowest-scoring schools that face the greatest pressure to raise scores, they may be the most severely affected (e.g., Rouse et al., 2007). Nonetheless, many policymakers and educators continue to view test score data as accurate measurements that track student performance as well as racial and socioeconomic inequities. While test scores are intended to provide more transparency about student performance, this chapter demonstrates the ways that predictability in test design supports test-specific instruction. Any score inflation caused by test-specific instruction interferes with our ability to accurately measure educational inequality.

References

Carnoy, M., & Loeb, S. (2002). Does external accountability affect student outcomes? A cross-state analysis. *Educational Evaluation and Policy Analysis 24*, 305-331.

Center on Education Policy (2008). *Has student achievement increased since 2002? State test score trends through 2006-07*. Washington, DC: author.

Darling-Hammond, L., & Wise, A.E. (1985). Beyond standardization: State standards and school improvement. *The Elementary School Journal*, *85*, 315-36.

Figlio, D., & Getzler, L. (2002). Accountability, ability and disability: Gaming the system. *National Bureau of Economic Research Working paper 9307*. Retrieved from: http://www.nber.org/papers/w9307

Fuller, B., Gesicki, K., Kang, E., & Wright, J. (2006). *Is the No Child Left Behind Act working? The reliability of how states track achievement.* University of California, Berkeley: Policy Analysis for California Education. Retrieved on October 11, 2009 from: http://www.eric.ed.gov/PDFS/ED492024.pdf

Gabriel, T. (2010). Under pressure, teachers tamper with tests. *The New York Times*, June 10. Retrieved July 8, 2010. http://www.nytimes.com/2010/06/11/education/11cheat.html?_r=2&src=mv

Grissmer, D., & Flanagan, A. (1998). *Exploring rapid achievement gains in North Carolina and Texas: Lessons from the states.* Washington, DC: National Education Goals Panel.

Hamilton, L. S., & Stecher, B. M. (2006). *Measuring instructional responses to standards-based accountability*. Santa Monica, CA: RAND Corporation.

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.

Hamilton, Laura S. and Brian M. Stecher. 2007. *Measuring Instructional Responses to Standards-Based Accountability.* Santa Monica, CA: RAND Corp.

Hamilton, L. S., Stecher, B. M., Marsh, J. A., McCombs, J. S., Robyn, A., Russell, J., et al. (2007). *How educators in three states are responding to standards-based accountability under No Child Left Behind. Research Brief.* Santa Monica, CA: RAND Corporation.

Hanushek, E. A., & Raymond, M. E. (2004). The effect of school accountability systems on the level and distribution of student achievement. *Journal of the European Economic Association, 2*, 406-415.

Hanushek, E.A., & Raymond, M.E. (2005). Does school accountability lead to improved student performance? *Journal of Policy Analysis and Management*, *24*(2), 297-327.

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

Jacob, B. A. (2005). Accountability, incentives, and behavior: Evidence from school reform in Chicago. *Journal of Public Economics*. 89, 761-796.

Jacob, B. A. (2007). *Test-based accountability and student achievement: An investigation of differential performance on NAEP and State Assessments.* NBER Working Paper w12817.

Jennings, J. L., & Beveridge, A. A. (2009). How does test exemption affect schools' and students' academic performance? *Educational Evaluation and Policy Analysis, 31*, 153-175.

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

Klein, S.P., Hamilton, L.S., McCaffrey, D.F., & Stecher, B.M. (2000). *What do test scores in Texas tell us?* (Issue Paper IP-202). Santa Monica, CA: RAND. Retrieved on November 15, 2011, from http://www.rand.org/publications/IP/IP202/

Koretz, D. (1986). *Trends in educational achievement*. Washington, D.C.: Congressional Budget Office, April.

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

Koretz, D. (2010). *Implications of current policy for educational measurement.* Princeton, N.J.: Center for K-12 Management and Performance Assessment, Educational Testing Service http://www.k12center.org/rsc/pdf/KoretzPresenterSession3.pdf

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

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

Koretz, D., Mitchell, K., Barron, S., & Keith, S. (1996). *The perceived effects of the Maryland School Performance Assessment Program (CSE Tech. Rep. No. 409)*. Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Koretz, D., & Hamilton, L. S. (2003). *Teachers' responses to high-stakes testing and the validity of gains: A pilot study*. CSE Technical Report 610. Los Angeles: Center for the Study of Evaluation, University of California.

