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
http://projects.iq.harvard.edu/eap

Hui Leng Ng
Harvard Graduate School of Education

Daniel Koretz
Harvard Graduate School of Education

Jennifer L. Jennings
Department of Sociology
New York University

Acknowledgement

The research reported here was supported by the Institute of Education Sciences, U.S. Department of Education, through Grant R305AI0420, and by the Spencer Foundation, through Grants 201100075 and 201200071, 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, the Spencer Foundation, or the New York State Education Department or its staff.

1 Hui Leng Ng is currently at the Singapore Ministry of Education. She may be contacted at hui_leng_ng@mail.harvard.edu or 285 Ghim Moh Road, Singapore 279622.
Abstract

A weakness of most previous score-inflation studies is that differences in performance between a high-stakes test and a readily available “audit” test could reflect factors other than inflation. Koretz and Beguin (2010) proposed a “self-monitoring assessment” (SMA) intended to eliminate such factors. Using pilot data from a first-ever designed SMA—comprising past state-test items (non-audit subtest) and specially-designed audit items (audit subtest)—administered to a random sample of 4th graders in New York State in 2011, we demonstrated empirically the feasibility of designing an SMA to measure score inflation associated with coaching. To examine the practical impact of using inflated scores for accountability, we also used the pilot data to investigate the consistency of school-performance ratings on the two subtests. We found that schools’ ratings, using various school-performance measures, varied substantially between the subtests, thereby suggesting that the ratings based on the non-audit subtest reflect schools’ relative engagement in coaching.
Sensitivity of School-Performance Ratings to Score Inflation:
An Exploratory Study Using a Self-Monitoring Assessment

Introduction

In recent years, student scores on standardized tests of academic achievement have grown in importance as indicators of schools’ success, in the US and internationally (Linn, 2004; Mathison, 2009; Organisation for Economic Cooperation and Development [OECD], 2008; Torrance, 2009). But how trustworthy are ratings of schools’ performance based on measures derived from scores obtained from high-stakes testing, where these scores could be inflated (i.e., higher than the actual achievement levels they represent)?

While the phenomenon of score inflation is well documented in the US (Koretz & Hamilton, 2006), existing research is limited in two ways with regard to its impact on school-performance measures. First, existing studies have investigated score inflation only at the state or district levels (e.g., Fuller, Gesicki, Kang, & Wright, 2006; Jacob, 2005, 2007; Klein, Hamilton, McCaffrey, & Stecher, 2000), but not at the school level. More importantly, these studies have relied typically on results from readily available “audit” tests, not designed specifically to detect score inflation, to evaluate performance on the high-stakes test. For example, all the studies cited above made use of scores on a test (e.g., NAEP) that may differ from the high-stakes test in ways other than the stakes for educators. These include differences in test-curriculum alignment, tested-student populations, and students’ motivational levels. Thus, any discrepancy in performances on the high- and lower-stakes tests could also be due to these factors, rather than score inflation. To address this, Koretz and Beguin (2010) proposed a “self-monitoring assessment” (SMA) that incorporates specially designed audit items into an
operational assessment directly to eliminate specific non-inflation-related factors, thereby providing a measure of score inflation that is free from such potential confounders.

In this study, we used data from the first pilot implementation of an SMA. This SMA, designed by the Education Accountability Project at the Harvard Graduate School of Education, was administered to a statewide random sample of 4th graders in New York State (NYS) in Spring 2011. Students were administered two sets of test items. The first (the “non-audit subtest”) comprised items from past NYS state tests that were deemed likely to be the focus of inappropriate test preparation. The second (the “audit subtest”) comprised items designed to test similar content but to be less susceptible to a particular type of test preparation—“coaching”—that we expected to be directed at the past items in the non-audit subtest. The two sets of items were interspersed in random order and were not distinguished to the students’ view. To the extent that the design of the audit items had eliminated the effects of other factors, discrepancies in performance between the two subtests is a measure of score inflation.

As this was a first pilot test of the SMA design, we first verified the extent to which the audit test provides a valid measure of score inflation. Then, to investigate the practical impact of using inflated scores for accountability, we examined the consistency of schools’ performance ratings on the two subtests. Specifically, we asked:

**RQ1.** *To what extent does the difference in student performance on the two subtests provide a measure of score inflation?*

**RQ2.** *How consistent are schools’ performance ratings when we use scores from the two subtests to derive the school-performance estimates?*

In the rest of the article, we first describe the theoretical set up underlying the use of two measures of the same outcome to detect score inflation, and how the SMA framework proposed by Koretz and Beguin (2010) provides a way to detect score inflation caused by inappropriate
test preparation that is free from potential confounding by specific non-inflation-related factors. Then, we review how coaching could inflate students’ scores, which formed the basis for designing the SMA used in the study. After briefly describing the context of the high-stakes school-accountability system in NYS, we set out the research design for the study and present our key findings. We conclude by discussing the implications and limitations of the study.

**Detecting Score Inflation using Two Measures of the Same Outcome**

There are five main factors associated with systematic differences between two standardized tests in the same academic subject which could result in inconsistent school ratings when one test is used rather than the other: (1) alignment between schools’ implemented curricula and the content mixes of the tests; (2) timing of the tests; (3) students’ motivational levels while taking the tests; (4) test-administration procedures; and (5) tested-student populations. Factors (2) and (3) are different aspects of the occasion of testing that are potentially confounded.

Further, for each factor, the differences between the tests could be due to either non-inflation-related sources or schools’ behavioral responses to high-stakes use of the results of one test but not the other (i.e., induced by stakes). For example, variations in test-curricula alignment among schools could arise when schools adopt different content focus or pedagogical approaches for non-stakes related reasons such as values orientation and curriculum-design model adopted (Marsh, 2009). They could also arise when some schools engage in certain inappropriate test-preparation activities focused on tested materials on the high-stakes test (see later).

For detecting and measuring score inflation, the difference in scores from two tests that differ in both non-inflation-related and stakes-induced aspects thus confounds the effects of the two sources of differences. This is a known weakness of previous score-inflation studies that
have relied on readily available “audit” tests, as discussed by other researchers (e.g., Applegate, Applegate, McGeehan, Pinto, & Kong, 2009; Center on Education Policy, 2010; Corcoran, Jennings, & Beveridge, 2011; Jacob, 2002; Jirka & Hambleton, 2004; Koretz & Beguin, 2010; Wei, Shen, Lukoff, Ho, & Haertel, 2006). Typically, past score-inflation studies addressed this limitation by introducing the dimension of time, thereby using discrepant score gains over time on the two tests as indicative of the presence of score inflation. This approach could not detect score inflation at any single time point.

The SMA framework proposed by Koretz and Beguin (2010) thus represents an important methodological breakthrough because it is intended to detect score inflation at a single time point. Before we describe how it does so, we first formalize the confounding in the use of the difference in scores between two tests as a measure of score inflation at a single time point.

Assume that we are interested in making an inference about students’ achievement in a particular content domain. Consider two tests that are designed to measure achievement in that domain, one of which is the target high-stakes test that we want to investigate the incidence of score inflation, and the other is an audit test that we want to use to make that evaluation. For simplicity of exposition, let us also assume that scores on the audit test are free from inflation.² Then, for student $i$ in school $s$, let

$$Y_{is}^{\text{Target}} = \theta_{is} + \eta_{is}^{\text{Target}} + \epsilon_{is}^{\text{Target}} \quad \text{(1.1)}$$

$$Y_{is}^{\text{Audit}} = \theta_{is} + \eta_{is}^{\text{Audit}} + \epsilon_{is}^{\text{Audit}} \quad \text{(1.2)}$$

² Without this assumption, the amount of school-level score inflation in the audit-test scores would add—since score inflation is positive by conception—to the amount of bias in the estimator for the school-level score inflation in the target test that we describe in the main text. This would result in a conservative bias in the estimator.
where $Y_{is}^{\text{Target}}$ and $Y_{is}^{\text{Audit}}$ are scores on the target and audit tests respectively; $\theta_{is}$ is the true achievement level on the content domain underlying the target inference; $\xi_{is}$ is the deviation between the observed score on the target test and the true achievement level that is attributable to the systematic, inflation-related factors influencing achievement on the target test (i.e., score inflation), and is positive by conception; $\eta_{is}^{\text{Target}}$ and $\eta_{is}^{\text{Audit}}$ are the deviations between the observed scores, on the target and audit tests respectively, and the true achievement level that are attributable to systematic, non-inflation-related factors influencing achievement on the respective tests; and $e_{is}^{\text{Target}}$ and $e_{is}^{\text{Audit}}$ are random sources of variation in the target and audit tests respectively, which are independent both within and between the two tests, and are homoscedastic with mean zero within each test, for all $i$ and $s$.

