Working Paper (14 Nov 2013)

Auditing for Score Inflation using Newly Tested Standards

A working paper of the Education Accountability Project at the Harvard Graduate School of Education
http://projects.iq.harvard.edu/eap

Daniel Koretz
Harvard Graduate School of Education

Carol Yu
Harvard Graduate School of Education

David Braslow
Harvard Graduate School of Education

Acknowledgements: The research reported here was supported by the Institute of Education Sciences, U.S. Department of Education, through Grant R305AII0420, to the President and Fellows of Harvard College. The authors also thank the New York State Education Department for providing the data used in this study. The opinions expressed are those of the authors and do not represent views of the Institute, the U.S. Department of Education, or the New York State Education Department or its staff.

Abstract

Score inflation is a well-known threat to validity under high-stakes conditions. Koretz & Beguin (2010) noted the weaknesses of evaluating score inflation using external tests and suggested instead using self-monitoring assessments (SMAs), which incorporate audit items that are sufficiently novel that they are not susceptible to test preparation aimed at more predictable items. This study investigated whether items selected only to assess previously untested standards can contribute to an audit component in a high-stakes test, using data from New York State's 2011 mathematics tests in grades 4, 7, and 8. Despite a severe conservative bias created by a number of aspects of the study design, we found that the audit component functioned as expected in two of three grades, although more weakly in one. The items did not function as an audit in the remaining grade. We discuss factors that may have contributed to the variation across grades. The findings suggest that items testing previously untested standards can contribute to an audit. However, they also indicate that merely testing previously untested standards is not sufficient to make items useful for this purpose. These findings underscore the need for additional research investigating the optimal characteristics of items used for auditing gains.

Introduction

Test-based accountability is the cornerstone of education policy in the United States. Beginning with the minimum-competency movement of the 1970s and the education reform movement of the 1980s, successive waves of reform have increased the pressure on educators to raise test scores. Many state programs in the 1990s established school-level rewards and punishments based on test scores. The enactment of No Child Left Behind in 2001 made this approach national policy (Koretz & Hamilton, 2006). The current Race to the Top program has moved the focus of evaluation from entire schools to individual teachers, increasing pressure yet further.

A substantial body of research has shown that these programs often induce various forms of inappropriate test preparation, much of which focuses attention unduly on the specifics of the test used for accountability rather than the broader domain of content and skills that the test is intended to represent (e.g., Stecher, 2002). One consequence of these behaviors is score inflation, that is, score increases substantially larger than the improvements in achievement that they are taken to measure. Studies have shown that the resulting bias can be very large (e.g., Jacob, 2007; Klein, Hamilton, McCaffrey, & Stecher, 2000; Koretz & Barron, 1998).

Most studies of score inflation have followed a single design. To be meaningful, increases in scores must generalize to the domain from which the test samples. If performance gains do generalize to the domain, they should show a reasonable degree of generalization to other tests that sample from the same domain and are intended to support similar inferences. Accordingly, most studies examine the consistency of gains

between a high-stakes test and a lower-stakes audit test, most often the National Assessment of Educational Progress (NAEP).

Koretz & Beguin (2010) noted numerous disadvantages of this approach. A suitable audit test may be unavailable or, like NAEP, may be available only for a few grades or only at high levels of aggregation. The substantive appropriateness of available audit tests may be arguable, as they may have been designed to support somewhat different inferences. Motivational differences may bias comparisons between the audit and high-stakes tests. Sample-based audit tests can create problems if samples differ over time.

As way to avoid these limitations, Koretz & Beguin (2010) suggested *self-monitoring assessments* (SMAs). SMAs incorporate audit components into the high-stakes test itself, using differences in performance between these audit components and routine operational items as a measure of score inflation. A first trial of an SMA was conducted using a stand-alone field test of New York State tests (Ng et al., 2013). This paper reports the first SMA conducted in an operational form of a high-stakes test, using data from New York State's 2011 mathematics tests in grades 4, 7, and 8.

The Problem of Score Inflation

The risk of score inflation has been noted in the measurement literature for more than half a century. For example, Lindquist (1951), writing in an era of low-stakes testing, noted that

> Because of the nature and potency of the rewards and penalties associated in actual practice with high and low achievement test scores of students,

> the behavior measured by a widely used test tends in itself to become the real objective of instruction, to the neglect of the (different) behavior with which the ultimate objective is concerned (p. 153).

Madaus (e.g., 1988) and Shepard (e.g., 1988) both warned that high-stakes testing was likely to create problems of score inflation.

The first empirical study of score inflation (Koretz, Linn, Dunbar, and Shepard, 1991) examined a system that, while high-stakes by the standards of the day, was very low-stakes by today's standards, entailing no concrete sanctions or rewards. Koretz et al. examined changes in scores when the district substituted one commercially produced achievement test for another, and they readministered the older test four years after its final high-stakes administration. They found inflation in mathematics of half an academic year by the end of third grade. Since that time, studies in a variety of different contexts have confirmed that score inflation is common, although not ubiquitous (e.g., Hambleton, et al., 1995; Haney, 2000; Ho, 2009; Ho & Haertel, 2006; Jacob, 2005, 2007; Klein, Hamilton, McCaffrey, & Stecher, 2000; Koretz & Barron, 1998; Koretz, Linn, Dunbar & Shepard, 1991). For example, Klein, Hamilton, McCaffrey, & Stecher (2000) found that gains on the high-stakes Texas TAAS test were roughly two to six times the size of the state's gains on NAEP. Jacob (2007) found that gains on several state high-stakes mathematics tests were roughly double those on NAEP. Koretz & Barron (1998) found that gains on Kentucky's high-stakes mathematics tests exceeded gains in NAEP roughly by a factor of four, and Hambleton et al. (1995) found that gains of roughly three-fourths

of a standard deviation on the state's fourth-grade reading test were accompanied by no gain whatever on NAEP.

Test Preparation that May Generate Score Inflation

A number of studies have examined educators' responses to testing and have found behaviors that may be linked to score inflation. Among the reported responses that may be related to inflation are narrowing of instruction, adapting instruction to the format of test items, focusing instruction on incidental aspects of tests, and cheating (e.g., Stecher, 2002). For example, in a comprehensive study of responses to test-based accountability in three states, Hamilton et al. (2007) found that 55 to 78 percent of teachers reported "emphasizing assessment styles and formats of problems," and roughly half reported spending more time teaching test taking strategies (p. 103). Such practices have been widely documented (e.g., Abrams, Pedulla, & Madaus, 2003; Luna & Turner, 2001; Pedulla et al., 2003; Stecher & Barron, 2001; Stecher et al., 2008).

Koretz & Hamilton (2006) distinguished among seven different types of test preparation, both desirable forms that are likely to produce meaningful gains in achievement and undesirable forms that are likely to produce score inflation. Educators may respond to the pressures to raise scores by allocating more time to instruction, finding more effective instructional strategies, or simply working harder. All of these, within limits, may produce meaningful gains in scores, that is, gains that reflect commensurate increases in student learning. At the other extreme, educators may cheat, as in the recent large-scale scandal in the Atlanta public schools (Severson, 2011), which can only produce inflated scores. Similarly, they may manipulate the population of test-

takers, which will not bias the scores of individual students but will inflate aggregate scores (e.g., Figlio & Getzler, 2006).

More relevant here are two categories of response that Koretz & Hamilton (2006) labeled *reallocation* and *coaching*. Reallocation entails better aligning instructional resources, such as time, to the content of the specific test used for accountability. Coaching refers to focusing instruction on minor, usually incidental characteristics of the test, including unimportant details of content and the types of presentations used in the items in the specific test. For example, item writers typically use Pythagorean triples in items assessing the Pythagorean Theorem because students are unable to compute square roots by hand. One common form of coaching in response to this is telling students to memorize the two Pythagorean triples that most often appear in test items, 3:4:5 ($3^2 + 4^2 = 5^2$) and 5:12:13 (e.g., Rubinstein, 2000). This allows students to answer the item correctly without actually learning the theorem or being able to apply it in real life.

Reallocation can be either desirable or not. For example, if a test reveals that a school's students are weak in proportional reasoning, one would want educators to bolster instruction in that area, and these responses might include increasing the allocation of time to it. This is desirable reallocation, because if it is effective, it will improve the mathematics achievement the test score is intended to proxy. However, if reallocation entails shifting resources away from content that is de-emphasized or omitted from the specific test but is nonetheless important to the inference based on scores, the result will be score inflation. Numerous studies have found that many teachers report decreasing their emphasis on elements of the curriculum that are de-emphasized by the

test (Pedulla et al., 2003; Stecher et al., 2002) and widely available test-preparation materials help educators do so (Haney, 2000; Stecher et al., 2002). Coaching is somewhat less ambiguous than reallocation. While there are cases in which coaching might induce meaningful gains—for example, if students are confronted with a format that is so novel that their performance would initially be biased downward without some familiarization—for the most part, it will either waste time or bias scores upwards.

Variations in test preparation and score inflation

Although most studies of score inflation and related inappropriate test preparation have examined only trends in the student population as a whole, a growing body of evidence suggests that these problems tend to be more severe in schools that are low-achieving and that disproportionately serve low-income and minority students. This is not surprising, as low-performing schools face the most severe pressure to raise scores, face more severe obstacles to improving performance, and often have a lower-quality instructional staff.

