Working Paper (25 Mar 2013)

Sensitivity of School-Performance Ratings to the Test Used

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

Hui Leng Ng[1]
Harvard Graduate School of Education
Phone: (+65) 9664-2439
Fax: -
Email: hui_leng_ng@mail.harvard.edu

Daniel Koretz
Harvard Graduate School of Education
415 Gutman Library
6 Appian Way
Cambridge, MA 02138
Phone: (617) 384-8090
Fax: (617) 496-3095
Email: daniel_koretz@gse.harvard.edu

Acknowledgement

We would like to thank Jennifer Jennings for her generosity in granting us access to the Houston Independent School District dataset that we used in the article.

[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

Standardized-test scores are increasingly important indicators of school success. But how robust are school-performance ratings when they are based on measures derived from such scores, especially under high-stakes conditions? Using data from Houston Independent School District, we investigated the sensitivity of school-performance ratings to the use of two different tests in the same academic subject. We found that the choice of test substantially affects both schools' ranks and their placement in performance bands. This applies to a variety of common school-performance measures, including both covariate-adjustment and gain-score measures with different covariates. We also found that test used matters more than the subject or model specification used, but less than the year used. Finally, we found evidence consistent with the view that the difference in stakes for schools between the tests contributed to the effects of choice of test. Particularly striking was the finding that the discrepancy in constructs of school performance as measured by the two tests in a particular subject is a school-level, rather than student-level, property. This is what one would expect if the difference in constructs arises in substantial part from responses to the high stakes attached to one of the tests.

Sensitivity of School-Performance Ratings to the Test Used

Introduction

School performance-rating lists are popular summaries of schools' educational success, both in the US and in other countries. This is because they appear easy to understand and interpret. Policymakers, researchers and other stakeholders in education use such lists for three common purposes: school accountability, school improvement, and information for parents' school-choice decisions (Organisation for Economic Cooperation and Development [OECD], 2008).

But, despite the centrality of standardized-test results in rating schools, there has been relatively little attention to the potential inconsistencies in schools' ratings associated with using different achievement tests in the same academic subject. A few recent studies have investigated the inconsistencies in *teachers'* value-added measures associated with tests that differ in content or stakes (Corcoran, Jennings, & Beveridge, 2011; Lockwood et al., 2007; Papay, 2011), but none has examined similar inconsistencies in *school*-performance measures.

The current prevalence of high-stakes testing in US education makes it important to extend the research on the sensitivity of school-performance ratings to the test used because high-stakes uses can induce score inflation, that is, increases in scores greater than true increases in achievement, and if variable, this inflation could bias school ratings. Although many studies in the US have documented discrepant performance gains based on scores on a high-stakes test (e.g., state test) versus those on a lower-stakes test (e.g., NAEP), these are at the state or district levels (e.g., Fuller, Gesicki, Kang, & Wright, 2006; Jacob, 2005, 2007), rather than at the school level. That is, if schools' ratings differ between high- and low-stakes tests, this suggests that schools' ratings on the high-stakes test reflect in part their relative degrees of engagement in inappropriate responses which distort the construct the test was designed to measure. Such responses include: (1) subjecting their students to inappropriate test-preparation activities; (2) irregularities in either test administration or scoring or both; and (3) other strategic responses that alter the composition of schools' tested-student populations, such as enforced grade retention and over-classification of lower-performing students for test exemption (e.g., Cullen & Rebeck, 2006; Figlio, 2002, 2006; Hamilton et al., 2007; Jacob, 2005; Jacob & Levitt, 2003; McMurrer, 2008; Nichols & Berliner, 2005). The first two types of responses bias school ratings by biasing the scores of individual students (Koretz & Hamilton, 2006). In contrast, the third type of response does not bias the scores of individuals but does bias aggregate measures of performance, including school ratings.

In this study, we investigated the impact of using two different standardized tests in the same subject, one high-stakes and one lower-stakes, on inferences about schools' relative performance. We used elementary-school reading and mathematics data from Houston Independent School District (HISD) on two standardized achievement tests, the Texas Assessment of Academic Skills (TAAS) and the Stanford Achievement Test (9th Edition) (SAT-9), in 2000 and 2001. Our analyses addressed three specific research questions:

RQ1. *How consistent are schools' performance ratings in a specific subject, grade, and year between the two different tests?*

RQ2. *How does the amount of inconsistency in schools' performance ratings associated with the use of different tests (i.e., test effect) compare to the amounts associated with other sources?*

RQ3. *Is there evidence that the estimated test effect on schools' ratings is associated with the differences in stakes for schools between the tests?*

We first define school performance and “value-added” measures of it, before we review the various mechanisms through which the test used could affect schools’ ratings in the same subject. Then we describe the context and school-accountability systems in HISD before we set out the research design for the study. We then address the three research questions sequentially. We conclude by discussing the implications of the results.