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

Koretz, D., Linn, R. L., Dunbar, S. B., & Shepard, L. A. (1991). The effects of high-stakes testing: Preliminary evidence about generalization across tests. In R.L. Linn (chair), *The effects of high-stakes testing*. Symposium presented at the annual meetings of the American Educational Research Association and the National Council on Measurement in Education, Chicago, April.

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

Koretz, D., Mitchell, K., Barron, S., & Keith, S. (1996b). *The perceived effects of the Maryland School Performance Assessment Program.* CSE Technical Report No. 409. Los Angeles: Center for the Study of Evaluation, University of California.

Lee, J. (2007). *The testing gap: Scientific trials of test-driven school accountability systems for excellence and equity*. Charlotte, NC: Information Age Publishing.

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

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

Massachusetts Department of Education. (2006). *MCAS Question Search*. Retrieved January 18, 2008 and November 10, 2011 from http://www.doe.mass.edu/mcas/search/.

McNeil, L. M. (2000). *Contradictions of school reform: The educational costs of standardized testing*. London: Routledge.

National Research Council. (2011). *Incentives and test-based accountability in education*. M. Hout and S. W. Elliott (Eds.) Committee on Incentives and Test-Based Accountability in Public Education, Board on Testing and Assessment. Washington, DC: The National Academies Press, 2011.

New York State Education Department (2005). *Mathematics Core Curriculum*. Albany: The University of the State of New York.

New York State Department of Education. *(2008). New York State Testing Program: Grade 7 mathematics test. Retrieved from http://nysedregents.org/Grade7/Mathematics/20080306book1.pdf*

New York State Department of Education. *(2009). New York State Testing Program: Grade 7 mathematics test. Retrieved from http://nysedregents.org/Grade7/Mathematics/20090309book1.pdf*

Pedulla, J. J., Abrams, L. M., Madaus, G. F., Russell, M. K., Ramos, M. A., Miao, J., et al. (2003). *Perceived effects of state-mandated testing programs on teaching and learning: Findings from a national survey of teachers*. Chestnut Hill, Massachusetts: National Board on Educational Testing and Public Policy. Retrieved July 15, 2010. http://www.bc.edu/research/nbetpp/statements/nbr2.pdf

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

Quincy Public Schools (2004). *MCAS Math Concordance*. Quincy, MA: author. Last retrieved on January 23, 2012 from http://66.241.201.139/powerpoint/MCAS/Math%20Concordance%20-%20Show.ppshttp://66.241.201.139/powerpoint/MCAS/Math%20Concordance%20-%20Show.pps

Ravitch, D. (2010). Is education on the wrong track? *A TNR Symposium*. Retrieved March 16, 2010. http://www.tnr.com/article/politics/education-the-wrong-track-0

Romberg, T. A., Zarinnia, E. A., & Williams, S. R. (1989). *The influence of mandated testing on mathematics instruction: Grade 8 teachers' perceptions*. Madison, WI: University of Wisconsin, Center for Educational Research, School of Education, and Office of Educational Research and Improvement of the United States Department of Education.

Rouse, C.E, Hannaway, J., Goldhaber, D., & Figlio, D. (2007) *Feeling the Florida heat? How low-performing schools respond to voucher and accountability pressure.* CEPS Working Paper No. 156 November 2007.

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

Shepard, L.A. (1990). "Inflated test score gains: Is the problem old norms or teaching the test?" *Education Measurement: Issues and Practice*, 9, p. 15-22.

Shepard, L. A., & Dougherty, K. D. (1991). The effects of high stakes testing. In R. L. Linn (Ed.), *Annual meetings of the American Education Research Association and the National Council of Measurement in Education*. Chicago, IL.

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

Smith, M. L., & Rottenberg, C. (1991). Unintended consequences of external testing in elementary schools. *Educational Measurement: Issues and Practice*. 10, 7-11.