Research has shown that teachers in more disadvantaged schools are likely to focus more than others on test preparation. These schools often have a stronger emphasis on assigning drills of test-style items and teaching test taking strategies (Cimbricz, 2002; Diamond & Spillane, 2004; Eisner, 2001; Firestone, Camilli, Yurecko, Monfils, & Mayrowetz, 2000; Herman & Golan, 1993; Jacob, Stone, & Roderick, 2004; Jones, Jones, & Hargrove, 2003; Ladd & Zelli, 2002; Lipman, 2002; Luna & Turner, 2001; McNeil, 2000; McNeil & Valenzuela, 2001; Taylor et al., 2002; Urdan & Paris, 1994). If such pedagogical practices result in score inflation, we would expect disadvantaged

students and schools to have relatively more score inflation than their advantaged counterparts, resulting in exaggerated reports of educational improvement for disadvantaged students.

Research on variations in score inflation is less abundant but is for the most part consistent with the research on behavioral responses in showing greater inflation in disadvantaged schools. Several reports show that the gains made by low-income and minority students relative to white students on state tests are not matched on audit tests (Klein et al., 2000; Jacob, 2007; Ho & Haertel, 2006). The link between pedagogical practices and score inflation has been further elaborated by Shen (2008), who examined differences in the performance of items that were more or less "teachable." Shen showed that schools had greater improvements over time in performance on more teachable items, and that this trend was more pronounced in disadvantaged schools. These limited studies suggest that the variations in test preparation noted above tend to produce the greatest score inflation for the most vulnerable students.

Another practice of concern is the reallocation of resources toward students whose anticipated scores are just under the proficiency cut score ("bubble students") and away from students whose anticipated scores are comfortably above or below proficient, a practice described as educational triage (Booher-Jennings, 2005; Gillborn & Youdell, 2000; Neal & Schanzenbach, 2010; Stecher et al., 2008). Accountability systems that only provide rewards or sanctions based on the percentage of students who score above the proficiency cut score incentivize teachers and schools with limited resources to focus those resources in ways that seem more likely to result in the greatest reward. If students

far beneath the cut score would require more resources than are available to get above the cut score, and if students comfortably above the cut score require fewer resources to demonstrate proficiency, then teachers may see that focusing their pedagogical effort on the bubble students may result in a larger reward. This phenomenon is particularly important in the present context because it creates particularly clear incentives to focus on test preparation for students near the cut score.

Predictable Sampling and the Design of SMAs

Successive operational test forms typically show predictable patterns. These patterns can include the amount of emphasis given to different content (including predictable omissions of content) and the ways in which content is presented. These predictable patterns in turn provide the opportunity to narrow test preparation to focus on the particulars of the test—that is, the opportunity for the behaviors that can generate score inflation. Test preparation materials often focus on these predictable patterns.

Holcombe, Jennings, & Koretz (2013) provided a framework for identifying the various types of predictable patterns that can enable inflation. They note that there are successive levels of narrowing in the design of a test. These are logically sequential, although they may not be carried out entirely in this order. First, one decides which elements from the domain will be represented in the standards or curriculum. Second, one decides which of the standards will be tested, and among those that will be tested, how frequently they will be tested and how much emphasis they will be given in a typical form. When standards are reasonably broad, one then has to decide how to sample from

the range of content and skills each standard implies. All of these stages of sampling represent a narrowing of the substantive range of the test.

The narrowing entailed in designing a test also includes predictable non-substantive elements, that is, elements that are not relevant to the intended inference. Many of these predictable patterns may be inadvertent. Some small details of content may be of this sort—for example, using only regular polygons in geometry items, or using only positive slopes in quadrant 1 in items about slopes, when the inference does not call for this narrowing. However, many of the predictable non-substantive elements in a test can be seen as aspects of presentation rather than content, e.g., item format, the particular graphics used with mathematics items, and so on.

Audit items in an SMA should assess material that is relevant to the inference, but without replicating the predictable patterns that create opportunities for inappropriate test preparation. For example, if all operational items assessing knowledge of the Pythagorean Theorem make use of common Pythagorean triples, a suitable audit item might be a calculator item with a non-integer solution. Students whose test preparation focused in Pythagorean triples would find the audit item much more difficult than would students receiving appropriate instruction on the Pythagorean Theorem.

The New York State Education Department (NYSED) offered us an opportunity to conduct a first pilot test of an SMA in the context of a stand-alone field test administered statewide in the spring of 2011. SMA items were administered in mathematics in grades 4, 7, and 8. In principle, all of the stages of narrowing described by Holcombe et al. (2013) could be represented in audit items, but we found that we

could not reliably code the difference between unimportant content details and aspects of presentation. For example, presenting slopes only as positive in quadrant 1 could be seen either as a detail of content or as a matter of presentation. Therefore, in this first SMA pilot, Ng et al. (2013) adapted this framework to generate four categories of audit items, combining representations with fine details of content. *Not-in-standards* (NIS) items represented content omitted from the state's standards but often included in the definition of the domain, e.g., in the NAEP frameworks. *Untested-standards* (US) items assessed state standards not previously tested. In cases in which the wording of standards appeared to narrow unduly a portion of the domain, Ng et al. (2013) administered *broadened at the standards level* (BS) items that very modestly broadened the range of the standard. Finally, some narrowing arose in the writing of operational items to represent a given standard. *Broadened at the item level* (BI) items broadened the within-standard diversity of test items, while adhering to the specific wording of the standards. Ng et al. (2013) designed the BI and BS items to match specific operational or anchor items included in the field test.

The results of the pilot study indicated that the SMA design can identify score inflation. However, those data had several important limitations. First, embedding the audit items in a stand-alone field test created a risk of bias from motivational effects, particularly given that audit items tended to be less familiar and more difficult. Second, the audit items were administered in the context of a complex matrix-sampled design, resulting in small sample sizes. Finally, the distribution of raw scores on the non-audit items was badly right-censored.

The 2011 Operational Tests and the Opportunity for an Audit

In the fall of 2009, the New York State Education Department (NYSED) publicly acknowledged that gains on some of the state's high-stakes test greatly outpaced gains on the NAEP and that this disparity suggested score inflation (Tisch, 2009). The standardized mean increase on the state test was seven times as large as the increase on the NAEP in grade 8 and three times as large in grade 4. (See Figure 1, which includes the state test results for grade 7, in which there is no NAEP, for reasons described below.)

In response, NYSED began taking steps to lessen the problem of score inflation. One step was to instruct the state's testing contractor to broaden the state's tests, partly to make them less predictable. The tests administered in the spring of 2011 were longer than those administered previously and incorporated items assessing standards that had not previously been tested.

This broadening of the assessment provided an opportunity for the SMA reported here, in which the items assessing previously untested standards served as the audit component. This study offered three advantages over the pilot reported by Ng et. al (2013). First, administering the audit items in the context of the operational high-stakes assessment eliminated the risk of motivational biases. Second, because the audit items were in the operational forms, the sample sizes approached 200,000. Third, because of changes made to the operational test in 2011, the distribution of raw scores on the non-audit component of the test was not as badly censored as had been the distribution of 2011 field test scores.

Nonetheless, the data used here have two important limitations. First, the audit is limited to one of the four categories of audit items administered by Ng et al. (2013). Second, none of the items in this study were specifically selected to serve as audit items, and they may be poorly suited for that purpose. For example, they may share attributes with other items that allow the effects of coaching to generalize to them.

These two limitations create a severe conservative bias—that is, they increase the risk that the audit will fail to identify real variations in score inflation. Therefore, this study evaluates the feasibility of using US items not designed for auditing as a component of an audit test, but the data are not sufficient to quantify inflation.

Data

Sample

In this study, we analyzed data provided by the New York State Education Department (NYSED) that contain information on all New York State public school students who took the New York State fourth, seventh, and eighth grade mathematics tests in 2011. Although there is no NAEP that could be used to audit the gains on the state test in grade 7, the implausibly rapid gain in mean scores on the state test—three-fourths of a standard deviation in only three years, roughly 30% larger than the gain in grade 8—strongly suggests score inflation and made this grade appropriate for this study. The dataset contains demographic data and both current- and prior-grade item responses. Students are also linked to their schools and districts, but not to their teachers. In our analysis, we focus primarily on the data for students in grade seven and then contrast results from the other grades.

Our final seventh-grade analytic sample consists of 93% of the original data provided by NYSED, with a total of 185,522 students nested within 1342 schools. In order to classify students by their prior performance, we merged the seventh-grade 2011 scores with all sixth-grade test scores from 2010. Deleting records that were missing either 2010 sixth-grade or 2011 seventh-grade scores resulted in the loss of about seven percent of our sample. In addition, we excluded a small number (less than 0.1 percent) of students with apparently anomalous scores.[1] The analytic sample differed little from the full sample in terms of demographic and other characteristics (Table 1).

Table 1 about here

New York City (NYC), which includes 34 percent of all tested seventh-graders in the original data, differs from the rest of the state in both its educational accountability system and its demographics. For example, a large majority of students in New York City are minority and from economically disadvantaged households, whereas the majority of the students from the rest of the state are white and not from economically disadvantaged households (Table 1). Accordingly, some of our analysis differentiates between NYC and the rest of the state. After the exclusions noted above, 90 percent of the original New York City sample remained in our analytic sample. In the rest of the state, 95 percent of the original sample remained in our analytic sample. For both of the two subsamples, as

[1] For our final analytic sample, we dropped 133 observations linked to building codes that had 10 or fewer students because they contributed extreme values for school mean variables and often represented unique learning conditions for students. We also dropped 45 observations belonging to P.S. 184 Shuang Wen School, a public school in New York City with an immersion program in Mandarin Chinese. 93% of the school's seventh graders were Asian, an extreme value relative to the rest of the schools in our sample that inflated coefficients for the school proportion-Asian variable.

in the state as a whole, the analytic samples differ only very slightly from the original sample (Table 1).