School Performance and “Value-Added” Measures of It

The shift from input-based accountability to outcome-based accountability in the past several decades fundamentally altered the definition of success for educational institutions. It is no longer adequate for educational institutions to demonstrate their provision of the minimal conditions for student learning to take place (Fuhrman, 2004). Instead, under outcome-based systems, educational institutions have to demonstrate that student learning has indeed taken place, and they are judged by the degree to which it did. In such systems, a school’s performance is determined by how well its students achieve the desired educational outcomes with which its stakeholders (e.g., parents; employers of future graduates; concerned voters) charge it.

Although designers of standardized tests have long warned that they measure only a portion of desired educational outcomes (e.g., Lindquist, 1951), standardized tests have become the ubiquitous metric for quantifying schools’ performance in many accountability systems internationally (Figlio & Loeb, 2010; Linn, 2004; Mathison, 2009; OECD, 2008; Torrance, 2009). Although test scores are used to derive a wide variety of school-performance measures, we used only normative measures in this study because our research questions are about schools’ relative performance. Such measures rate a school based on the difference between its observed performance—typically the mean—to its anticipated performance. The measures are distinguished by the specification of the statistical models used to predict the anticipated performance. The simplest approach compares each school’s aggregate performance to the mean obtained over all schools. More complex models predict the anticipated performance using a function of students’ prior achievement, other achievement-related characteristics (e.g., demographics), or both.

School-performance measures based on prediction models that control for students’ prior achievement are often called “value-added” measures. Policymakers are interested in these because of their purported ability to isolate educators’ contributions to student achievement more accurately than status measures (Doran, 2003; Mayston, 2006; McCaffrey, Lockwood, Koretz, & Hamilton, 2003; OECD, 2008). A school’s “value added” is intended to measure its contribution “to students’ progress towards stated or prescribed education objectives…net of other factors” (OECD, 2008, p.17). However, many researchers have cautioned against making causal inferences using these measures when students are not assigned randomly to schools since important unmeasured achievement-related differences among schools threaten such inferences (Braun, 2005; Fuhrman, 2010; Koretz, 2008; McCaffrey et al., 2003; McCaffrey & Lockwood, 2008; Reardon & Raudenbush, 2009; Rothstein, 2008, 2009; Rubin, Stuart, & Zanutto, 2004). Moreover, Castellano & Ho (2012) have pointed out that many of these models, including some examined here, would be more accurately called “conditional status” models, as they rate schools or teachers on the current performance of their students conditioned on past performance. For simplicity only, we will refer to the models as value-added.

Mechanisms Through Which Test Choice Could Affect Schools’ Ratings

Several factors could result in schools being rated differently on two different standardized tests that measure the same academic subject. Some of these factors are stochastic in nature.[1] For example, an unexpected event on the day on which one test, but not the other, was administered, which affected only students in some schools but not those in other schools. Others are systematic. Of these, some are related to the design of the tests themselves and thus exist even under no-stakes conditions. For example, the tests could cover different mathematics topics in different proportions, and as a result, one might be better aligned with the curriculum of a given school independent of any responses to testing. In contrast, others could be induced or accentuated by differences between the tests in the stakes involved for schools. For example, some schools might manipulate the composition of their tested-student populations for a high-stakes test used to determine their “effectiveness” by sidelining lower-performing students strategically, or by re-focusing instruction on the specifics of the test used for accountability. We limit our discussion to both types of systematic factors.

Further, only systematic factors that affect schools *non-uniformly* would contribute to inconsistent school ratings when one test was used rather than the other. For example, consider two mathematics tests that cover topics in different proportions. If all schools implement the same mathematics curriculum and thus have the same degree of test-curricular alignment with each test, then this difference in test design would not lead to inconsistent school ratings. Similarly, if all schools respond to high-stakes pressures by uniformly adopting practices that inflate the scores on the high-stakes test, then the difference in stakes would also not lead to inconsistent school ratings.

[1] Only stochastic events at the school level apply here since random measurement error associated with random events at the student level would be greatly reduced through aggregation of the scores to the school level.

Therefore, we further limit our discussion to systematic factors that affect schools non-uniformly. They include (1) the alignment between schools' implemented curricula and the content mixes of the tests; (2) the timing of the tests; (3) the students' motivational levels while taking the tests; (4) the test-administration procedures; and (5) the tested-student populations.