Stecher, B. (2002). Consequences of large-scale, high-stakes testing on school and classroom practice. In L. Hamilton, et al., *Test-based accountability: A guide for practitioners and policymakers*. Santa Monica: RAND. http://www.rand.org/publications/MR/MR1554/MR1554.ch4.pdf

Stecher, B. M., Chun, T.J., Barron, S.I, & Ross, K.E. (2000). *The effects of the Washington State Education Reform on schools and classrooms: Initial findings*. Santa Monica, CA: RAND Corporation.

Figures

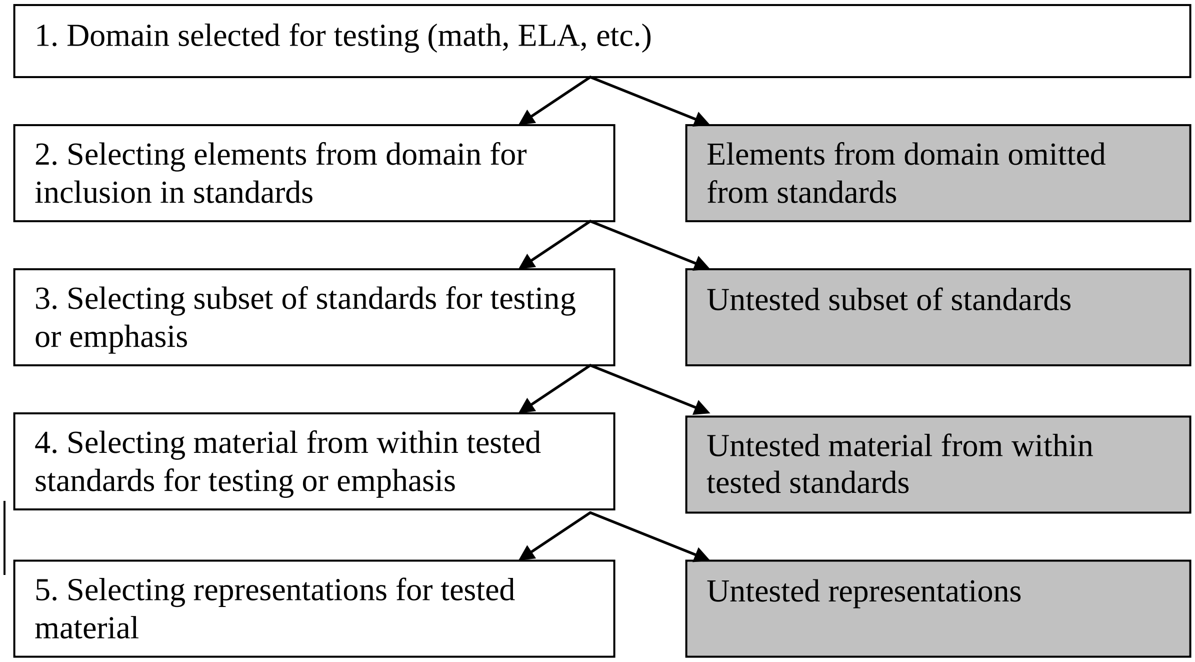

1. Domain selected for testing (math, ELA, etc.)
2. Selecting elements from domain for inclusion in standards
Elements from domain omitted from standards
3. Selecting subset of standards for testing or emphasis
Untested subset of standards
4. Selecting material from within tested standards for testing or emphasis
Untested material from within tested standards
5. Selecting representations for tested material
Untested representations

Figure 7.1. Taxonomy of Opportunities for Score Inflation.

NOTE: “tested” vs. “untested” also represents emphasized vs. de-emphasized.

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

33 A right rectangular prism has the following dimensions:

- The height is 5 feet.
- The length is 6 feet.

The volume of the prism is 60 cubic feet. What is the width of the prism?

A. 2 feet

B. 3 feet

C. 5 feet

D. 6 feet

b. 6th-grade Singapore Primary School Leaving Exam: no formula given

In Figure 1, Tank A is completely filled with water and Tank B is empty. Water is poured from Tank A into Tank B without spilling. The heights of the water level in the two tanks are now equal as shown in Figure 2.