Methods

Outcome

The outcome is measured by the difference in performance on the non-audit and audit portions of the NYS Grade 7 Mathematics Test. Our audit measure is the raw score, calculated as a proportion of possible credit, on the items assessing standards ("performance indicators" in the terminology used in New York at the time) not previously tested (see Appendix A). Our non-audit measure is the raw score on the remaining operational test items. In grade 7, which is the primary focus of the paper, we excluded only one item that the state dropped because of a negative correlation with test scores. In each of the other grades, we dropped three items that assessed previously untested standards but that were extremely easy (p-values ranging from .79 to .96).

A difficulty in designing SMAs is that audit items may share characteristics with non-audit items, as a result of which the effects of test preparation focused on non-audit items may generalize to the audit items. This would produce an underestimate of score inflation. Where standards are used to identify audit items, as in this study, this risk may arise if the content covered by previously tested and previously untested standards is similar or if the items share non-substantive features that have been the focus of inappropriate test preparation. Although the bias caused by similarities between audit and non-audit items would be conservative, we explored this possibility by comparing the content of the performance indicators assessed by audit and non-audit items. For earlier

work, other members of our team had created groupings of performance indicators with highly similar content. Using this classification, we found that no seventh-grade audit items were in a grouping with any non-audit items. One fourth-grade item was in a grouping with several non-audit items, but the concept it assessed appeared sufficiently distinct, so we retained the item. A single eighth-grade audit items was similarly questionable but was ultimately retained. While this analysis helped to confirm that the audit and non-audit tests did not have overlapping content, it did not address the risk that audit and non-audit items shared other attributes, e.g., aspects of presentation or content not captured by labeling of the performance indicators.

The dependent variable in our analyses is based on the simple difference between the raw proportion correct on both the non-audit portion of the test and the untested standards audit component:

(1) $$Y_{is} = p_{is}^{non} - p_{is}^{audit}$$

where *i* indexes individuals and *s* indexes schools. We standardized this difference to mean 0, standard deviation 1. The sensitivity of our findings to this approach is discussed below.

The audit portions of the grade 4, grade 7 and grade 8 assessments varied by length and the characteristics of items. The audit portion in grade 7, which is the primary focus of this paper, comprised eight items, including one constructed-response item. This was the most diverse group of audit items, as it assessed performance indicators from four of the five seventh-grade mathematics content strands: algebra, number sense and operations, geometry, and statistics and probability. The grade 4 audit test was very short

and narrow, including only three items assessing performance indicators from a single strand (number sense and operations). The grade 8 audit portion was the longest, with ten multiple choice items, but all but one item assessed performance indicators from algebra and geometry, the strands that also constituted a very large portion of the non-audit test. However, as we explain later, the different representation of items assessing strands does not appear to explain discrepancies in the audit functioning across grades.

Predictors

Demographics. We included student-level dummy variables for black, Hispanic, and Asian students, leaving White and other-race students as the omitted comparison group.[2] We also included a student-level dummy variable for low-income status, indicating a student's participation in free- or reduced-price lunch programs or other economic assistance programs, such as food stamps or Supplemental Security Income.[3] The means of these variables were then computed for schools.

Bubble status. Dummy variables were created to flag bubble students. We defined these as students whose scores in the previous year (2010) were up to three raw score points below the cut score between Level 2 and Level 3 ("Proficient"). We used this as a proxy for educators' anticipated scores for the students in 2011. The proportions of bubble students were then computed for schools. A sensitivity analysis comparing alternative definitions of this variable is discussed below.

---

[2] In most cases, race was reported by parents. When parents do not report race, districts are responsible for assigning classifications.

[3] For detailed information about the criteria for the low-income variable, see University of the State of New York (2011), p. 44.

Region. We created dummy variables to separate districts into three categories, based on differences in district size, urbanicity, and demographics: New York City; other urban (Buffalo, Rochester, Syracuse, and Yonkers, often called "the big four," the four largest districts after NYC); and all other districts (the omitted category).

Analytic Strategy

We used a difference-in-differences approach to detect potential score inflation. The audit test score, p_{is}^{audit}, is an unbiased estimator of the student's uninflated achievement, θ:

$$p_{is}^{audit} = \theta_{is} + \epsilon_{is}, \qquad E(\epsilon_{is}) = 0. \tag{2}$$

However, p_{is}^{non} is potentially biased by inflation, ζ_{is}, which is expected to vary both within and between schools. In addition, p_{is}^{non} may differ from p_{is}^{audit} by a difficulty factor, δ, that is unrelated to inflation:

$$p_{is}^{non} = \theta_{is} + \delta + \zeta_{is} + \nu_{is}. \tag{3}$$

Therefore one can re-express the audit measure (1) as:

$$Y_{is} = \delta + \zeta_{is} + (\nu_{is} - \epsilon_{is}). \tag{4}$$

Note that the non-inflationary difference in difficulty, δ, is shown as a constant. Student-level variations in difficulty unrelated to inflation contribute to measurement error. School-level variations in difficulty pose a more complex issue. In the general case, school performance may differ between parts of a test for systematic reasons unrelated to inflation. For example, when another test that reflects a somewhat different framework is used as an audit, there is a risk that schools will vary in the emphasis they give to certain aspects of the audit test because of curricular differences that are independent of their

responses to the high-stakes test (Ng, Koretz, & Jennings, 2013). In this case, however, all of the content of both subtests is explicitly included in the target of inference, and therefore in the content that schools are expected to teach. Therefore, if some schools show strong performance on the tested standards that does not generalize to the untested standards included in the inference, we can consider that score inflation. Accordingly, we can treat school-level differences in performance between the two subtests as comprising only inflation and school-level error and not non-inflationary differences in difficulty, although we cannot know whether the inflation is a result of deliberate responses to testing. Our analytical models, unlike the simpler presentation here, incorporate estimates of school-level error.

This outcome (equation 4) poses two difficulties for our analysis. First, the student-level error in the difference score, $(\nu_{is} - \epsilon_{is})$, will be large because of the short length of the audit test. Below we show that the estimated reliability of $\left(p_{is}^{non} - p_{is}^{audit}\right)$ based on the internal consistency reliabilities of the two subtests is extremely low. This creates a severe conservative bias—that is, a large risk of a Type II error. We rely primarily on aggregate (school-level) findings, which ameliorates this problem to some degree, but it remains an important limitation of the data that is likely to produce an underestimate of effects.

An even more important limitation is that δ cannot be estimated from our data. To estimate δ, we would need uninflated estimates of the difficulties of both subtests, but we do not have such an estimate for the non-audit subtest, which was administered only

under conditions vulnerable to inflation. Therefore, δ and ζ are confounded, and the simple difference $\left(p_{is}^{non} - p_{is}^{audit}\right)$ cannot be interpreted as an indicator of inflation.

We address this by using a difference-in-differences approach, which removes the effects of δ. Specifically, we investigate whether variations in $\left(p_{is}^{non} - p_{is}^{audit}\right)$ are systematically associated with student- and school-level variables that have been shown in prior studies to be associated with either score inflation or inflation-inducing instructional behaviors: ethnicity, economic disadvantage, and bubble status. Because we expect that inflation-inducing behaviors vary across schools, we examine relationships with both student-level variables and school-level aggregates of these variables.

We began with a two-level random effects model (students at level 1, schools at level 2):

(5)
$$Y_{is} = \beta_{0s} + \mathbf{X}\boldsymbol{\beta_{10}} + \epsilon_{is}$$
$$\beta_{0s} = \gamma_{00} + \mathbf{Z}\boldsymbol{\gamma_{01}} + u_{0s}$$

where **X** is a vector of student-level variables and **Z** is the corresponding vector of school-level means. This model appropriately adjusts standard errors for clustering and also accommodates our hypothesis that inflation-inducing behaviors are to some degree a school-level variable. We did not include cross-level interactions between levels 1 and 2—that is, within-school slopes were fixed across schools.

Although it is likely that inflation-related variables also vary systematically across districts, we could not fit three-level models nesting schools within districts because 95 percent of districts in New York State have five or fewer schools in them. However, in our final regression models, we included a level-3 fixed effect for NYC schools, as well

as cross-level interactions with the NYC dummy. This strategy was motivated by demographic and contextual differences between New York City schools and those in the rest of the state. First, New York City had implemented a unique, high-stakes accountability system (New York City Department of Education, 2007). Second, schools in New York City are markedly different in demographic composition. The confounding of minority composition and the NYC dummy can be seen in Figure 2, which plots the proportion Black or Hispanic against the proportion low income, separately for the two regions. Most NYC schools have a large proportion of low-income students (many reporting 100 percent low-income enrollments), and many also have a large proportion of black or Hispanic students (top right corner of the left-hand panel). In contrast, the rest of the state has far fewer such schools, with most schools in the lower left hand corner (low percent black or Hispanic, less than half low-income). Even more striking, almost all schools statewide with fewer than half low-income or black or Hispanic students are outside of NYC.

In response, we fitted models that included a NYC fixed effect (with the rest of the state as the omitted category) and the interactions between this dummy and level-1 and level-2 predictors. Initially, we also estimated models that included a fixed effect for "other urban" (Buffalo, Rochester, Syracuse, and Yonkers) because those districts are more similar in demographics and urbanicity to NYC than to the rest of the state. However, our results show that these other-urban districts had patterns of performance similar to those for other non-NYC schools, so the final models presented here include only the NYC fixed effect.