Further, assuming that the scores on the two tests are on the same scale, the student-level score inflation is then given by:

\[
(2) \quad \xi_{is} = \left( Y_{is}^{\text{Target}} - Y_{is}^{\text{Audit}} \right) - \left( \eta_{is}^{\text{Target}} - \eta_{is}^{\text{Audit}} \right) - \left( e_{is}^{\text{Target}} - e_{is}^{\text{Audit}} \right)
\]

Similarly, by defining school $s$’s score inflation to be the average score inflation in its students’ scores, the corresponding school-level aggregate score inflation is given by:

\[
(3) \quad \bar{\xi}_{s} = \left( \bar{Y}_{s}^{\text{Target}} - \bar{Y}_{s}^{\text{Audit}} \right) - \left( \bar{\eta}_{s}^{\text{Target}} - \bar{\eta}_{s}^{\text{Audit}} \right) - \left( \bar{e}_{s}^{\text{Target}} - \bar{e}_{s}^{\text{Audit}} \right)
\]

Being the only observed quantity in equation (3), the difference between the average scores on the two tests for school $s$, represented by $\left( \bar{Y}_{s}^{\text{Target}} - \bar{Y}_{s}^{\text{Audit}} \right)$, is a natural estimator for its aggregate score inflation. However, it is clear from equation (3) that it is a biased estimator, with amount of bias equals to $E\left( \bar{\eta}_{s}^{\text{Target}} - \bar{\eta}_{s}^{\text{Audit}} \right)$. This quantity denotes the expected difference between school $s$’s average scores on the two tests associated with the systematic,
non-inflation-related sources of differences, such as the non-stakes-induced parts of the five factors above.

The SMA framework proposed by Koretz and Beguin (2010) addresses this bias in \( \left( \bar{y}_{s}^{\text{Target}} - \bar{y}_{s}^{\text{Audit}} \right) \) as a measure of the school-level score inflation in two ways. First, an SMA incorporates audit items directly into an operational assessment so that both the audit and operational items are administered at the same time to the same students. In this way, it eliminates sources of differences—both non-inflation-related and induced by stakes—between the two types of items that are associated with the timing of the tests, students’ test-taking motivational levels, test-administration procedures, and tested-student populations (i.e., factors (2) to (5) above). Thus, in theory, an SMA provides a measure of score inflation that is independent of these four sources of differences, even if the differences were induced by stakes.

Secondly, and more critically, by designing the audit items carefully to be “less susceptible to inflation [caused by inappropriate test preparation]...compared to more vulnerable items in the operational assessment” (Koretz & Beguin, 2010, p. 98), an SMA attempts to isolate the influence of differences in test-curricula alignment—i.e., source (1) earlier—that are induced by stakes from those that are not. More specifically, as the authors explained, audit items can be designed to measure aspects that are important to the target inference (i.e., the construct denoted by \( \theta_{is} \)), but which are neglected by inappropriate test-preparation activities. Students who received only inappropriate test preparation would find such audit items relatively more difficult than the operational items that were targets of their inappropriate test preparation. They would thus perform worse on the audit items than on the corresponding targeted operational items. In this case, to the extent that the non-inflation-related sources of differences between the two sets of items have been eliminated effectively (i.e., \( E\left( \bar{y}_{s}^{\text{Target}} - \bar{y}_{s}^{\text{Audit}} \right) = 0 \)), the difference in
average performance on the two sets of items for school \( s \), \( \left( \bar{\gamma}_s^{\text{Target}} - \bar{\gamma}_s^{\text{Audit}} \right) \), provides a measure of its aggregate score inflation associated with inappropriate test preparation.

The key to designing an SMA that could detect and measure score inflation associated with inappropriate test preparation thus lies in understanding what such activities entail and how they could inflate scores. This allows us to design suitable audit items that would counteract the effects of such activities. Therefore, we next review inappropriate test-preparation activities that schools had adopted under pressures from high-stakes testing conditions, focusing on the type that is relevant to the design of the particular SMA used in the study: coaching.

**Inappropriate Test-Preparation Activities that Could Lead to Score Inflation**

Test preparation is not always problematic nor does it necessarily result in score inflation under all circumstances (Koretz & Hamilton, 2006; Mehrens & Kaminski, 1989; Miyasaka, 2000). For example, defining test preparation as “all steps educators take, both desirable and undesirable, to prepare students for tests” (p. 548), Koretz & Hamilton (2006) discussed different types of test preparation. According to them, three of these—“teaching more”, “working harder”, and “working more effectively”—could lead to scores that are commensurate with real gains in actual student learning. However, one of them, “cheating”, always results in inflated scores, while two others, “reallocating” and “coaching”, could result in score inflation under certain circumstances.

An SMA is intended to detect score inflation caused by certain test-preparation activities involving reallocation and coaching. Specifically, it targets activities that shift instructional emphasis within a particular tested subject area (i.e., within-subject reallocation), and those that focus narrowly on specific features of items that recur on the tests (i.e., coaching) (Koretz & Beguin, 2010). When such within-subject reallocation or coaching results in students’
performance on the tested materials being unrepresentative of their performance on materials in the broader domain underlying the target inference, score inflation occurs, albeit through conceptually different mechanisms (Koretz & Hamilton, 2006). We focus only on coaching for the rest of the section because of its relevance to the present study.

**Coaching**

Many past studies have documented that educators responded to high-stakes testing by narrowing their instructional emphasis to the specific features of items that recur on the high-stakes tests (Firestone, Mayrowetz, & Fairman, 1998; Herman & Golan, 1993; Koretz & Barron, 1998; Lomax et al., 1995; Nichols & Berliner, 2007; Pedulla et al., 2003; Shen, 2008; Sloan, 2004; Smith, 1991). Such item features could be content related or simply idiosyncratic surface characteristics with little to do with the construct being measured. Koretz and Hamilton (2006) referred to excessive instructional emphases on these two types of item features as “substantive coaching” and “non-substantive coaching” respectively.

For instance, consider an inference about students’ understanding of linear relationships between two variables. If the teachers noticed that all past items on this topic on the high-stakes test involved only positive linear relationships, they could narrow their instruction to these and ignore negative linear relationships. Students who had been thus coached (substantively) may not necessarily know how to solve problems involving negative linear relationships. If the teachers further noticed that all past items on this topic involved asking students to identify the slope when given an equation written in the form “\( y = a + bx \)”, where the two variables were invariably “\( y \)” and “\( x \)” and the value of \( b \) was invariably positive, then they could further narrow their instruction to a simple rule that says “when you see the word ‘slope’, the answer is the number beside ‘\( x \)’”. In this case, as long as the future items appear in the same exact form
where the rule applies, students who had been thus coached (non-substantively) would be able to answer such items correctly, even without understanding anything substantive about slopes and linear equations. However, such students would likely have difficulties with an item involving a linear relationship such as “\(m + 5n = 4\)”, for there is no “\(x\)” in it, nor is the slope “5”.

Both forms of coaching could take place through several types of activities. As Nichols and Berliner (2007) wrote,

“We found numerous examples from schools across the country that had dedicated hours upon hours preparing students for the test—drilling, emphasizing rote memorization, teaching students how to take tests, reviewing over and over again the concepts that will be represented on the test, and giving multiple practice tests, all at the expense of other content, ideas, and curricula that may not be represented on the test.” (p. 122)

This is consistent with what other researchers in the studies cited earlier have documented.

Such activities lead to decontextualized learning that at best develops “highly restricted skills that may be of little use in life activities” (Anastasi, 1981, p. 1089). They might be effective for short-term gains on tests but are unhelpful for actual learning, especially of conceptually more complex content. For example, Koretz and Barron (1998) provided evidence that item-specific coaching was associated with gains on mathematics tests in KIRIS. More recently, Shen (2008) found that teachers in her study tended to focus on the non-substantive aspects (e.g., item format, syntax of problems, step-by-step rules, algorithms) when describing how they teach the content of items that they had classified as “difficult to teach”. In contrast, she found them discussing the actual content matter more frequently when approaching the content of items that they had classified as “easy to teach”.

In all cases, narrowed instruction focused only on specific features of items that recur on past tests makes students’ performance on such items that appear on future tests unrepresentative of their actual understanding of the specific content area tested by the items. As Koretz and
Hamilton (2006) explained, unlike within-subject reallocation, which makes students’ performance on the tested content areas unrepresentative of their performance on the untested ones, both substantive and non-substantive coaching make students’ observed performance unrepresentative of their actual understanding of even the particular content area that is tested.