1. Alignment between Schools' Implemented Curricula & Tests' Content Mix

Schools could be rated differently on two tests due to differences in the degrees of alignment between their implemented curricula and the mixes of content on the two tests. This is because students' performance depends on whether, and the intensity to which, they had been exposed to the test content (D'Agostino, Welsh, & Corson, 2007; Porter, Smithson, & Blank, 2007).

Such variations in test-curricula alignment among schools might arise even in the absence of any responses to the fact of testing, but they also could arise because of responses to high stakes, even when the two tests share the same broader content domain. In particular, certain inappropriate test-preparation activities—namely, (1) shifting instructional emphasis *within* a particular academic subject from infrequently-tested materials to those tested frequently on the high-stakes tests (i.e., within-subject reallocation); and (2) focusing narrowly on the specific features of items that recur on the high-stakes tests (i.e., coaching) (Koretz & Hamilton, 2006)—would result in differences in test-curricula alignment for two tests with different stakes. Further, research shows that schools differ systematically in the incidence of such behaviors, with schools serving higher proportions of traditionally-underserved students tending to demonstrate more inappropriate test preparation than those serving lower proportions of such students (Firestone, Camilli, Yurecko, Monfils, & Mayrowetz, 2000; Herman & Golan, 1993; Lomax, West, Harmon, Viator, & Madaus, 1995; McNeil & Valenzuela, 2000; Monsaas &

Engelhard Jr., 1994; Shen, 2008). To the extent to which this test preparation is effective, it would contribute to inconsistencies in ratings between tests with different stakes.

2. Test Timing

Differences in the timing of test administration could affect schools' ratings if schools differ in their average rates of learning. For example, consider the extreme case where one test was administered at the beginning of fall and the other at the end of spring of the same school year. By virtue of any learning that has taken place during the school year, students' achievement on the two tests would be different. The longer the time between the administrations of two tests, the larger the expected impact on schools' ratings.

3. Students' Motivational Levels While Taking the Tests

Schools could be rated differently on two tests if students have differences in motivation that vary among schools because these differences in motivation will affect their test scores. For example, in a review of studies investigating the effects of student-level incentives (e.g., extra course credit, monetary reward) on students' test scores in the K-12 context, Wise and DeMars (2005) reported estimated effect sizes between the incentivized and un-incentivized group ranging from .07 to 1.49 (13 estimates; average = .54). Students' test-taking motivational levels are thus a source of "construct-irrelevant variance"—i.e., variation in students' scores due to factors that affect the scores but are extraneous to the construct being measured. Even student-level variation in motivation could affect school rankings if students with different gaps in motivation were distributed non-uniformly among the schools, although we found no relevant literature addressing this possibility.

Past research does document a variety of strategies that some schools used to motivate their students to do well on tests that hold high stakes for the schools, but that might not necessarily have any consequence for the students themselves. Such strategies include holding pep rallies near to the testing day, promising students days off or field trips, and rewarding good performance with prizes (Hollingsworth & Sockett, 1994; Pedulla et al., 2003). In contrast, although measurement experts are concerned about test-takers' motivational levels as a source of construct-irrelevant variance on tests that are low stakes to them (Barry, Horst, Finney, Brown, & Kopp, 2010; Haladyna & Downing, 2004; Wise & DeMars, 2005), there is no evidence that schools had done anything to influence their students' motivational levels systematically when the test has no consequence for them or their students. Therefore, any between-school variation in students' gaps in motivational levels on two tests is more likely to be induced by differential schools' responses to differences in school-level stakes.

4. Test-Administration Procedures

Test-administration procedures include aspects such as test security (e.g., access to test forms before test; handling of answer booklets after test), test duration, proctoring rules (e.g., proctors' responses to students' content-based questions during the test), and testing accommodations for special populations where applicable (e.g., large-font test forms for visually-impaired students).

When schools deviate from the standardized administration procedures differently for the two tests, and if that difference varies among schools, this may alter school ratings. A modest number of studies show varying adherence to standardized administration procedures among teachers in aspects such as enforcing the stated test duration, giving verbal instructions, and providing assistance to students under low-stakes conditions (Horne & Garty, 1981; White, Taylor, Carcelli, & Eldred, 1981). In contrast, there are many more documented instances of stakes-induced administration irregularities. These include unauthorized access to the test forms before testing day, providing students with answers before or during the test, and alteration of students' answers after the test (Amrein-Beardsley, Berliner, & Rideau, 2010; Jacob & Levitt, 2003; Lai & Waltman, 2008; Nichols & Berliner, 2005, 2007; Pedulla et al., 2003; Sorenson, 2006; Wilson, Bowers, & Hyde, 2011). Such behaviors inflate students' scores on the high-stakes test. However, we found no studies directly contrasting the incidence of violations of standardization under high- and low-stakes conditions.