Thus, our primary final models, where *N* denotes the NYC dummy, were:

(6) $$Y_{is} = \beta_{0s} + \mathbf{X\beta_{10}} + \beta_{20}N + N\mathbf{X\beta_{30}} + \epsilon_{is}$$

$$\beta_{0s} = \gamma_{00} + \mathbf{Z\gamma_{01}} + N\boldsymbol{Z\gamma_{02}} + u_{0s}$$

We grand-mean centered the student-level predictors. This makes the β_{0s} values interpretable as adjusted group means. In addition, this yields parameter estimates for level-2 variables that are direct estimates of context effects (Raudenbush & Bryk, 2002). That is, the parameter estimates for level-2 variables indicate the extent to which the school means differ by more than the level-1 model would predict. We also group-mean centered level-2 variables for ease of interpretation. We did not center the NYC dummy variable, allowing us to interpret the non-interacted terms as the parameter estimates for schools outside NYC and to interpret the interaction terms as additional effects for students and schools in NYC.

Interpreting the practical magnitude of the findings is not straightforward for the aggregate level-2 variables. The practical strength of the relationships between each aggregate predictor and the outcome depends on its distribution in the population, not just the magnitude of the coefficients. For example, the proportion of bubble students is small in most schools. If the coefficient for the proportion of bubble students was large but the observed range of the proportions for NYS schools was small, the practical importance of the school-level bubble effect would be modest despite the large coefficient. Therefore, although the aim of this study is to investigate the feasibility of using untested-standards items for auditing, not to estimate the magnitude of inflation, we quantified the estimate

level-2 effects by calculating the difference scores for observations at the 25th and 75th percentiles on each predictor (Table 2).

Results

We first present our primary results from grade 7 and then compare findings in grades 4 and 8.

Descriptive statistics and item analysis

In grade 7, the audit component was more difficult than the non-audit component (mean proportions correct of .50 and .66, respectively), and scores on the audit were somewhat more variable (Table 3). In addition, non-audit scores showed a substantial left skew (skewness = -0.40). This was unsurprising, as the raw score distributions on high-stakes tests often quickly develop ceiling effects (Ho & Yu, 2013). Because of other changes made to the test in an effort to address possible score inflation, the distribution of the 2011 raw scores, even after deleting the items assessing previously untested standards, was less severely censored than the 2010 scores (skewness = -0.60).

We conducted classical item analyses for both the audit and non-audit tests (Table 4). On average, the grade 7 audit items were substantially more difficult (.49) than the non-audit items (.66). The audit items showed a narrower range of difficulties, and the easiest audit item was more difficult (.73) than the easiest non-audit item (.98). The 2011 non-audit component was quite similar in this regard to the 2010 test, and patterns were largely similar across the three grades. Internal consistency reliability (as measured by Cronbach's alpha) was .90 for the non-audit test and .66 for the audit test. The relatively low reliability on the audit test is a result of its shortness, not greater item

heterogeneity: the Spearman-Brown prophecy formula predicts that the internal consistency reliability of the audit test would be .91 if it were the same length as the non-audit test. Similarly, item-rest point-biserial correlations between items and total scores were within normal ranges for both tests. The correlation between scores on the audit and non-audit components was moderately high (r = .76) despite attenuation from measurement error. Because of this correlation and the modest internal consistency reliability of the audit measure, the student-level difference measure was highly unreliable, $r_{xx'} = .08$. As noted earlier, this creates a potentially severe conservative bias in our findings, although our focus on school-level relationships lessens this problem somewhat.

For the most part, collinearity was not severe in these data. At the student level, only a single correlation between predictors exceeded .35: low-income students were more common in NYC ($r = .46$; Table 5). As expected, correlations were higher at the aggregate (school) level, but most of these were also modest (Table 6). Apart from the correlations with the NYC dummy discussed above, only two school-level correlations exceeded .50: proportion low income with proportion Black ($r = .56$) and proportion Hispanic ($r = .62$).

Results in grade 7

We hypothesized that score inflation would vary across schools. Most of the variation in the outcome was within schools, and the intraclass correlation from an unconditional model was low ($\rho = .06$). This is partly due to the large amount of student-level measurement error in the outcome, which increases dispersion of students' scores

and is random with respect to schools. However, this is a sufficient amount of clustering that each of our multilevel models provided significantly better fit than the corresponding single-level models ($p < .001$ in each comparison). Examination of residuals from the multilevel models indicated no substantial heteroskedasticity.

The student-level effects in grade 7 were mostly small, as we anticipated in the light of the very low reliability of the student-level difference score, although some were nonetheless statistically significant in our very large sample. Outside of NYC, the largest effect was for bubble students, whose difference scores were, on average, a fourth of a standard deviation larger than that of other students (Table 7, column 2). Asian students outside NYC had difference scores roughly 0.1 SD smaller than non-Asian students. Hispanic and low-income students outside NYC had difference scores roughly 0.1 SD larger than white and other-race students, holding other variables constant. The corresponding estimate for black students, however, was both trivial and nonsignificant.

Outside of NYC, effects at the school level were considerably larger and all highly significant. Outside of NYC, the difference score was positively associated with the proportion of bubble students in the school (0.74) and the proportion of low-income students (0.20), but was negatively associated with the proportion of Asian students (-0.58), holding all other variables constant. However, the difference score was negatively associated with the proportions of both Hispanic (-0.20) and black (-0.31) students after controlling for the other variables in the model.

At the student level, the main effect of the NYC variable was substantial and negative (-0.25). However, this is not the estimated difference in overall means for

students inside and outside of NYC. Rather, it is the adjusted mean difference for students inside and outside of NYC with zero values for all other student- and school-level variables. Given the demographic differences between NYC and the rest of the state, this comparison is not of substantive interest. However, the interactions of the NYC dummy with the level-1 and level-2 predictors provide important additional information about how the relationships between the difference score and student- or school-level predictors differ within and outside New York City. The student-level interactions with the NYC dummy were generally very small, although three were statistically significant. In NYC, inflation was a bit larger for black students compared to the rest of the state and trivially smaller for both low-income and bubble students. In contrast, two interactions between school-level variables and the NYC dummy were statistically significant and large: the interactions with proportion Asian and proportion black. In both cases, the relationship between these proportions and the difference score were much weaker for schools in NYC than for comparable schools outside of the city. The estimated relationship between proportion Asian and the difference score was approximately -0.16 in NYC, compared to -0.58 elsewhere in the state. Similarly, the negative relationship between the difference score and proportion black was very weak in NYC (-0.10, compared to -0.31 outside of NYC).

As noted earlier, the practical impact of these school-level relationships depends on not only the strength of the relationship between the predictor and the outcome, but also on the distributions of the predictors. We quantified this by estimating the difference in the outcome for schools at the 25^{th} and 75^{th} percentiles on each of the predictors

(Table 2). Two of the interquartile differences were sizable and consistent with our original hypotheses. The difference score was substantially larger in schools at the 75th percentile on proportion low income than at the 25th percentile, by 0.22 SD in NYC and 0.32 SD elsewhere (Table 2). The practical impact of the proportion of bubble students was somewhat smaller despite the large level-2 regression coefficient because of the limited spread of this predictor: the interquartile difference was slightly under 0.2 SD in both regions. The interquartile difference for proportion Asian was very small. For both proportion black and proportion Hispanic, the interquartile differences were inconsistent with expectations, ranging from -0.09 SD to -0.21 SD.

Results from grades 4 and 8

Results for the fourth-grade audit test were largely consistent with the seventh-grade results described above, although in most instances, the parameter estimates outside NYC were smaller in grade 4 (Table 7, column 1). In some instances, the differences were modest, but in others they were very large. In particular, the parameter estimate for the proportion of bubble students outside of NYC was less than half as large in grade 4 as in grade 7. In contrast, the eighth-grade results show little systematic association between the outcome and the predictors in our model. Many of the estimates are very small, and some of those that are substantively large failed to reach significance despite our large sample (Table 7, column 3).

There are no obvious patterns in the data that clearly explain the differences in results between grades 4 and 7. Because the fourth-grade audit measure was extremely short (only 3 items), one might speculate that the observed differences between grades

result from lower reliability in grade 4 producing greater attenuation, but this is not the case. While the audit scores are indeed somewhat less reliable in grade 4 than in grade 7 ($r_{xx'} = .47$, compared with $r_{xx'} = .66$), the student-level reliability of the difference score is actually a bit higher in grade 4 ($r_{xx'} = .11$, compared with $r_{xx'} = .08$) because the non-audit component is more reliable in grade 4 than in grade 7 and the correlation between the audit and non-audit tests is lower ($r = .69$). The distributions of scores also differed between grades 4 and 7, but the impact of these differences is not clear. Recall that in grade 7, the non-audit component showed a substantial ceiling effect (skewness = -0.40; Table 3), while the audit component showed no ceiling and indeed showed modest positive skew (0.14). In contrast, in grade 4, both the audit and non-audit distributions were right-skewed, although the audit component was much less so (Table 3). One might expect the ceiling in the audit measure to undermine its effectiveness. However, the differences between the components in both difficulty and standard deviation were considerably *larger* in grade 4, which one might expect to increase their potential for auditing.

Other considerations suggest that the difference between grades 4 and 7 may be substantive, although our data are not sufficient to confirm this. First, the comparison with NAEP suggests relatively less severe inflation in grade 4, which might lead one to predict that smaller effects would be found via audit testing. Second, the fourth-grade audit measure was very narrow as well as short, comprising only three items assessing a single strand (numbers and operations). This makes it more likely that the US item set failed to include content that contributed to the audit function. Finally, the differences

between the grades in the relations with the NYC dummy also suggest the possibility of substantive explanations.