As such, audit items designed to test the neglected content aspects of recurrent topics on past tests, and those that use novel representations to elicit responses on such topics, could detect potential score inflation due to substantive and non-substantive coaching respectively.

**Test-Based Accountability in NYS and Opportunity to Administer SMA**

NYS has a long history of test-based accountability. It administered its first state assessments to 8th and 12th graders in 1865 and 1878 respectively (New York State Education Department [NYSED], 2010). Originally optional and used only for student accountability, the state’s implementation of an equity-driven “Action Plan” in 1984 raised the stakes of the state assessments considerably for both students and schools (Johnson, 2009). First, it made the High School Regents Examinations in several academic subjects—English, mathematics, science, global studies, and US history and government—mandatory for high-school graduation for all students. It also implemented its first test-based accountability system for schools in 1989 (NYSED, 2009a). The system required schools to publish annual reports on the aggregate achievements of their students on several indicators (e.g., dropout rates, results on the Regents Competency Tests). It also subjected schools to a registration review process, under which those whose students were assessed to be performing below expectation faced de-registration and closure if their students did not improve adequately within a given timeframe. Both the public school reports and the school-evaluation system persisted to today, taking the forms of the New York State Report Cards and a complex tiered system of “Differentiated Accountability” catering
to schools assessed to be in need of different levels of improvement interventions (NYSED, 2009b, 2012).

The opportunity to design and administer an SMA in NYS came as part of the state’s efforts to restore public confidence in its testing program. Growing skepticism of the credibility of rapidly rising state-test scores over the past decade, especially when viewed in relation to the absence of comparable gains of the state’s performance on the NAEP, came to a head in 2009. Accusations such as state officials lowering proficiency cut-scores deliberately in order to raise state-test results were played out prominently in the public media, fueling increasing distrust in the system (e.g., Hernandez, 2009; Medina, 2010; Ravitch, 2009a, 2009b; “Shock”, 2010; Stern, 2010; Winerip, 2011). In response, the state’s then-newly appointed Chancellor and Commissioner took a series of unprecedented steps to restore public confidence in the state’s testing program. These included commissioning the Education Accountability Project at the Harvard Graduate School of Education to conduct an audit study to investigate whether the state-test scores were indeed inflated. The present study grew out of that larger study.

Data

Source

The NYS 2011 audit dataset that we have used is a subset of the data from the first trial of the SMA developed by the Education Accountability Project and administered in a statewide mathematics field-test to a random sample of students in Spring 2011. In the field test, the SMA items were placed in 12-item audit-test booklets as specified by NYS’s testing contractor, the overseer of the entire field test. These audit-test booklets, together with other field-test booklets containing items that the testing contractor was developing, were allocated randomly to students using an overall matrix-sampling schedule that the testing contractor drew up. Each student in
the field test who was administered an audit-test booklet was administered only one of them. For our study, we used data from two 4th grade audit-test booklets (i.e., 24 SMA items in total). Each audit-test booklet comprised an audit subtest and a non-audit subtest. We describe the construction of the items in the two audit-test booklets later.

**Measures**

**Outcome variables.** The outcome variables were students’ mathematics achievement measured by (1) scores on the audit subtest ($A_{SCORE}$); and (2) scores on the non-audit subtest ($N_{SCORE}$). Both were raw proportion-correct scores recording the proportion of the subtest items that a student answered correctly.

We had to use raw proportion-correct scores because of the idiosyncrasies of the data-collection process that generated the pilot data. As each student who was allocated an audit-test booklet in the field test took exactly one audit-test booklet, and there is no overlap between items on the different audit-test booklets, we could not create a common scale for items on the subtests.

In addition, as each audit-test booklet is further divided into the audit and non-audit subtests, each 6-item long, there is considerable measurement error in the raw proportion-correct scores associated with item sampling on either subtest. This is evident from the estimated internal-consistency reliability (Cronbach’s alpha) for the subtests, ranging from .57 to .67 across subtests, and booklets. We discuss how we addressed this later in the “Methods” section.

**Control predictors.** In deriving our school-performance measures, we used models that included various combinations of the following:

*Achievement in Previous Year.* Students’ mathematics achievement in 2010 was measured by scale scores on the 2010 mathematics state test ($STATE_{2010}$).
Audit-Test Booklet. A vector of dichotomously-coded indicators ($\textit{BOOK}$) recorded the particular audit-test booklet that a student took.

Student Background. Three sets of dichotomously-coded covariates recorded selected student-background characteristics: (1) gender, family low-income status, immigrant status, and several race/ethnicity categories; (2) limited-English-proficient (LEP), and disability statuses; and (3) whether a student had received testing accommodations while taking the state tests in 2011. We denote these covariates collectively by the vector $B$.

School-level Aggregate Variables. We also derived aggregate school-level measures by averaging up the corresponding student-level variables for all 4th graders (i.e., instead of just those who had participated in the field-test and were administered an audit-test booklet) in each school in 2011. We denote these collectively by the vector $S$.

Construction of Analytic Sample

The total field-test sample in the 4th grade audit dataset that we used comprised 4,555 students in 1,050 schools. They constituted 2.3% of the state’s 4th grade population in 2011—henceforth, “population”—and 43.9% of the schools that have 4th graders in 2011.

The total field-test sample is comparable to the population in terms of most control predictors. However, white students were slightly over-represented in sample (52%) compared to the population (49%), and African-American/Hispanic students were correspondingly under-represented in the sample (38%) compared to the population (41%). Students from low-income families were also slightly under-represented—52% in the sample compared to 56% in the population. Nonetheless, at the school level, the total field-test sample has means that are comparable along all the observed characteristics to those in the population.
We constructed the analytic sample by first retaining student-level observations that were non-missing on all variables required to compute the school-performance measures. Then we retained only schools with at least three students satisfying the student-level inclusion criterion.

The need to set such an arguably small minimum sample size is another consequence of the idiosyncrasies of the data-collection process that generated the pilot data. The 2 audit-test booklets that we used for the study were only part of a total of 15 field-test mathematics booklets of all types that were allocated randomly to students in the field test. Consequently, the random sample of students who were administered the audit-test booklets within each school was only a subset of the entire sample of students who participated in the field test in the school. This resulted in some schools having very few students with audit-test data. In fact, the largest sample size in any school in the analytic sample was only 17. This in turn necessitated adopting the small minimum sample-size in order to retain a reasonable number of schools.

The resulting analytic sample comprised 3,918 students in 778 schools. These represent attrition rates of 14.0% and 25.9% at the student and school levels respectively. Nonetheless, the analytic sample is comparable to the total field-test sample with respect to all outcome and control predictors, both at the student and school levels.

**Methods**

**Design of the SMA**

The SMA that was pilot-tested was designed with the intention to eliminate potential confounding by non-inflation-related factors as follows. First, each audit-test booklet comprised 6 audit items and 6 non-audit items, all of which were in multiple-choice format and were administered to the same students on a single occasion. Although the two sets of items were analytically labeled as the audit subtest and non-audit subtest respectively, the items were
interspersed in random order in each booklet and were not distinguished to the students’ view. All these eliminated confounding arising from differences in tested-student population, timing, administration conditions, and motivation. Scoring effects were not germane because all items were in multiple-choice format.

In addition, the two booklets included in the present study were specially designed with the intention to further eliminate potential confounding from content differences. This was done by closely matching each non-audit item to an audit item that tested the same specific content. The two items in each non-audit/audit pair differed only in that the audit item lacked one or more predictable aspects in its closely matched non-audit item. These aspects could be predictable in terms of the recurrence of specific content aspects or surface features of items on past tests, which facilitate substantive and non-substantive coaching respectively. For instance, suppose all previous state-test items that evaluated students’ understanding of graphical representations of simple linear relationships presented only positive-sloping lines in the region where both variables are positive—“quadrant 1” of the four-quadrant two-dimensional space. In this case, the non-audit item would present a positive-sloping line in quadrant 1, while a possible matched audit item might use a line that was negative sloping, in a different quadrant, or both. We hypothesized that students who received only inappropriate instruction on positive linear relationships in quadrant 1 would find the audit item more difficult than the non-audit item.