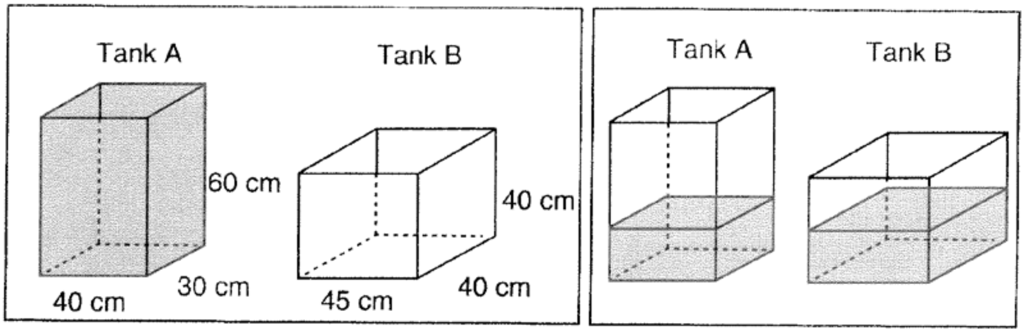


Figure 1

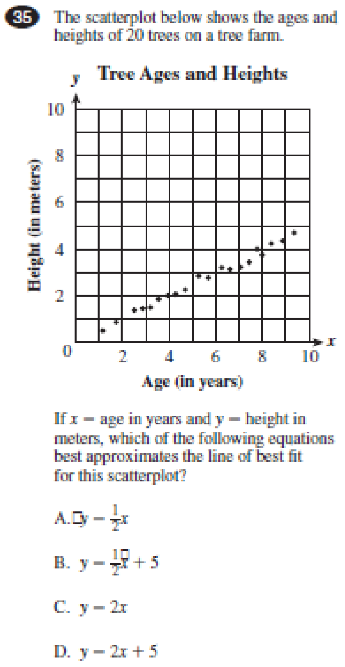
Figure 2

What is the height of the water level in Tank A in Figure 2?

Answer ________________

Figure 7.2. Comparison of volume items from 10th-grade MCAS and Singapore 6th-grade tests. Sample MCAS retrieved from MA Department of Education's MCAS Question Search (2011). http://www.doe.mass.edu/mcas/search/ Sample Singapore item is from Singapore Examinations and Assessment Board (2009).

D.2 (Data Analysis and Interpretation)



35 The scatterplot below shows the ages and heights of 20 trees on a tree farm.

Tree Ages and Heights
y
10
8
6
4
2
0
Height (in meters)
2
4
6
8
10
x
Age (in years)

If $x = $ age in years and $y = $ height in meters, which of the following equations best approximates the line of best fit for this scatterplot?

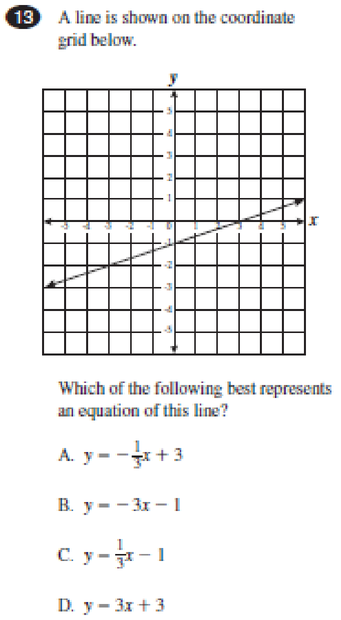
A.□$y = \frac{1}{2}x$

B. $y = \frac{1\square}{2}x + 5$

C. $y = 2x$

D. $y = 2x + 5$

P.2 (Patterns and Algebra)



13 A line is shown on the coordinate grid below.

y
x
5
4
3
2
1
0
-1
-2
-3
-4
-5

Which of the following best represents an equation of this line?