In contrast, it seems that the weak observed relationships in grade 8 arose from technical factors. The comparison with NAEP suggests very large inflation in grade 8, so that possible explanation can be ruled out. The first technical factor is low reliability. While the audit component in eighth grade is the most reliable of the three ($r_{xx'} = .72$), the correlation between the audit and non-audit components is high ($r = .81$), so that the student-level reliability of the difference is near zero ($r_{xx'} = .02$).Second, there was a near-zero difference in mean scores (0.02) between the audit and non-audit test. In both grade 4 and grade 7, the change in the test in 2011 resulted in the addition of more difficult items, but the distribution of raw scores remained otherwise quite similar. In contrast, in grade 8, the entire distribution of p-values was shifted downward, and the negative skewness was largely eliminated (Table 3). This suggests that the operational test had been made less predictable and therefore that the audit test may not have been sufficiently different from the non-audit test to yield noticeable differences in performance. Third, the variability of audit and difference scores was smaller in grade 8 than in the other grades, making it more difficult to capture any potential variation in inflation. The grade 8 audit was also narrow, comprising items from only two strands, but the grade 4 results indicate that this alone would not preclude systematic differences in performance.

Sensitivity analyses

We analyzed the sensitivity of our findings to two decisions: the definition of the bubble variable and the decision not to standardize the audit and non-audit components separately.

We compared three definitions of bubble status. The definition used in the analysis above, a band of three raw score points below the Proficient standard the previous year, categorized 9.2 percent of observations as bubble students. A broader alternative used a band of two points above and below the Proficient cut and classified 13.7 percent of students as in the bubble. A third alternative used a band of two points below Proficient and categorized 6.3 percent of cases as bubble students. The choice among these three definitions had no appreciable effects on the results. The first two yielded essentially the same results, while the third differed only in that the small coefficient for the student-level interaction between NYC and bubble status became nonsignficant (Appendix B).

In contrast, the choice of standardization method was consequential. As noted above, our outcome variable was the simple difference in proportion of possible points between the non-audit and audit portions of the test. We then standardized these difference scores for ease of interpretation. A plausible alternative would be to standardize the two components separately and then difference the standardized scores.

This decision has an impact because of differences between the distributions of the audit and non-audit scores. In grade 7, for example, the audit scores are somewhat right-skewed, while the non-audit scores are left-skewed and show a substantial ceiling

effect (Table 3; see also Figure 3). The ceiling may be the reason that the standard deviation of non-audit scores is smaller (0.20) than that of the audit scores (0.27).

The choice whether to standardize at the component level hinges on one's interpretation of these differences in distributions. If one believes that they are simply an artifact of scaling, then it would be reasonable to standardize at the component level and thus remove the artifact. On the other hand, if one believes that the differences in distributions reflect the phenomenon under investigation, then standardizing would be inappropriate in that it would remove the subject of study. Our interpretation is that the variance of the non-audit component is reduced by the phenomenon under investigation, score inflation, as the rapid gains in scores on high-stakes tests are often accompanied by severe right-side censoring (Ho & Yu, 2013). Thus, we interpret the greater standard deviation in the non-audit scores as substantively meaningful. If that is correct, standardizing at the component level would not be appropriate.

This decision had a large impact on the results. In grade 7, if one standardizes at the component level, most of the relevant coefficients shrink dramatically, and some change sign. For example, the most important coefficient, the effect of school proportion bubble students, shrinks to near zero (Appendix C).

Discussion

This study, which is the first evaluation of a self-monitoring assessment (SMA) in an operational administration of a high-stakes test, explored the feasibility of using items assessing previously untested standards (US items) as one component of a self-monitoring assessment (SMA). It was not designed to estimate the severity of score

inflation. The study is premised on the assumption that US items cannot be sufficient to measure inflation because they cannot capture the effects of some types of in appropriate test preparation. Moreover, for reasons noted earlier, our data create a severe downward bias in our estimated effects.

Despite the limitations of our data, our findings in grades 4 and 7 suggest that US items can contribute to an audit. Despite the very low reliability of the difference scores, the performance differences between US items and the remainder of the test in these grades showed clear systematic patterns at both the student and school levels. As noted earlier, the US items represent material given weight in the target of inference, so a failure to generalize to these items can be considered inflation. However, it is more difficult to determine the extent to which this performance difference reflects the effect of inflation-inducing responses to testing, rather than differences in instruction that are not a response to testing but that happen to create inflation. Although the difference-in-differences approach required by the nature of our data provide only an indirect test of inflation, we consider the results to suggest that the performance difference between the audit and non-audit components represents inflationary responses to testing. Many of the observed relationships were consistent with expectations. The most important evidence that these results do reflect responses to testing is that the bubble-student variables, the predictors most likely to create strong incentives to engage in inflation-inducing behavior, showed the strongest relationships to the difference score, both at the student and school levels and both in NYC and elsewhere in the state.

On the other hand, the results in grade 8 clearly indicate that measuring previously untested standards is not sufficient to indicate that items will be useful in auditing. The disparity in score gains between NAEP and the state test was very large in grade 8, and the presence of score inflation in that grade is widely acknowledged, but we found virtually no systematic patterns in the difference scores. We noted several factors that we believe contributed to this finding with these data, but US items could fail as audits for other reasons as well, for example, if they share with other items attributes that are the focus of inappropriate test preparation, such as teaching to the rubric (Stecher & Mitchell, 1995).

Because of two important limitations of our data, the findings presented here understate the potential utility of US items for auditing gains. The first limitation is the attenuation of all systematic relationships by error. In no instance did the student-level reliabilities of our difference scores exceed .11. A reliability of .11 attenuates correlations by roughly two-thirds and therefore creates a severe downward bias in the estimates from regression models such as ours. The second limitation is that our US items were a happenstance selection, not intended for auditing. The between-grade differences in results suggest that the selection of US items is important, and a set well-designed for auditing might produce substantially stronger effects than we found.

The conventional call for additional research is particularly appropriate here. First, given the lack of similar studies, there is a clear need for replications of this study in other contexts. Second, the field has as yet produced very little research bearing on the optimal characteristics of items for auditing. To date, almost all studies of score inflation

have focused on disparities in trends in total scores between high-stakes tests and external audit tests that happened to be available, such as NAEP or lower-stakes tests already administered by a state or locality. Additional work is needed to identify the characteristics of items that contribute to these disparities in trends. Moreover, there is as yet almost no research investigating the design of items intended specifically for auditing. For example, to our knowledge, there have been no efforts to date to design audit items to reflect the specific types of inappropriate test preparation that are most commonly used in response to a particular accountability test. Given both the likely continuation of the current emphasis on test-based accountability and the clear limitations of auditing with second tests that happen to be available, these are important areas for additional research.

References

Abrams, L. M., Pedulla, J. J., & Madaus, G. F. (2003). Views from the classroom: Teachers' opinions of statewide testing programs. *Theory into Practice, 42*(1), 18-29.

Booher-Jennings, J. (2005). Below the bubble: "Educational Triage" and the Texas Accountability System. *American Educational Research Journal, 42(2), 231-268.*

Cimbricz, S. (2002). State-mandated testing and teachers' beliefs and practice. *Education Policy Analysis Archives, 10*(2). Retrieved from http://epaa.asu.edu/epaa/v10n2.html

Diamond, J. B., & Spillane, J. P. (2004). High-stakes accountability in urban elementary schools: Challenging or reproducing inequality? *Teachers College Record, 106*(6), 1145-1176.

Eisner, E. (2001). What does it mean to say a school is doing well? *Phi Delta Kappan, 82*(5), 367-372.

Figlio, D. and Getzler, L. (2006). Accountability, ability and disability: Gaming the system? In T. Gronberg & D. Jansen (Eds.), *Advances in Applied Microeconomics* (Vol. 14, pp. 35–49). Elsevier.

Firestone, W. A., Camilli, G., Yurecko, M., Monfils, L., & Mayrowetz, D. (2000). State standards, socio-fiscal context and opportunity to learn in New Jersey. *Education Policy Analysis Archives, 8*(35). Retrieved from http://olam.ed.asu.edu/epaa/v8n35.

Gillborn, D. & Youdell, D. (2000). *Rationing education: policy, practice, reform, and equity*. Buckingham, UK: Open University Press.

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, L. S., Stecher, B. M., Marsh, J. A., McCombs, J. S., Robyn, A., Russell, J., Naftel, S., and Barney, H. (2007). *Standards-based accountability under No Child Left Behind: Experiences of teachers and administrators in three states*. Santa Monica, CA: RAND Corporation.

Haney, W. (2000). The myth of the Texas miracle in education. *Education Policy Analysis Archives, 8*(41). Retrieved from http://epaa.asu.edu/ojs/article/view/432/828

Herman, J. L., & Golan, S. (1993). The effects of standardized testing on teaching and schools. *Educational Measurement: Issues and Practice, 12*(4), 20-25, 41-42.

Ho, A. D. (2009). A nonparametric framework for comparing trends and gaps across tests. *Journal of Educational and Behavioral Statistics, 34*(2), 201-228.

Ho, A. D., & Haertel, E. H. (2006). *Metric-free measures of test score trends and gaps with policy-relevant examples* (CSE Report No. 665). Los Angeles, CA: Center for the Study of Evaluation. Retrieved from http://www.cse.ucla.edu/products/reports/r665.pdf

Ho, A. and Yu, C. (2013). *Descriptive statistics for modern test score distributions: Skewness, kurtosis, discreteness, and ceiling effects.* Unpublished working paper.

Holcombe, R., Jennings, J. L., & 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

Jacob, B. A. (2005). Accountability, incentives and behavior: The impact of high-stakes testing in the Chicago public schools. *Journal of Public Economics*, *89*(5-6), 761-796. doi: 10.1016/j.jpubeco.2004.08.004

Jacob, B. (2007). *Test-based accountability and student achievement: An investigation of differential performance on NAEP and state assessments.* Cambridge, MA: National Bureau of Economic Research (Working Paper 12817).