In sum, the two subtests used in the study were matched, by design, in several important respects—namely, tested-student population, timing, administration conditions, students’ motivational levels, scoring, and content—with the intention of eliminating potential confounding from these sources of differences. The audit items were also carefully constructed with the intention to measure score inflation due to substantive or non-substantive coaching.
Investigating the Functioning of the SMA as a Measure of Score Inflation

We investigated two sources of evidence that would evaluate the validity of the audit-test data as a measure of score inflation (RQ1). Since these analyses were conducted to validate the score-inflation measure rather than to estimate the subtest effects, we conducted the analyses using the total field-test samples instead of the analytic samples.

First, we checked whether the audit items had functioned as designed. If they had, students would exhibit systematically poorer performance on the audit subtest versus the non-audit subtest. This is because the audit items were designed to be expectedly more difficult, relative to the matched non-audit items, for students who have received only coaching. Therefore, we computed the average first-order difference, $\bar{D}$, given by:

$$
\bar{D} = \frac{1}{N} \sum_{s=1}^{N} \sum_{i=1}^{n_s} (D_{is}) = \frac{1}{N} \sum_{s=1}^{N} \sum_{i=1}^{n_s} (N_{-SCORE_{is}} - A_{-SCORE_{is}})
$$

A positive $\bar{D}$ indicates that the audit items had functioned as designed, a requisite for using the audit-test data to detect score inflation.

However, a positive $\bar{D}$ is necessary but insufficient to validate that the audit-test data provide a valid measure of score inflation. This is because, although the two subtests were closely matched to eliminate potential confounding from several important sources of differences, we could not eliminate potential confounding associated with incidental differences between the non-audit and audit items, unrelated to the differences designed to detect inflation, that affect the relative difficulties of the two types of items. This is inevitable because without additional data, we cannot determine the “true” relative difficulties of the two subtests in a testing situation void of inappropriate test preparation. Koretz and Beguin (2010) suggested
evaluating this by administering the non-audit and audit items to students for whom neither test is high-stakes. However, this is not available in the pilot data. Therefore, in the absence of information about the “true” relative difficulties between the two types of items, students could perform differently on the audit and non-audit items simply because the two types of items were truly differently difficult, not just because of score inflation.

To address this potential confounding, we used a “difference-in-differences” strategy where we compared the first-order differences across groups of students we expect to have different degrees of score inflation based on past research (i.e., second-order differences). Past research suggests that minority students, poor students, and students in schools with high proportions of minority or poor students, tended to be exposed to more inappropriate test preparation than others (Firestone, Camilli, Yurecko, Monfils, & Mayrowetz, 2000; Herman & Golan, 1993; Lomax et al., 1995; McNeil & Valenzuela, 2000; Monsaas & Engelhard Jr., 1994; Shen, 2008). In addition, past research has also documented consistently the “bubble-kid syndrome” whereby students near the boundary of proficiency cut-scores were more often targeted for inappropriate test preparation than those at either ends of the proficiency spectrum (Booher-Jennings, 2005; Jacob, 2005; Neal & Schanzenbach, 2009; Reback, 2008).

Insofar as the inappropriate test preparation inflates students’ scores on the content targeted by such activities, we would thus expect the students who had been exposed to more of such activities to show larger score inflation on tests containing the targeted content. This is supported by some empirical evidence showing that minority students have larger average score gaps between two tests with different stakes than white students in NYS. In a talk at the *Harvard Inequality and Social Policy Seminar* on March 5, 2012, J. Jennings presented evidence
from NYS showing that among 8th graders, African-American and Hispanic students have larger average score gaps between the state tests and NAEP than white students in 2007 and 2009.

Recall that the non-audit subtest comprised items from past NYS state tests that were likely targets of inappropriate test preparation. In addition, all the non-audit items included in our study were matched to audit items that tested the same specific content but that removed one or more predictable aspects in their matched non-audit items that could facilitate coaching. Hence, if the score difference between the non-audit and audit subtests did provide a measure of score inflation, we expect that students with any of the above characteristics associated with greater exposure to inappropriate test preparation—minority, poor, near the proficient cut-score, in schools with high proportions of minority or poor students—to have larger score gaps between the subtests than students without these characteristics. The absence of such discrepant score gaps between the groups of students would thus cast doubts on the score difference between the two subtests as a measure of score inflation.

Therefore, we examined the relationships between students’ non-audit versus audit score gaps and: (1) race/ethnicity (non-white versus white); (2) low-income status; (3) proficiency level on the 2010 mathematics state test; (4) school’s proportion of non-white students; and (5) school’s proportion of students with low-income status. We estimated the hypothesized relationships by fitting the generic 2-level (students nested in schools) random-intercepts model:

$$D_{is} = \alpha_{00} + \alpha_{10} \text{NONWHITE}_{is} + \alpha_{20} \text{LOW}_I \text{INC}_{is} + \alpha_{30} \text{PROF10}_{is} + \alpha_{01} \text{SM}_I \text{NONWHITE}_{s} + \alpha_{02} \text{SM}_I \text{LOW}_I \text{INC}_{s} + (\epsilon_{is} + u_s)$$

for student $i$ in school $s$, where $\text{NONWHITE}$ and $\text{LOW}_I \text{INC}$ are dichotomously-coded variables coding for non-white students and students from low-income families respectively, and $\text{SM}_I \text{NONWHITE}$ and $\text{SM}_I \text{LOW}_I \text{INC}$ are the corresponding school-level proportions of non-
white students and students from low-income families respectively. \( PROF10 \) is a set of three binary variables coding for the top three of four proficiency levels on the 2010 mathematics state test, with proficiency levels II and III being the levels separated by the proficient cut-score, and the lowest level (Level I) being omitted and used as the reference level. All variables were grand-mean-centered. \( e_{is} \), the student-level residual for student \( i \) in school \( s \), was assumed to be independently and normally distributed with mean zero and variance \( \sigma^2_e \), for all \( i \), and \( s \). Finally, \( u_s \), the school-level residual for school \( s \), was assumed to be independent of \( e_{is} \) for all \( i \), and \( s \), and drawn from a normal distribution with mean zero and variance \( \sigma^2_u \).

The parameters of interest are \( \alpha_{10} \), \( \alpha_{20} \), \( \alpha_{30} \), \( \alpha_{01} \), and \( \alpha_{02} \):

- Parameters \( \alpha_{10} \) and \( \alpha_{20} \) represent the population difference in average non-audit versus audit score gap between (1) non-white and white students; and (2) students from low-income and non-low-income families, respectively, controlling for everything else. Based on past research, we expect both parameters to be positive, indicating that on average, in the population, everything else being equal, non-white students and students from low-income families have larger score gaps than white students and students from non-low-income families respectively.

- The components of parameter \( \alpha_{30} \) represent the population differences in average score gaps between students with proficiency Levels II, III, or IV on the 2010 mathematics state test and those with proficiency Level I, controlling for everything else. Since proficiency Level II is the level immediately below the “proficient” level, based on past research on the “bubble-kid syndrome”, we expect this component of \( \alpha_{30} \) to be positive, indicating that, on average, in the population, everything else being equal, students with proficiency Level II on the previous year’s
state test have larger score gaps than students with proficiency Level I. Past research also suggests that students with proficiency Level II would have received more inappropriate test-preparation activities than those in proficiency Levels III and IV. This could be inferred by comparing the respective components of \( \alpha_{30} \).

- Parameters \( \alpha_{01} \) and \( \alpha_{02} \) represent the population relationships between schools’ average score gap and their proportions of non-white students and students from low-income families respectively, controlling for everything else. Based on past research, we expect both parameters to be positive, indicating that, on average, in the population, everything else being equal, students in schools with higher proportions of non-white or low-income students have larger score gaps than those in schools with lower proportions of such students.

**Estimating the Impact on Schools’ Performance Ratings**

**Creating School-Performance Measures.** To investigate the impact of the subtest used on schools’ performance ratings (RQ2), we first generated a set of normative school-performance estimates. Such measures compare a school’s observed performance to its anticipated performance, and identify as high-performing those schools with higher observed performance than predicted. The different measures are distinguished by the specification of the statistical models used to predict the anticipated performance. The simplest model compares each school’s aggregate performance to the mean obtained over all schools. This transforms the status measure into one taking reference from the “average” school’s performance. More complex models predict the anticipated performance using a function of students’ prior achievement, other achievement-related characteristics (e.g., demographics), or both.

We specified statistical models that differed along two dimensions: (1) types of measures; and (2) types of background covariates. We used two types of measures: status and covariate-
adjustment. The latter included a measure of prior achievement as a covariate. We included three types of covariates defined by the levels of inclusion (if at all) of the control predictors. Crossing these two dimensions generated a set of 3 status measures (S1-S3) and 3 covariate-adjustment measures (CA1-CA3) (Table 1).