A. $y = -\frac{1}{3}x + 3$

B. $y = -3x - 1$

C. $y = \frac{1}{3}x - 1$

D. $y = 3x + 3$

Figure 7.3. Similar 10th-grade MCAS items used to sample nominally different standards. Item #35 on MCAS 10th-grade test in 2005 assesses standard D.2 by having students determine the equation from a linear scatterplot. Item #13 on MCAS 10th-grade test in 2006 assesses standard P.2 by having students determine the equation from a line. Retrieved from MA Department of Education’s MCAS Question Search (2008). http://www.doe.mass.edu/mcas/search/

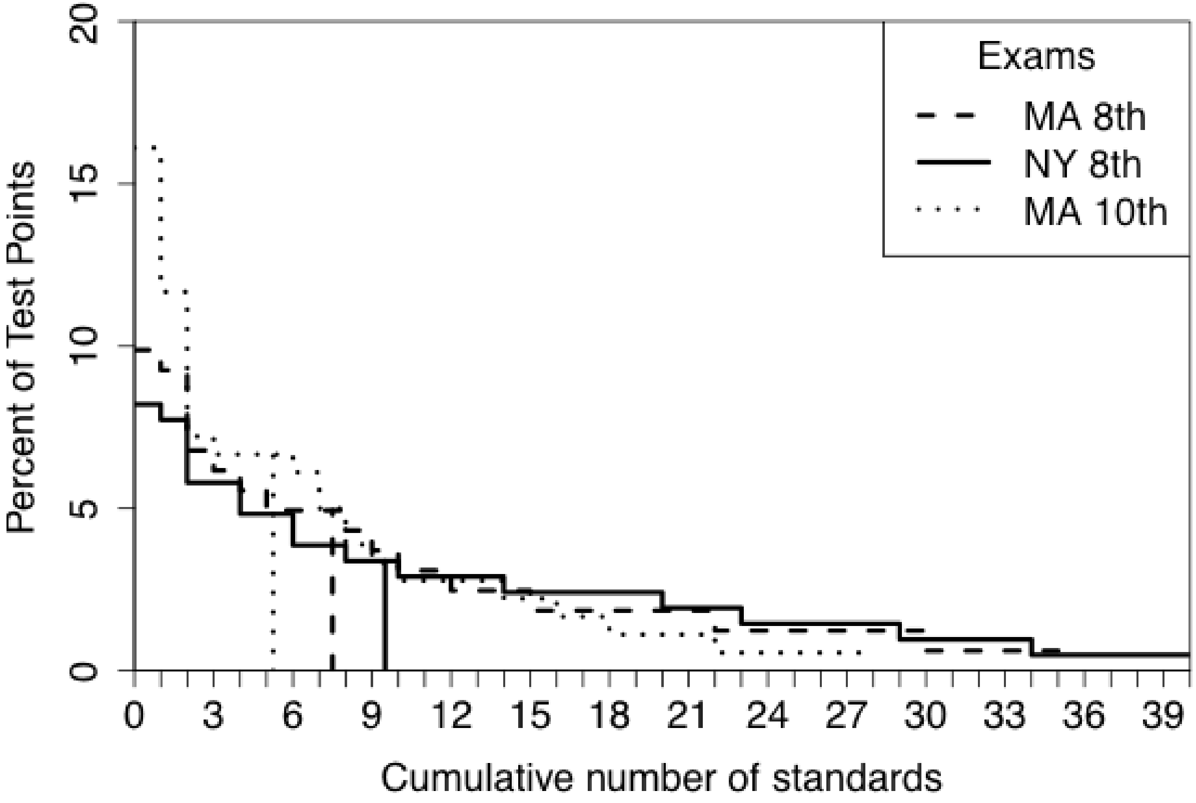


Figure 7.4. Distribution of standards tested on NY and MA Mathematics Tests

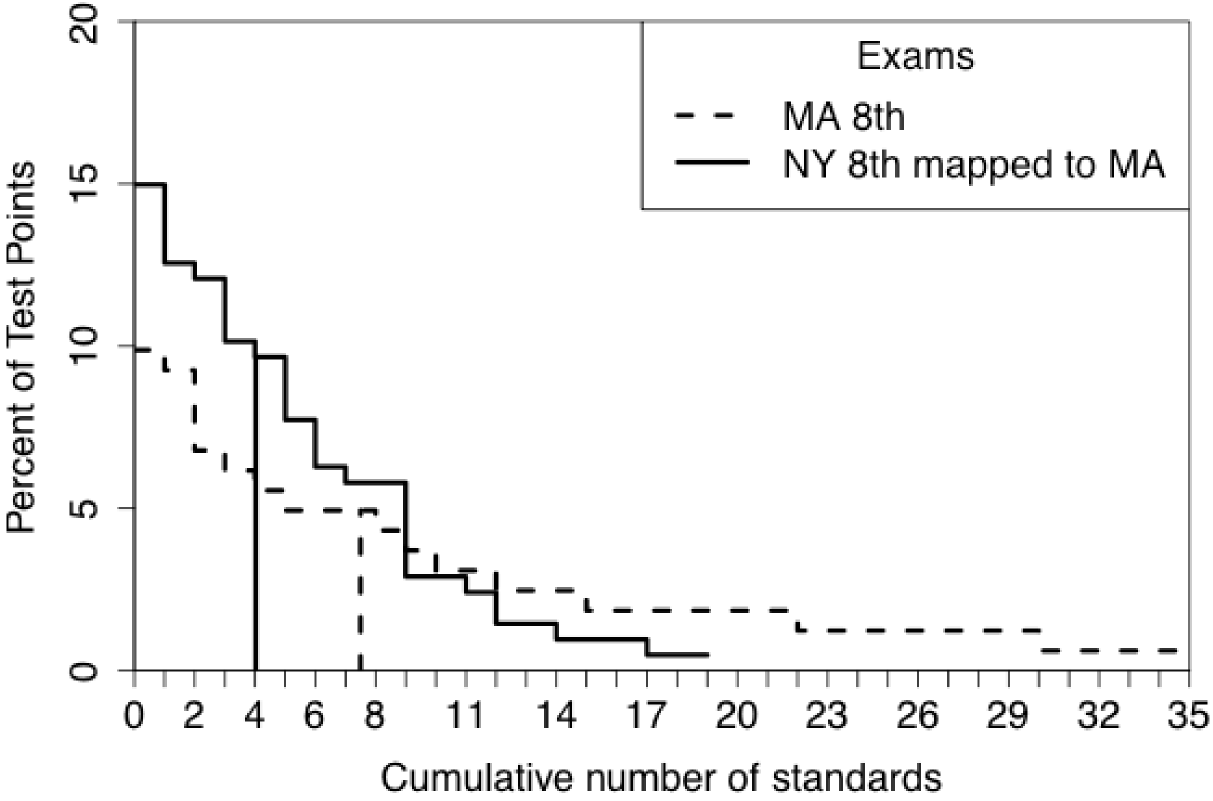
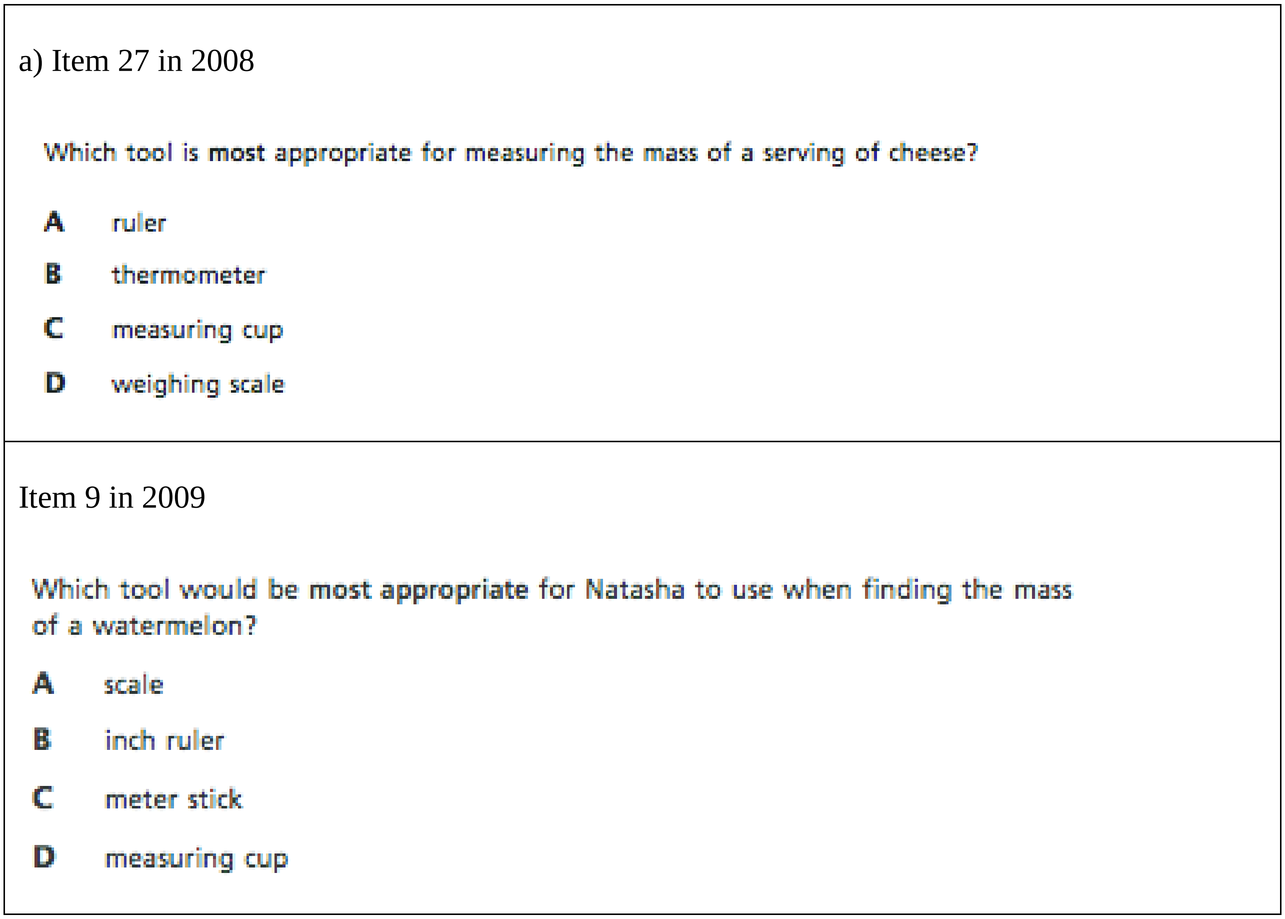