Jacob, R. T., Stone, S., & Roderick, M. (2004). *Ending social promotion: The response of teachers and students*. Chicago, IL: Consortium on Chicago School Research. Retrieved March 29, 2011, from http://www.eric.ed.gov/PDFS/ED483823.pdf

Jones, M. G., Jones, B. D., & Hargrove, T. Y. (2003). *The unintended consequences of high-stakes testing*. Lanham, MD: Rowman & Littlefield Publishers, Inc.

Klein, S. P., Hamilton, L.S., McCaffrey, D.F., and Stecher, B.M. (2000). *What do test scores in Texas tell us?* Santa Monica, CA: RAND (Issue Paper IP-202). Last accessed from http://www.rand.org/publications/IP/IP202/ on June 4, 2013.

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

Koretz, D., & Béguin, A. (2010). Self-Monitoring Assessments for Educational Accountability Systems. *Measurement: Interdisciplinary Research & Perspective*, *8*, 92–109. doi:10.1080/15366367.2010.508685

Koretz, D., and 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.

Ladd, H. F., & Zelli, A. (2002). School-based accountability in North Carolina: The responses of school principals. *Educational Administration Quarterly, 38*(4), 494-529. doi: 10.1177/001316102237670

Lindquist, E. F. (1951). Preliminary considerations in objective test construction. In E. F. Lindquist (Ed.), *Educational Measurement* (pp. 119-184). Washington, DC: American Council on Education.

Lipman, P. (2002). Making the global city, making inequality: The political economy and cultural politics of Chicago school policy. *American Educational Research Journal, 39*(2), 379-419.

Luna, C., & Turner, C. L. (2001). The impact of the MCAS: Teachers talk about high-stakes testing. *The English Journal, 91*(1), 79-87.

Madaus, G. F. (1988). The distortion of teaching and testing: High-stakes testing and instruction. *Peabody Journal of Education*, 65(3), 29-46.

McNeil, L. M. (2000). *Contradictions of school reform: Educational costs of standardized testing*. New York, NY: Routledge.

McNeil, L. M. & Valenzuela, A. (2001). The harmful impact of the TAAS system of testing in Texas: Beneath the accountability rhetoric. In M. Kornhaber & G. Orfield (Eds.), *Raising standards or raising barriers? Inequality and high-stakes testing in public education* (pp. 127-150). New York, NY: Century Foundation.

Neal, D. and Schanzenbach, D. (2010). Left behind by design: Proficiency counts and test-based accountability. *Review of Economics & Statistics, 92(2), 263–283.*

New York City Department of Education (2007). Mayor Bloomberg and Chancellor Klein release first-ever public school Progress Reports [Press release]. Retrieved from http://schools.nyc.gov/Offices/mediarelations/NewsandSpeeches/2007-2008/20071105_progress_reports.htm.

Ng, H. L., Koretz, D., & Jennings, J. L. (2013). *Sensitivity of school-performance ratings to score inflation: An exploratory study using a self-monitoring assessment*. A working paper of the Education Accountability Project at the Harvard Graduate School of Education. Last retrieved on June 3, 2013 from http://projects.iq.harvard.edu/files/eap/files/audit_paper_g4only_wp_draft_031013_1.pdf.

Pedulla, J. J., Abrams, L. M., Madaus, G. F., Russell, M. K., Ramos, M. A., & Miao, J. (2003). *Perceived effects of state-mandated testing programs on teaching and*

learning: Findings from a national survey of teachers. Boston, MA: National Board on Educational Testing and Public Policy. Retrieved from http://www.bc.edu/research/nbetpp/statements/nbr2.pdf

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods, second edition*. Thousand Oaks, CA: Sage.

Rubinstein, J. (2000). *Cracking the MCAS grade 10 math*. New York: Princeton Review Publishing.

Severson, K. (2011). A scandal of cheating, and a fall from grace. *The New York Times*, September 7, p. A16. Last retrieved on June 5, 2013 from http://www.nytimes.com/2011/09/08/us/08hall.html?pagewanted=all&_r=0.

Shen, X. (2008). Do unintended effects of high-stakes testing hit disadvantaged schools harder? (Doctoral dissertation, Stanford University).

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

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. Retrieved from http://www.rand.org/publications/MR/MR1554/MR1554.ch4.pdf..

Stecher, B. M., & Barron, S. I. (2001). Unintended consequences of test-based accountability when testing in “milepost” grades. *Educational Assessment, 7*(4), 259-281.

Stecher, B. M., Epstein, S., Hamilton, L. S., Marsh, J. A., Robyn, A., McCombs, J. S., Russell, J., & Naftel, S. (2008). *Pain and gain: Implementing NCLB in three states, 2004 – 2006*. Santa Monica, CA: RAND. Retrieved from http://www.rand.org/pubs/monographs/2008/RAND_MG784.pdf

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

Taylor, G., Shepard, L., Kinner, F., & Rosenthal, J. (2002). *A survey of teachers' perspectives on high-stakes testing in Colorado: What gets taught, what gets lost* (CSE Technical Report 588). Los Angeles, CA: University of California. Retrieved September 20, 2010, from http://eric.ed.gov/PDFS/ED475139.pdf

Tisch, M. (2009). What the NAEP results mean for New York. Latham, N.Y.: New York School Boards Association (November 9). Last retrieved on June 24, 2012 from http://www.nyssba.org/index.php?src=news&refno=1110&category=On%20Board%20Online%20November%209%202009.

University of the State of New York (2011). *New York State Student Information Repository System (SIRS) manual.* Albany, N. Y.: Author. Last accessed June 24, 2013 from age 244 of http://www.p12.nysed.gov/irs/sirs/archive/2010-11SIRSManual6-2.pdf.

Urdan, T. C., & Paris, S. G. (1994). Teachers' perceptions of standardized achievement tests. *Educational Policy, 8*(2), 137-157.

Tables

Table 1

Demographic Characteristics of the Original and Analytic Samples, Grade 7

	Means					
	New York State		New York City		Non-New York City	
Category	Original Data	Analytic Sample	Original Data	Analytic Sample	Original Data	Analytic Sample
White	.51	.52	.14	.14	.69	.70
Asian	.08	.08	.14	.15	.05	.05
Black	.19	.18	.30	.30	.13	.12
Hispanic	.22	.21	.41	.41	.12	.12
Other	.01	.01	.01	.01	.01	.01
Low Income	.52	.51	.86	.85	.36	.35
Near Level 3 Cut Score (2010)	.09	.09	.10	.10	.09	.09
Total Observations (N)	199,276	185,522	65,194	58,691	134,082	126,831

Table 2

Difference in Outcome for Schools Between the 75th and 25th Percentiles on School-Level Predictors for Schools Within and Outside New York City, Grade 7

	NYC	Outside NYC
Proportion Asian	-0.06	-0.09
Proportion Black	-0.21	-0.18
Proportion Hispanic	-0.19	-0.09
Proportion Low Income	0.22	0.31
Proportion Bubble	0.18	0.16

Table 3

Descriptive Statistics for 2011 Non-Audit Component, Audit Component, and Difference Score

Grade	Statistic	Non-audit	Audit	Difference
4	Mean	0.71	0.55	0.16
	Standard deviation	0.19	0.33	0.25
	Skewness	-0.73	-0.18	0.17
7	Mean	0.66	0.50	0.16
	Standard deviation	0.20	0.27	0.18
	Skewness	-0.40	0.14	0.14
8	Mean	0.58	0.56	0.02
	Standard deviation	0.22	0.25	0.15
	Skewness	-0.15	-0.07	0.15

Table 4

Descriptive Statistics of the Proportions of Possible Points Earned on Items from the New York State Mathematics Tests

				Proportion of Points Earned			Internal Consistency
Grade	Year	Assessment	Items	Mean	Minimum	Maximum	Alpha
4	2010	Entire	48	.77	.44	.96	.93
	2011	Non-audit	54	.72	.23	.97	.93
	2011	Audit	3	.55	.33	.68	.47
7	2010	Entire	38	.70	.24	.92	.90
	2011	Non-audit	44	.67	.36	.98	.90
	2011	Audit	8	.49	.32	.73	.66
8	2010	Entire	45	.73	.52	.96	.94
	2011	Non-audit	44	.61	.33	.83	.92
	2011	Audit	10	.55	.40	.71	.72

Table 5

Pearson Correlations among Student Characteristics and Region, Grade 7

	Non-audit audit difference	NYC	Asian	Black	Hispanic	Other	Low Income	Bubble Status
Non-audit audit difference	1.00							
NYC	-.06	1.00						
Asian	-.07	.17	1.00					
Black	.00	.21	-.14	1.00				
Hispanic	.03	.33	-.15	-.24	1.00			
Other	.00	-.02	-.03	-.05	-.06	1.00		
Low Income	.03	.46	.09	.26	.33	.01	1.00	
Bubble Status	.08	.01	-.04	.03	.03	.00	.05	1.00

Table 6

Pearson Correlations among School Characteristics and Region, Grade 7

	Non-audit audit difference	NYC	Proportion Asian	Proportion Black	Proportion Hispanic	Proportion Other	Proportion Low Income	Proportion Bubble
Non-audit audit difference	1.00							
NYC	-.24	1.00						
Proportion Asian	-.28	.24	1.00					
Proportion Black	-.09	.38	-.12	1.00				
Proportion Hispanic	-.09	.60	.05	.09	1.00			
Proportion Other	.04	-.08	-.03	-.08	-.10	1.00		
Proportion Low Income	-.01	.66	.10	.56	.62	-.04	1.00	
Proportion Bubble Status	.22	.09	-.23	.19	.14	.00	.27	1.00