For each subtest, we generated the school-performance estimates by fitting the following generic 2-level random-intercepts multilevel model:

\[
ST_{score_{is}} = \mu + \alpha STATE_{2010_{is}} + \beta' B_{is} + \gamma' S_{s} + \varphi' BOOK_{is} + \psi_s + \epsilon_{is}
\]

for student \(i\) in school \(s\), where \(ST_{score} = A_{score}\) or \(N_{score}\) for measures based on the audit and non-audit subtests respectively.\(^3\) All control variables were grand-mean-centered. For each subtest, we assumed \(\epsilon_{is}\), the residual error term for student \(i\) in school \(s\), to be independent and normally distributed with mean zero and variance \(\sigma^2_{\epsilon}\), for all \(i\), and \(s\). \(\psi_s\), the empirical Bayes residual for school \(s\), is an estimate of the deviation of the empirical Bayes estimate of the school’s mean performance from its predicted performance using the model specified in equation (6) (Raudenbush & Bryk, 2002). It represents an estimate of the school’s performance based on the measure defined by the particular model specification. The use of these empirical Bayes estimates adjusts for both the student-level measurement errors due to the short, 6-item subtests, and the relative sampling errors for schools with small sample sizes. This is because both these errors determine the amount of shrinkage in the empirical Bayes estimate for each school: the larger the errors, the larger the shrinkage for the school. We also assumed that these school-performance estimates were drawn from a normal distribution with mean zero and variance \(\sigma^2_{\psi}\).

---

\(^3\) Following the argument by Thum and Bryk (1997), we fitted only multilevel models with fixed coefficients for all predictors because we were only interested in a school’s performance averaged over all its students.
Using these school-performance estimates, we estimated the impact of the subtest used on two common uses of school-performance measures: (1) to create rank-ordered lists of schools; and (2) to classify schools into broad performance bands.

**Impact on Schools’ Ranks.** We computed the Spearman’s \( \rho \) (rank correlations) between school-performance estimates obtained from the two subtests:

\[
(7) \ r^S (BT) = \text{Corr}\left(\text{Rank}(\hat{\psi}_s | \text{Non-audit}), \text{Rank}(\hat{\psi}_s | \text{Audit})\right)
\]

where \( \text{Rank}(\hat{\psi}_s | \text{Non-audit}) \) and \( \text{Rank}(\hat{\psi}_s | \text{Audit}) \) denote school ranks on the school-performance estimates derived from the non-audit and audit subtests respectively.

In addition, to assess the size of the estimated subtest effects, we compared them to the amounts of inconsistency induced in schools’ ratings by the different model specifications. This is because past research shows that schools’ ratings are only moderately correlated across status and covariate-adjustment measures, and also among covariate-adjustment measures with different covariates (e.g., Bosker & Witziers, 1995; Darmawan & Keeves, 2006; Keeves, Hungi, & Afrassa, 2005; OECD, 2008; Raudenbush, 2004; Tekwe, Carter, & Ma, 2004).

More specifically, we compared the estimated subtest effects to three groups of model-specification effects (i.e., between-model \( r^S \)) based on each subtest: (a) between-measure-type, within-covariate-type (Group A; i.e., measure effects); (b) within-measure-type, between-covariate-type (Group B; i.e., covariate effects); and (c) all other pair-wise combinations of measure- and covariate-types (Group C).

**Impact on Schools’ Assignment to Performance Bands.** The impact of subtest change on the assignment of schools to performance bands will depend on the classification scheme employed. In general, the proportion of schools changing classifications will be lower if there are fewer cut scores, if the cuts are in parts of the distribution with low density, or if the marginal
distributions are substantially non-uniform. Therefore, specific classification schemes served only as illustrations of possible impact. We examined three schemes, selected for their uses in past research or school-accountability systems. In Scheme 1, we classified schools with school-performance estimates that are at least one posterior SD (i.e., the SD of the distribution of empirical Bayes residuals obtained from equation [1]) below the average as “below average”; those with estimates that are at least one posterior SD above the average as “above average”; and all other schools as “average”. This scheme was used frequently in effective-schools research to identify “outlier” schools (Crone, Lang, & Teddlie, 1995). Other researchers have also used it to estimate the amount of inconsistency based on different school-performance estimates (e.g., Briggs & Weeks, 2009). Similarly, quantiles are often used in teacher/school value-added studies to illustrate the amount of classification inconsistency associated with a correlation between alternative value-added measures (e.g., Ballou, 2009; Corcoran et al., 2011; Papay, 2011). For Scheme 2, we used quintiles. Finally, unequal proportion, asymmetric classification schemes are sometimes used in practice. For Scheme 3, we adapted the system used in New York City’s Progress Report for schools (New York City [NYC] Department of Education, 2011a, 2011b). For the 2010-11 school year, NYC’s elementary and middle schools were assigned letter grades according to the following percentile ranks: “A”—top 25%; “B”—next 35%; “C”—next 30%; “D”—next 7%; “F”—bottom 3% (NYC Department of Education, 2011b).

We computed the observed percentage agreement in schools’ ranks between the two subtests for each scheme, separately for each model specification. We compared these percentages to the respective percentage of chance agreement (i.e., the agreement rate expected with random assignment of schools to the performance bands), which is a function of the number of cut scores and the marginal distributions.
Results

Functioning of the SMA as a Measure of Score Inflation

First-Order Differences. On average, students in the total field-test sample had poorer performance on the audit subtest than on the non-audit subtest. The estimated value of $\bar{D}$ was .28, suggesting that the corresponding average difference is positive in the population ($p < .001$). The audit items had thus functioned as designed, thereby meeting the necessary but insufficient requisite for using the audit data to detect score inflation.

Second-Order Differences. But to what extent was the positive value of $\bar{D}$ a result of coaching rather than non-inflation-related, incidental differences between the non-audit and audit items? In Table 2, we present the taxonomies of the fitted relationships between students’ non-audit versus audit score gaps and characteristics associated with differential exposure to inappropriate test preparation. In columns 1 through 5, we display the results separately for each of the characteristics, while in columns 6 through 10, we display the results when groups of characteristics were included in the model simultaneously.

The results are consistent with what past research would predict. Looking at the fitted bivariate relationships separately for each of the characteristics, on average, in the population:

1. A non-white student had larger estimated non-audit versus audit score gap than a white student ($p < .01$).
2. A student from a low-income family had a larger estimated score gap than one from a non-low-income family ($p < .001$).
3. A student with proficiency Level II on the previous year’s mathematics state test had larger estimated score gap than one with proficiency Level I on the same test ($p < .001$). In contrast, the estimated score gap of a student with proficiency Level III did not differ
from that of a student with proficiency Level I ($p > .05$), and the estimated score gap of a student with proficiency Level IV was smaller than that of a student with proficiency Level I ($p < .001$). Thus, by inference, on average, a student with proficiency Level II also had larger estimated score gap than one with Level-III or Level-IV proficiency level.

A student in a school with a higher proportion of non-white students had larger estimated score gap than one in a school with a lower proportion of white students ($p < .001$).

A student in a school with a higher proportion of students from low-income families had larger estimated score gap than one in a school with a lower proportion of such students ($p < .001$).

However, not all of these estimated bivariate relationships are independent sources of validity evidence for the score-inflation measure. Overall, students’ low-income status, at both the student and school levels, and their previous year’s proficiency levels appeared to be more important factors associated with the size of their score gaps, compared to race/ethnicity at either the student or school level. When both NONWHITE and LOW_INC were included in the model, only the coefficient on covariate LOW_INC remained the same and non-zero ($p < .001$), while that for NONWHITE became indistinguishable from zero (column 6 in Table 2). Similarly, the score gap was no longer related to schools’ proportions of non-white students once all the student-level characteristics were included in the model (column 8). In addition, controlling for everything else, the schools’ proportions of students from low-income families appeared to have more influence on score gaps than the students’ own low-income status since the coefficient on the latter became indistinguishable from zero when the former was included in the model (columns 9 & 10). Finally, students’ previous year’s proficiency levels appeared to be an independent source of association with score gaps, with no change in the associated hypothesis
tests and little change in the magnitudes of the associated parameter estimates when other characteristics were included in the model (compare column 3 to columns 7-10).

Nonetheless, the consistency of the results obtained across the different student characteristics we examined signals the presence of a systematic source. This is because incidental differences between the non-audit and audit items—a potential confounder of the inflation-related sources—are unlikely to persist over multiple comparisons.