Figure 7.5. Distribution of Standards Tested on NY and MA 8th-Grade Mathematics Tests (NY standards mapped to MA standards)



a) Item 27 in 2008

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

A ruler

B thermometer

C measuring cup

D weighing scale

Item 9 in 2009

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

A scale

B inch ruler

C meter stick

D measuring cup

Figure 7.6. Clone items assessing NY's standard 7.M.9: "Determine the tool and technique to measure with an appropriate level of precision: mass." (New York State Department of Education, 2008, p.21 and 2009, p.11).

Tables

Table 7.1. Fraction of Total NY and MA Standards Tested by Year and Over 2006-2008

Set of Standards	Average per year	Ever, 2006-2008
New York eighth-grade standards	42	58
Massachusetts eighth-grade standards	67	90
Massachusetts tenth-grade standards	60	83

Note. Number of standards tested in parentheses. MA excludes out of grade standards that are sampled on the test, because these are extraneous to the validity of inferences about student mastery of the standards in the tests' corresponding grade level framework.

Notes:

[1] A majority of states have committed to adopting the Common Core Standards, and this could provide a more widely accepted operational definition of the target of inference. However, Porter et al. (2011) reported that alignment between the Common Core and NAEP is only modest, albeit a bit higher than the average alignment between the Common Core and the average state's standards. The disparity between NAEP and the Common Core is not yet widely understood, and it remains unclear whether educators, policymakers, and others will decide that NAEP content omitted from the Common Core is relevant to the domain.

[2] Because the New York state math test was administered in March during this time period, New York identified a subset of standards that were eligible for inclusion in the state test. These standards included seventh-grade standards scheduled to be taught after March of the previous year, as well as eighth-grade standards taught before March. All told, 68.8 percent of the eighth-grade standards are eligible for testing. In addition, the eighth-grade test includes standards taught in the seventh-grade curriculum after the March test. To provide comparability across states, we focus on the percent of the total 8th standards that were tested each year. Were we to use a denominator that includes seventh-grade post-March standards and eighth-grade pre-March standards, 83 percent of standards were covered between 2006-2008.

[3] Beginning in 2011, New York stopped releasing test forms. In most grades and subjects, Massachusetts began releasing only half of the common items in 2009.