Table 7

Final Multilevel regression models

	Grade 4	Grade 7	Grade 8
Constant	0.034***	0.058***	0.012
	(0.01)	(0.01)	(0.01)
Asian	-0.082***	-0.118***	-0.031*
	(0.02)	(0.01)	(0.01)
Black	0.043***	0.02	-0.004
	(0.01)	(0.01)	(0.01)
Hispanic	0.096***	0.119***	-0.008
	(0.01)	(0.01)	(0.01)
Low Income	0.065***	0.082***	0.01
	(0.01)	(0.01)	(0.01)
Bubble	0.166***	0.259***	0.099***
	(0.01)	(0.01)	(0.01)
Proportion Asian	-0.290***	-0.578***	-0.280*
	(0.06)	(0.12)	(0.13)
Proportion Black	-0.113***	-0.305***	-0.282***
	(0.03)	(0.05)	(0.05)
Proportion Hispanic	-0.01	-0.199***	-0.216**
	(0.03)	(0.06)	(0.07)
Proportion Low Income	0.112***	0.203***	0.209***
	(0.03)	(0.04)	(0.05)
Proportion Bubble	0.217*	0.735***	0.391
	(0.09)	(0.20)	(0.21)
NYC	-0.182***	-0.252***	-0.028
	(0.02)	(0.03)	(0.04)
NYCxAsian	0.012	0.022	0.036
	(0.02)	(0.02)	(0.02)
NYCxBlack	0.028	0.081***	0.047*
	(0.02)	(0.02)	(0.02)
NYCxHispanic	0.01	0.009	0.062***
	(0.02)	(0.02)	(0.02)
NYCxLow Income	0.051**	-0.031*	-0.027
	(0.02)	(0.02)	(0.02)
NYCxBubble	0.009	-0.043*	-0.024
	(0.01)	(0.02)	(0.02)
NYCxProportion Asian	0.358***	0.421*	0.315
	(0.09)	(0.17)	(0.19)
NYCxProportion Black	0.229***	0.206*	0.264*
	(0.06)	(0.09)	(0.11)
NYCxProportion Hispanic	0.169**	0.101	0.219
	(0.06)	(0.10)	(0.12)
NYCxProportion Low Income	-0.058	0.102	-0.164
	(0.06)	(0.09)	(0.10)
NYCxProportion Bubble	0.054	0.037	0.31
	(0.17)	(0.34)	(0.36)
N	185,047	185,522	186,550

Table 8

Hierarchy of Multilevel Regression Models, Grade 7

	Model 1	Model 2	Model 3	Model 4	Model 5
Constant	0.009	0.009	0.007	0.006	0.058***
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
Asian	-0.107***		-0.121***	-0.115***	-0.118***
	(0.01)		(0.01)	(0.01)	(0.01)
Black	0.074***		0.055***	0.051***	0.02
	(0.01)		(0.01)	(0.01)	(0.01)
Hispanic	0.140***		0.118***	0.114***	0.119***
	(0.01)		(0.01)	(0.01)	(0.01)
Low Income	-0.634***		-0.747***	-0.620***	-0.578***
	(0.07)		(0.07)	(0.07)	(0.12)
Bubble status	-0.181***		-0.336***	-0.324***	-0.305***
	(0.03)		(0.03)	(0.03)	(0.05)
Proportion Asian	-0.193***		-0.377***	-0.361***	-0.199***
	(0.03)		(0.04)	(0.04)	(0.06)
Proportion Black		0.095***	0.081***	0.075***	0.082***
		(0.01)	(0.01)	(0.01)	(0.01)
Proportion Hispanic		-0.102***	0.201***	0.151***	0.203***
		(0.02)	(0.04)	(0.04)	(0.04)
Proportion Low Income				0.246***	0.259***
				(0.01)	(0.01)
Proportion Bubble Students				0.880***	0.735***
				(0.17)	(0.20)
NYC					-0.252***
					(0.03)
NYCxAsian					0.022
					(0.02)
NYCxBlack					0.081***
					(0.02)
NYCxHispanic					0.009
					(0.02)
NYCxLow Income					-0.031*
					(0.02)
NYCxBubble					-0.043*
					(0.02)
NYCxProportion Asian					0.421*
					(0.17)
NYCxProportion Black					0.206*
					(0.09)
NYCxProportion Hispanic					0.101
					(0.10)
NYCxProportion Low Income					0.102
					(0.09)
NYCxProportion Bubble					0.037
					(0.34)
N	185,522	185,522	185,522	185,522	185,522

Figures

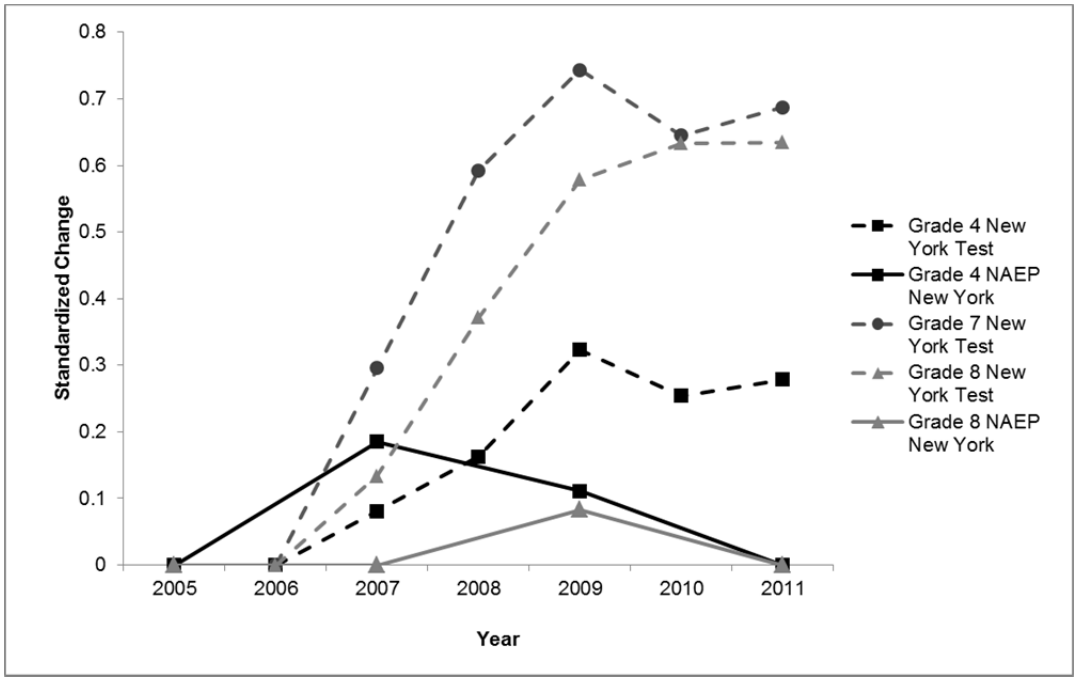

Figure 1. Trends (standardized mean change) in Mathematics on New York Tests and NAEP.

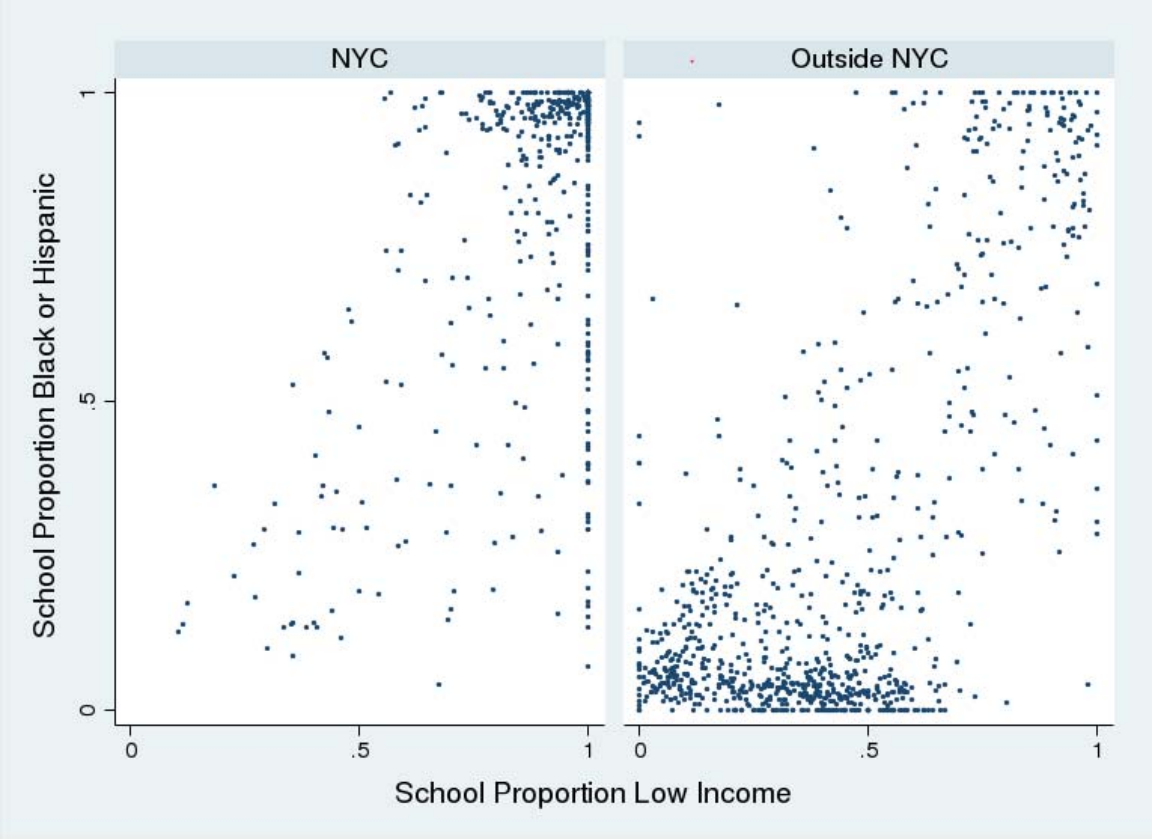

Figure 2. School proportion Black or Hispanic plotted against proportion low income, within and outside New York City.