These results, coupled with the design of the audit items—particularly that they were each closely matched to their respective non-audit items in the specific content tested, differing only in aspects that make them less vulnerable to typical coaching—strongly suggest that the score difference between the two subtests provides a measure of score inflation.

**Estimated Impact on Schools’ Performance Ratings**

**Impact on Schools’ Ranks.** Schools’ ranks varied considerably with the subtest used, for both status and covariate-adjustment measures, but more so for the latter than the former. The between-subtest correlations ranged from .53 to .68 for the three status measures, and from .42 to .48 for the three covariate-adjustment measures (Table 3).

In addition, these estimated subtest effects are at least as large as the estimated model-specification effects from all three groups, for both subtests. The maximum between-subtest correlation of .68 across all model specifications is comparable to the minimum between-model correlation across all groups and subtests of .71 (Table 4). This means that the inconsistency in schools’ ranks associated with a switch in subtest is at least as large as, and frequently larger than, that associated with a switch between either (1) a status and a covariate-adjustment measure with the same covariate type (i.e., measure effect; Group A); or (2) two measures of the same type but with a different set of control predictors (i.e., covariate effect; Group B).
Impact on Classification of Schools in Broad Performance Bands. Schools’ assignments to performance bands also depended substantially on the subtest used. In fact, in most cases, a switch in subtest led to an observed amount of inconsistency in assigned bands close to that arising from ignoring schools’ performance estimates on either subtests, and assigning them randomly to performance bands on both subtests. This is particularly so for covariate-adjustment measures, and applies to all three classification schemes.

For performance bands defined by cut scores at ±1 posterior standard deviation (Scheme 1), the average observed percentage agreement was 65%, across model specifications (Table 5). These observed percentages of agreement are fairly close to their corresponding chance agreement rates in general, with smaller gaps for covariate-adjustment measures than for status measures. The percentages of observed agreement above chance agreement rates ranged from 7% to 17% (average = 11%; Cohen’s kappa ranging from .12 to .36) across all model specifications, with the average percentage above chance agreement rates for the covariate-adjustment measures (8%) being smaller that for the status measures (13%).

Consistent with our expectations, for each model specification, the subtest effect was larger for either of the two 5-band schemes (Schemes 2 and 3) than that for the 3-band scheme (Scheme 1). While the minimum observed percentage agreement across model specifications for Scheme 1 was 62%, the corresponding maximum observed percentages of agreement were 37% and 50% for Schemes 2 and 3 respectively.

Between the two particular 5-band schemes that we used the subtest effect for the equal-proportion scheme (Scheme 2) was consistently larger than that for the unequal proportion and asymmetric scheme (Scheme 3) across all model specifications. For example, the observed
percentage agreement for the covariate-adjustment measure with all the covariates (i.e., model CA3) was 30% on Scheme 2 and 38% on Scheme 3.

**Sensitivity of Results to Ceiling Effects**

One consequence of not being able to create a common scale for all the items on the two subtests is that it precludes using scale scores to address the severe ceiling effects—i.e., right-censoring of scores at upper end of the score distribution—observed for the non-audit raw-score distribution (Figure 1).

The observed ceiling effects on the non-audit raw-score distributions could in turn potentially affect the results in two ways. First, insofar as the characteristics used for validating the score-inflation measures were correlated with performance, the ceiling effect could affect the validation results by understating the non-audit versus audit score gaps for students near the test-score ceiling on the non-audit subtests. Secondly, since past research suggests that schools’ value-added estimates are sensitive to ceiling effect in raw-score distributions (Koedel & Betts, 2008), the observed ceiling effect could also affect the estimated subtest effects directly.

To investigate these potential impacts of ceiling effect on our results, we conducted two sets of sensitivity analyses using a restricted sample that excluded students with perfect scores on the non-audit subtests: (1) we repeated the validation analyses; and (2) we re-estimated the subtest effects on schools’ ranks.

**Impact of Ceiling Effects on Validation Results.** In general, the earlier conclusions about the validity of the score difference between the subtests as a measure of score inflation are unaffected by ceiling effects. First, students in the restricted sample that excluded the perfect scorers still had poorer performance on the audit subtest than on the non-audit subtest, with an estimated $\bar{D}$ of $.26 (p < .001), compared to .28 obtained with the full analytic sample.
Similarly, results from the analyses based on second-order differences remained largely intact in the restricted sample. In Table 6, we display these results, parallel to those using the full analytic sample in Table 2. While the actual parameter estimates were somewhat different, the earlier conclusions about the bivariate relationships between score gaps and each of the characteristics in the population remained intact and consistent with what past research would predict. However, the strength of the estimated relationship for the low-income status at either the student or school level was weaker in the restricted sample than in the full analytic sample. No relationship remained between students’ score gaps and their own low-income status when the previous year’s proficiency level and the schools’ proportion of white students were included in the model ($p > .05$) (columns 7 & 8 in Table 6). Similarly, schools’ proportion of students from low-income families was also no longer related to score gaps when other characteristics (except schools’ proportion of white students) were included in the model ($p > .05$) (column 9).

**Impact of Ceiling Effects on Estimated Subtest Effects.** Results from the second set of sensitivity analyses suggest that the ceiling effect on the non-audit subtest could have attenuated the earlier estimated subtest effects. In Table 7, we display the estimated between-subtest correlations using the restricted sample, by model specification, paralleling those derived from the full analytic samples in Table 3. However, some schools were eliminated in the process of excluding students with perfect scores on the non-audit tests because they had fewer than three students who did not have perfect scores. Thus, for comparability, we re-estimated the subtest effects using the full analytic sample but only for the restricted set of schools that still had at least three students after excluding the perfect scorers. In Table 7, we display these estimates in the panels labeled “I. With Perfect Scorers”, and those using the restricted sample in the panels labeled “II. Without Perfect Scorers”.
For each model specification, the exclusion of the perfect scorers resulted in lower estimated between-subtest correlations (panels I and II in Table 7). For example, for the status measure without any covariate (i.e., model S1), the estimated between-subtest correlation in the full analytic sample was .66, compared to .47 in the restricted sample.

As the estimated subtest effects using the restricted set of schools but with perfect scorers included are very close to those obtained for the full set of schools in the main study (panels I in Table 7 and Table 3 respectively), these results suggest that the estimated subtest effects that we obtained in the main study are likely to be underestimates due to the severe ceiling effect on the non-audit subtest score distribution.

**Caveat.** In both sets of sensitivity analyses, the use of the restricted sample only provided evidence for whether ceiling effects had any impact on the results from the main study. It did not correct for the ceiling effects and provide unbiased estimates of either the subtest effect or the second-order differences underlying the validation analyses. For example, in an investigation of ways to address ceiling effects for latent growth-curve modeling, Wang, Zhang, McArdle, and Salthouse (2008) reported that parameter estimates obtained from using restricted samples remained biased. Further research using methods that address the ceiling effect is necessary to obtain unbiased estimates of the second-order differences and the subtest effects.

**Discussion**

We found strong supporting evidence that the SMA items included in the study provide a measure of score inflation, thereby demonstrating empirically the feasibility of designing an SMA to detect and potentially measure inflation due to coaching. We also found that the subtest used matters considerably for inferences about schools’ relative performance: schools’ ranks and their assignment to performance bands shifted substantially when one subtest was used rather
than the other. Across model specifications, the median between-subtest correlation was .56. Similarly, across model specifications and classification schemes, an estimated average of 54% of schools experienced a shift in their assigned performance bands with just a switch in subtest.

Together, these results suggest that schools’ ratings based on the non-audit subtest—and hence the past state tests from which the non-audit items were drawn—reflect their relative levels of engagement in coaching activities that inflate student scores.

However, several idiosyncrasies of the data-collection process that generated the pilot data led to important limitations in the study. First, the testing contractor overseeing the field test restricted all test booklets, including the audit-test booklets, to lengths of 12 items each. Each audit-test booklet was further divided into 6-item non-audit and audit subtests, with no overlapping item between the booklets. As we mentioned earlier, these resulted in (1) considerable measurement errors associated with item sampling in the students’ non-audit and audit subtest scores; and (2) the need to use raw scores—with the attendant issue of ceiling effect on the non-audit subtest score distribution—as we could not create a common scale for all the items. Further, each student was administered exactly one mathematics booklet, audit-test or otherwise. As we also mentioned earlier, because the audit-test booklets that we used for the study were only part of several field-test mathematics booklets of all types, some schools ended up having very few students with audit-test data. This in turn contributed to substantial student-sampling errors in these schools with small sample sizes.

Measurement errors due to either item or student sampling, and the observed ceiling effect on the non-audit subtest score distribution, potentially biased the estimated subtest effects in opposite directions, with ambiguous net effect. On one hand, our use of empirical Bayes residuals as estimates of schools’ aggregate performance would have adjusted for both the item-
sampling errors due to the short subtests, and the relative student-sampling errors. Nonetheless, it remains possible that all the estimated correlations were attenuated—and the corresponding subtest effects over-estimated—due to the small sample sizes for all schools. On the other hand, results from our analyses of the impact of ceiling effects suggest that the estimated subtest effects based on the full analytic sample are likely to be under-estimates.

In contrast, the validation results are robust against both the measurement errors due to item sampling, and the observed ceiling effect on the non-audit subtest distribution. We have presented evidence on the latter earlier. Measurement errors due to item sampling would result in larger standard errors for the parameter estimates from the regression analyses used for validation, thereby making it more difficult to detect statistically significant relationships between the score gaps and the various characteristics. Despite this, we still detected statistically significant relationships in directions consistent with what past research would predict.

In sum, our results demonstrated that an SMA might be a useful tool for identifying, and potentially for quantifying, score inflation. In addition, they demonstrated the potential bias in school-performance measures arising from score inflation on high-stakes tests. However, this study reflects only the first attempt to design an SMA. More research and development are thus needed to determine effective ways to design items that serve an audit function.
Bibliography


Center on Education Policy. (2010). *State test score trends through 2008-09, part 1: Rising scores on state tests and NAEP*. Center on Education Policy.


www.nypost.com/p/news/opinion/opedcolumnists/item_EF91F0Q4g70eT5y9CCXV1K;jsessionid=BD70062077631DCF7856B01A52AB0473


Table 1

Classification and model specification of school-performance measures, by types of measures and covariates

<table>
<thead>
<tr>
<th>Type of Covariates</th>
<th>Type of Measures</th>
<th>Status</th>
<th>Covariate-Adjustment</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. None</td>
<td></td>
<td>S1</td>
<td>CA1</td>
</tr>
<tr>
<td>2. Student-level background only</td>
<td></td>
<td>S2</td>
<td>CA2</td>
</tr>
<tr>
<td>3. Student- and school-level background</td>
<td></td>
<td>S3</td>
<td>CA3</td>
</tr>
</tbody>
</table>
Table 2

Taxonomy of fitted relationships between the non-audit versus audit score-gap and various characteristics indicative of differential incidence of inappropriate test preparation (n = 4,555 students; N = 1,050 schools)

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Fixed Effects</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NONWHITE ( (\hat{\alpha}_{10}) )</td>
<td>0.021**</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LOW_INC ( (\hat{\alpha}_{20}) )</td>
<td></td>
<td>0.042***</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PROF10_II ( (\hat{\alpha}_{30}) )</td>
<td></td>
<td></td>
<td>0.064***</td>
<td></td>
<td></td>
</tr>
<tr>
<td>PROF10_III ( (\hat{\alpha}_{30}) )</td>
<td></td>
<td></td>
<td></td>
<td>0.011</td>
<td></td>
</tr>
<tr>
<td>PROF10_IV ( (\hat{\alpha}_{30}) )</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>-0.083***</td>
</tr>
<tr>
<td>SM_NONWHITE ( (\hat{\alpha}_{01}) )</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.043***</td>
</tr>
<tr>
<td>SM_LOW_INC ( (\hat{\alpha}_{02}) )</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.068***</td>
</tr>
<tr>
<td>Intercept</td>
<td>0.278***</td>
<td>0.278***</td>
<td>0.278***</td>
<td>0.278***</td>
<td>0.278***</td>
</tr>
<tr>
<td><strong>Random Effects</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Between-school ( (\hat{\sigma}_u^2) )</td>
<td>0.003***</td>
<td>0.002***</td>
<td>0.002***</td>
<td>0.003***</td>
<td>0.002***</td>
</tr>
<tr>
<td>Within-school ( (\hat{\sigma}_e^2) )</td>
<td>0.056***</td>
<td>0.056***</td>
<td>0.054***</td>
<td>0.056***</td>
<td>0.056***</td>
</tr>
<tr>
<td><strong>Goodness-of-Fit</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>( \Delta(\text{Deviance}) )</td>
<td>7.550</td>
<td>30.646</td>
<td>223.894</td>
<td>16.332</td>
<td>36.756</td>
</tr>
</tbody>
</table>
## Fixed Effects

<table>
<thead>
<tr>
<th></th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>NONWHITE ($\hat{a}_{10}$)</td>
<td>−0.000</td>
<td>0.002</td>
<td>−0.011</td>
<td>−0.010</td>
<td>−0.008</td>
</tr>
<tr>
<td>LOW_INC ($\hat{a}_{20}$)</td>
<td>0.042***</td>
<td>0.023**</td>
<td>0.018*</td>
<td>0.007</td>
<td>0.006</td>
</tr>
<tr>
<td>PROF10_II ($\hat{a}_{30}$)</td>
<td>0.067***</td>
<td>0.067***</td>
<td>0.068***</td>
<td>0.068***</td>
<td></td>
</tr>
<tr>
<td>PROF10_III ($\hat{a}_{30}$)</td>
<td>0.018</td>
<td>0.018</td>
<td>0.019</td>
<td>0.019</td>
<td></td>
</tr>
<tr>
<td>PROF10_IV ($\hat{a}_{30}$)</td>
<td>−0.074***</td>
<td>−0.074***</td>
<td>−0.072***</td>
<td>−0.072***</td>
<td></td>
</tr>
<tr>
<td>SM_NONWHITE ($\hat{a}_{01}$)</td>
<td>0.027</td>
<td></td>
<td></td>
<td>−0.006</td>
<td></td>
</tr>
<tr>
<td>SM_LOW_INC ($\hat{a}_{02}$)</td>
<td></td>
<td>0.049**</td>
<td>0.052*</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Intercept</td>
<td>0.278***</td>
<td>0.278***</td>
<td>0.278***</td>
<td>0.278***</td>
<td>0.278***</td>
</tr>
</tbody>
</table>

## Random Effects

<p>| | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Between-school ($\hat{\sigma}_u^2$)</td>
<td>0.002***</td>
<td>0.002***</td>
<td>0.002***</td>
<td>0.001***</td>
<td>0.001***</td>
</tr>
<tr>
<td>Within-school ($\hat{\sigma}_e^2$)</td>
<td>0.056***</td>
<td>0.054***</td>
<td>0.054***</td>
<td>0.054***</td>
<td>0.054***</td>
</tr>
</tbody>
</table>

## Goodness-of-Fit

| $\Delta$(Deviance) | 30.646 | 234.030 | 237.060 | 243.318 | 243.404 |

* $p < .05$; ** $p < .01$; *** $p < .001$
### Table 3

*Between-subtest Spearman’s rank correlations, by model specification (N = 778 schools)*

<table>
<thead>
<tr>
<th>Type of Covariates</th>
<th>Type of Measures</th>
<th>Status</th>
<th>Covariate-adjustment</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. None</td>
<td></td>
<td>.68</td>
<td>.48</td>
</tr>
<tr>
<td>2. Student-level background only</td>
<td></td>
<td>.55</td>
<td>.42</td>
</tr>
<tr>
<td>3. Student- and school-level background</td>
<td></td>
<td>.53</td>
<td>.42</td>
</tr>
</tbody>
</table>
Table 4

Between-model Spearman’s rank correlations, by subtest, and group of model-specification comparisons (N = 778 schools)

<table>
<thead>
<tr>
<th>Subtest</th>
<th>Group</th>
<th>No. of Correlations</th>
<th>Min</th>
<th>Median</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>Audit</td>
<td>Group A</td>
<td>3</td>
<td>.87</td>
<td>.91</td>
<td>.92</td>
</tr>
<tr>
<td></td>
<td>Group B</td>
<td>S</td>
<td>.76</td>
<td>.86</td>
<td>.94</td>
</tr>
<tr>
<td></td>
<td></td>
<td>CA</td>
<td>.89</td>
<td>.94</td>
<td>.97</td>
</tr>
<tr>
<td>Audit</td>
<td>Group C</td>
<td>6</td>
<td>.71</td>
<td>.81</td>
<td>.88</td>
</tr>
<tr>
<td>Non-audit</td>
<td>Group A</td>
<td>3</td>
<td>.85</td>
<td>.90</td>
<td>.91</td>
</tr>
<tr>
<td></td>
<td>Group B</td>
<td>S</td>
<td>.78</td>
<td>.84</td>
<td>.95</td>
</tr>
<tr>
<td></td>
<td></td>
<td>CA</td>
<td>.92</td>
<td>.94</td>
<td>.98</td>
</tr>
<tr>
<td>Non-audit</td>
<td>Group C</td>
<td>6</td>
<td>.73</td>
<td>.80</td>
<td>.89</td>
</tr>
</tbody>
</table>