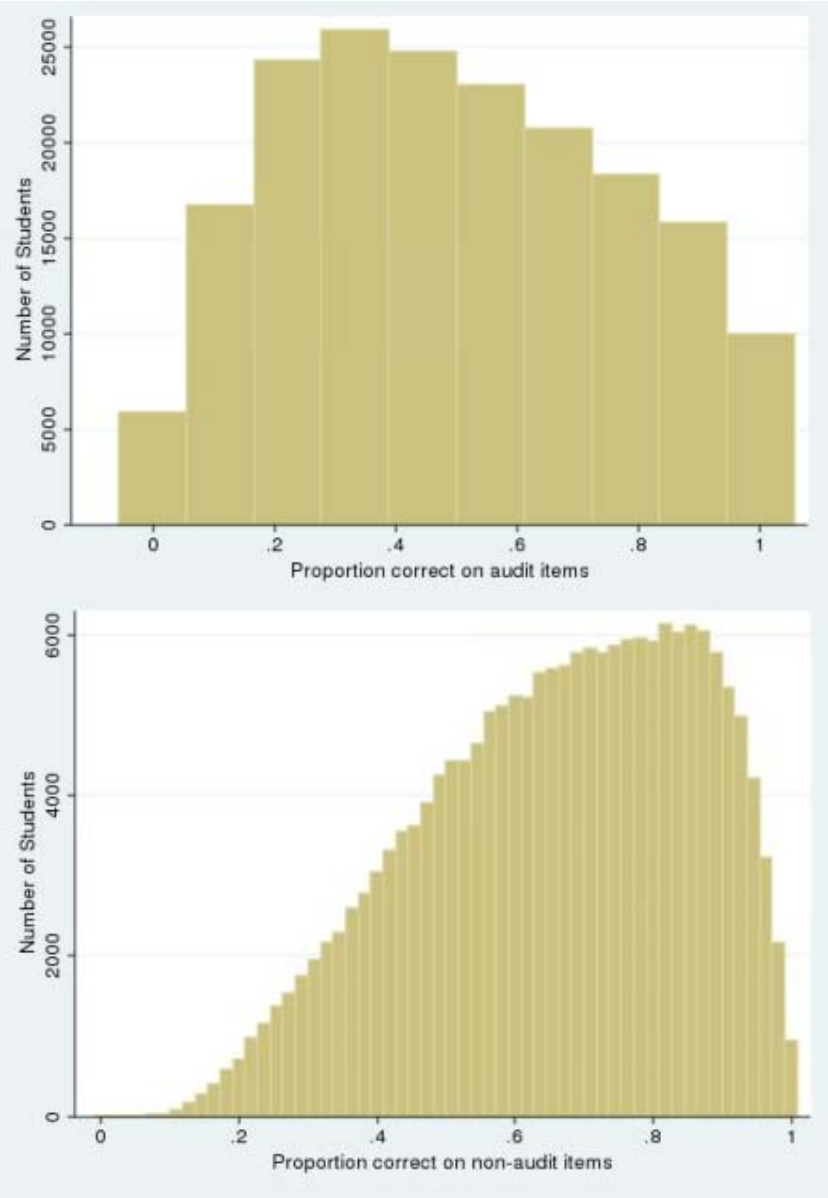

Figure 3. Student-level distributions of proportion-correct on the audit test (top panel) and non-audit test (bottom pane), grade 7.

Appendix A: The Grade 7 Audit Test

Table A1

Characteristics of Items Used to Construct the Grade 7 Audit Test

Item	Strand	Standard	Standard Description	P-Value	Item-Rest Correlation
3	Number Sense and Operations	7N14	Develop a conceptual understanding of negative and zero exponents with a base of ten and relate to fractions and decimals (e.g., 10^{-2} = .01 = 1/100)	.32	.38
6	Geometry	7G5	Identify the right angle, hypotenuse, and legs of a right triangle	.65	.40
8	Statistics and Probability	7S1	Identify and collect data using a variety of methods	.40	.22
16	Statistics and Probability	7S5	Select the appropriate measure of central tendency	.50	.36
25	Algebra	7A5	Solve one-step inequalities (positive coefficients only)	.73	.31
38	Number Sense and Operations	7N17	Classify irrational numbers as non-repeating/non-terminating decimals	.44	.39
45	Algebra	7A8	Create algebraic patterns using charts/tables, graphs, equations, and expressions	.38	.46
49 (Constructed Response)	Algebra	7A8	Create algebraic patterns using charts/tables, graphs, equations, and expressions	.98	.42

Appendix B. Sensitivity Tests for the Bubble Student Classification

Table B1

Comparison of Multi-level Regression Results Using Three Different Specifications for the Bubble Predictor, Grade 7

	3 Raw Score Points Below Proficient[a]	2 Raw Score Points Above and Below Proficient	2 Raw Score Points Below Proficient
Number of Bubble Students	17058	25348	11738
Proportion of Analytic Sample	9.19%	13.66%	6.33%
Constant	0.058***	0.059***	0.060***
	(0.01)	(0.01)	(0.01)
Asian	-0.118***	-0.114***	-0.120***
	(0.01)	(0.014)	(0.014)
Black	0.02	0.019	0.022*
	(0.01)	(0.011)	(0.011)
Hispanic	0.119***	0.117***	0.120***
	(0.01)	(0.011)	(0.011)
Low Income	0.082***	0.084***	0.085***
	(0.01)	(0.007)	(0.007)
Bubble	0.259***	0.267***	0.250***
	(0.01)	(0.008)	(0.011)
Proportion Asian	-0.578***	-0.578***	-0.608***
	(0.115)	(0.114)	(0.114)
Proportion Black	-0.305***	-0.297***	-0.305***
	(0.045)	(0.045)	(0.046)
Proportion Hispanic	-0.199***	-0.175**	-0.197***
	(0.058)	(0.058)	(0.059)
Proportion Low Income	0.259***	0.197***	0.214***
	(0.01)	(0.042)	(0.042)
Proportion Bubble	0.735***	0.689***	0.887***
	(0.2)	(0.163)	(0.255)
NYC	-0.252***	-0.248***	-0.261***
	(0.03)	(0.03)	(0.03)
NYCxAsian	0.022	0.019	0.022
	(0.02)	(-0.022)	(0.022)
NYCxBlack	0.081***	0.080***	0.081***
	(0.02)	(0.02)	(0.02)
NYCxHispanic	0.009	0.008	0.009
	(0.02)	(0.018)	(0.018)
NYCxLow Income	-0.031*	-0.031*	-0.033*

	(0.02)	(0.015)	(0.015)
NYCxBubble	-0.043*	-0.035*	-0.029
	(0.02)	(0.014)	(0.02)
NYCxProportion Asian	0.421*	0.446**	0.443**
	(0.17)	(0.167)	(0.168)
NYCxProportion Black	0.206*	0.211*	0.227*
	(0.09)	(0.09)	(0.09)
NYCxProportion Hispanic	0.101	0.087	0.118
	(0.1)	(0.103)	(0.104)
NYCxProportion Low Income	0.102	0.087	0.098
	(0.09)	(0.09)	(0.09)
NYCxProportion Bubble	0.037	0.026	-0.117
	(0.34)	(0.271)	(0.418)
N	185,522	185,522	185,522

[a] This is the bubble specification reported in the body of the paper.

Appendix C. Sensitivity Tests for Standardization of the Outcome Variable

Table C1

Comparison of Grade 7 Multi-level Regression Results Using Two Different Standardization Methods for the Outcome Variable

	Standardized Difference of Non-audit and Audit Scores	Difference of Standardized Audit and Standardized Non-audit Scores
Constant	0.058***	0.006
	(0.01)	(0.01)
Asian	-0.118***	0.016
	(0.01)	(0.01)
Black	0.02	-0.092***
	(0.01)	(0.01)
Hispanic	0.119***	0.010
	(0.01)	(0.01)
Low Income	0.082***	-0.060***
	(0.01)	(0.00)
Bubble	0.259***	0.078***
	(0.01)	(0.01)
Proportion Asian	-0.578***	-0.166
	(0.12)	(0.09)
Proportion Black	-0.305***	-0.139***
	(0.05)	(0.03)
Proportion Hispanic	-0.199***	-0.058
	(0.06)	(0.04)
Proportion Low Income	0.203***	-0.031
	(0.04)	(0.03)
Proportion Bubble	0.735***	0.046
	(0.20)	(0.15)
NYC	-0.252***	-0.066**
	(0.03)	(0.02)
NYCxAsian	0.022	0.025
	(0.02)	(0.02)
NYCxBlack	0.081***	0.057***
	(0.02)	(0.01)
NYCxHispanic	0.009	-0.004
	(0.02)	(0.01)
NYCxLow Income	-0.031*	0.033**
	(0.02)	(0.01)
NYCxBubble	-0.043*	0.025*
	(0.02)	(0.01)
NYCxProportion Asian	0.421*	0.004

	(0.17)	(0.13)
NYCxProportion Black	0.206*	-0.115
	(0.09)	(0.07)
NYCxProportion Hispanic	0.101	-0.204**
	(0.10)	(0.08)
NYCxProportion Low	0.102	0.170*
Income	(0.09)	(0.07)
NYCxProportion Bubble	0.037	-0.269
	(0.34)	(0.25)
N	185,522	185,522

*p<0.05, **p<0.01, ***p<0.001