Note. Group A – between-measure-type, within-covariate type; Group B – between-covariate-type, within-measure-type; Group C – between all other pairs of measure- and covariate-types; S – status measures; CA – covariate-adjustment measures.
Table 5

*Between-subtest percentage agreement, percentage change agreement (in parentheses) and Cohen’s kappa (in italics), by type of measures and covariates, and classification scheme*

\( (N = 778 \text{ schools}) \)

<table>
<thead>
<tr>
<th>Classification Scheme</th>
<th>3</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Type of Measure</strong></td>
<td><strong>Type of Covariates</strong></td>
</tr>
<tr>
<td>Status</td>
<td>(± 1 posterior SD)</td>
</tr>
<tr>
<td>1</td>
<td>70 (53)</td>
</tr>
<tr>
<td></td>
<td>.36</td>
</tr>
<tr>
<td>2</td>
<td>64 (54)</td>
</tr>
<tr>
<td></td>
<td>.23</td>
</tr>
<tr>
<td>3</td>
<td>64 (53)</td>
</tr>
<tr>
<td></td>
<td>.23</td>
</tr>
<tr>
<td>Covariate adjustment</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>.18</td>
</tr>
<tr>
<td>2</td>
<td>65 (55)</td>
</tr>
<tr>
<td></td>
<td>.20</td>
</tr>
<tr>
<td>3</td>
<td>62 (55)</td>
</tr>
<tr>
<td></td>
<td>.17</td>
</tr>
</tbody>
</table>

*Note.* Type of covariates: 1. None; 2. Student-level background only; 3. Student- and school-level background
Table 6

Taxonomy of fitted relationships between the non-audit versus audit score-gap and various characteristics indicative of differential incidence of inappropriate test preparation, using restricted samples that excluded students with perfect scores on the non-audit subtest (n = 2,490 students; N = 961 schools)

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Fixed Effects</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NONWHITE ($\hat{a}_{10}$)</td>
<td>0.023*</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LOW_INC ($\hat{a}_{20}$)</td>
<td></td>
<td>0.038***</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PROF10_II ($\hat{a}_{30}$)</td>
<td></td>
<td></td>
<td>0.034*</td>
<td></td>
<td></td>
</tr>
<tr>
<td>PROF10_III ($\hat{a}_{30}$)</td>
<td></td>
<td></td>
<td>−0.026</td>
<td></td>
<td></td>
</tr>
<tr>
<td>PROF10_IV ($\hat{a}_{30}$)</td>
<td></td>
<td></td>
<td>−0.130***</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SM_NONWHITE ($\hat{a}_{01}$)</td>
<td></td>
<td></td>
<td></td>
<td>0.033*</td>
<td></td>
</tr>
<tr>
<td>SM_LOW_INC ($\hat{a}_{02}$)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.063***</td>
</tr>
<tr>
<td>Intercept</td>
<td>0.257***</td>
<td>0.257***</td>
<td>0.257***</td>
<td>0.255***</td>
<td>0.254***</td>
</tr>
<tr>
<td><strong>Random Effects</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Between-school ($\hat{\sigma}_u^2$)</td>
<td>0.002***</td>
<td>0.002***</td>
<td>0.002***</td>
<td>0.002***</td>
<td>0.006***</td>
</tr>
<tr>
<td>Within-school ($\hat{\sigma}_e^2$)</td>
<td>0.060***</td>
<td>0.059***</td>
<td>0.057***</td>
<td>0.060***</td>
<td>0.056***</td>
</tr>
<tr>
<td><strong>Goodness-of-Fit</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\Delta$(Deviance)</td>
<td>4.990</td>
<td>13.694</td>
<td>106.396</td>
<td>5.872</td>
<td>4.226</td>
</tr>
<tr>
<td></td>
<td>6</td>
<td>7</td>
<td>8</td>
<td>9</td>
<td>10</td>
</tr>
<tr>
<td>------------------</td>
<td>-----</td>
<td>-----</td>
<td>-----</td>
<td>-----</td>
<td>-----</td>
</tr>
<tr>
<td><strong>Fixed Effects</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NONWHITE ($\hat{\alpha}_{10}$)</td>
<td>0.005</td>
<td>0.006</td>
<td>0.003</td>
<td>-0.004</td>
<td>0.006</td>
</tr>
<tr>
<td>LOW_INC ($\hat{\alpha}_{20}$)</td>
<td>0.036**</td>
<td>0.015</td>
<td>0.014</td>
<td>0.003</td>
<td>0.002</td>
</tr>
<tr>
<td>PROF10_II ($\hat{\alpha}_{30}$)</td>
<td>0.036*</td>
<td>0.036*</td>
<td>0.037**</td>
<td>0.037**</td>
<td></td>
</tr>
<tr>
<td>PROF10_III ($\hat{\alpha}_{40}$)</td>
<td>-0.020</td>
<td>-0.020</td>
<td>-0.019</td>
<td>-0.018</td>
<td></td>
</tr>
<tr>
<td>PROF10_IV ($\hat{\alpha}_{50}$)</td>
<td>-0.122***</td>
<td>-0.122***</td>
<td>-0.121***</td>
<td>-0.120***</td>
<td></td>
</tr>
<tr>
<td>SM_NONWHITE ($\hat{\alpha}_{01}$)</td>
<td></td>
<td></td>
<td>0.006</td>
<td></td>
<td>-0.030</td>
</tr>
<tr>
<td>SM_LOW_INC ($\hat{\alpha}_{02}$)</td>
<td></td>
<td></td>
<td></td>
<td>0.038</td>
<td>0.058*</td>
</tr>
<tr>
<td>Intercept</td>
<td>0.257***</td>
<td>0.257***</td>
<td>0.257***</td>
<td>0.256***</td>
<td>0.256***</td>
</tr>
<tr>
<td><strong>Random Effects</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Between-school ($\hat{\sigma}_u^2$)</td>
<td>0.002***</td>
<td>0.002***</td>
<td>0.002***</td>
<td>0.002***</td>
<td>0.002***</td>
</tr>
<tr>
<td>Within-school ($\hat{\sigma}_e^2$)</td>
<td>0.059***</td>
<td>0.057***</td>
<td>0.057***</td>
<td>0.057***</td>
<td>0.057***</td>
</tr>
<tr>
<td><strong>Goodness-of-Fit</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\Delta$(Deviance)</td>
<td>13.908</td>
<td>109.566</td>
<td>109.634</td>
<td>112.488</td>
<td>113.668</td>
</tr>
</tbody>
</table>

* $p < .05$; ** $p < .01$; *** $p < .001$
Table 7

*Between-subtest Spearman’s rank correlations using restricted samples, with and without the perfect scorers, by model specification (N = 744 schools)*

<table>
<thead>
<tr>
<th>Type of Measures</th>
<th>Type of Covariates</th>
<th>Status</th>
<th>Covariate-adjustment</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>I. With Perfect Scorers</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>1. None</td>
<td>.66</td>
<td>.45</td>
</tr>
<tr>
<td></td>
<td>2. Student-level background only</td>
<td>.52</td>
<td>.40</td>
</tr>
<tr>
<td></td>
<td>3. Student- and school-level background</td>
<td>.50</td>
<td>.40</td>
</tr>
<tr>
<td></td>
<td>II. Without Perfect Scorers</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>1. None</td>
<td>.47</td>
<td>.33</td>
</tr>
<tr>
<td></td>
<td>2. Student-level background only</td>
<td>.36</td>
<td>.29</td>
</tr>
<tr>
<td></td>
<td>3. Student- and school-level background</td>
<td>.34</td>
<td>.28</td>
</tr>
</tbody>
</table>
Figure 1

*Sample distribution of proportion-correct scores, by subtest*

<table>
<thead>
<tr>
<th>Non-Audit Subtest</th>
<th>Audit Subtest</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image1.png" alt="Histogram for Non-Audit Subtest" /></td>
<td><img src="image2.png" alt="Histogram for Audit Subtest" /></td>
</tr>
</tbody>
</table>