5. Tested-Student Population

Schools could also be rated differently on two tests if the students who took one test differed in average achievement from those who took the other and this difference varied among schools.

Systematic differences in tested-student populations might arise because of policy. For example, in Texas, recent migrant-students are exempted from state tests due to their limited English proficiency (Texas Education Agency [TEA], 2006). If these students differ in achievement from tested students, and if the policy is not imposed on other tests, this policy could contribute to differences in ratings across tests.

High-stakes conditions could accentuate differences in tested-student populations if schools manipulate their tested-student populations on the high-stakes test to different degrees. For example, research shows that some schools alter the composition of their tested-student populations by sidelining lower-performing students through practices such as grade retention, strategically-timed disciplinary actions, and *exploitation* of exemption rules by over-classifying students with special-education (SPED) or limited-English-proficient (LEP) statuses (Cullen & Rebeck, 2006; Figlio, 2002, 2006; Haney, 2000; Jacob, 2005; McGill-Franzen & Allington, 2006; Nichols & Berliner, 2005). Such actions result in the performance estimates of these schools being higher than what they would be if all students were tested (Jennings & Beveridge, 2009).

Context and Test-based School-Accountability Systems in HISD

HISD is the largest public-school district in Texas and the seventh largest nationwide. In the school year ending in Spring 2012, it served a total of 203,066 students in 279 schools (HISD, 2012). The student population was 62% Hispanic, 25% African-American, 8% White, 3% Asian, and 1% students of other race/ethnicity. In addition, 80% of the students met the federal criteria for free and reduced-price lunches and were classified as economically disadvantaged, while 30% and 8% of the students were in programs tailored for LEP or SPED students, respectively.

Texas's high-stakes test-based accountability system predates the federal *No Child Left Behind* (NCLB) legislation by two decades, with its first public release of school-level performance in 1983 (Cruse & Twing, 2000). Besides the state-level school-accountability system, HISD also has had its own district-level system since 1993 (HISD, 2011). The federal-level system that came with the implementation of NCLB from 2003 thus constitutes the third system in the District but is not relevant to our findings because our data were from Spring 2000 and 2001, the last school years before the implementation of NCLB.

Use of TAAS and SAT-9 for School Accountability

The state test used in our study is the TAAS. It was administered state-wide from 1990 to 2002 and held high stakes for schools because both the state and district accountability systems used the results from these tests to assess schools, albeit in different combinations of subjects, measures, and other indicators (e.g., dropout/completion rates) (HISD Department of Research and Accountability, 2005; HISD, 2001; TEA, 2007a). For example, besides the reading, writing, and mathematics results used in both systems, the state system included results in social studies and science for selected grades. Nonetheless, based on their students' results on the state tests and the other indicators where applicable, high-performing schools were rewarded (e.g., monetary awards) while low-performing ones were sanctioned (e.g., placed on probation; re-constituted; closed) in each of the three accountability systems.

From 1996 to 2003, HISD also mandated that schools administer the nationally normed Stanford Achievement Tests, SAT-9. It first used the SAT results only to monitor district-level performance (i.e., no direct stakes for schools). From 2002, it also used these results for school accountability in its district-level system (HISD, 2001, nd.).

As such, in the two years included in our study (2000 and 2001), the TAAS and the SAT-9 differed distinctly in the levels of stakes for schools because the state and district accountability systems used only the TAAS results, but not the SAT-9 results.

Other Differences between TAAS and SAT-9

Besides differing in stakes for schools, in 2000 and 2001, the TAAS and the SAT-9 also differed in terms of content, testing date, score scales, and tested-student populations.

First, the TAAS and the SAT-9 have overlapping but non-identical content domains. In 2000 and 2001, the TAAS was aligned to Texas's curriculum, the Texas Essential Knowledge and Skills (TEA Austin Division of Student Assessment, 2000; TEA, 2008). In contrast, as the SAT-9 was targeted at the student population nationwide, it was aligned to content domains that were applicable more broadly across states. For example, the testing company described the SAT-9 reading and mathematics tests as being aligned to the NAEP framework and the *Curriculum and Evaluation Standards for School Mathematics* published by the National Council of Teachers of Mathematics respectively (Pearson Education Inc, 2012). Further, as there is evidence that states differed considerably in their content standards in reading (or English language arts) and mathematics (Porter, Polikoff, & Smithson, 2009), we expect the TAAS and SAT-9 to be aligned to somewhat different content domains in these two subjects.

One unpublished study provided an estimate of the degree of overlap between the TAAS and the SAT-9 in grade-4 mathematics. Hoey, Campbell, and Perlman (2001) estimated that 83% of the TAAS objectives were “either partially or totally matched” with the SAT-9 objectives while 74% of the SAT-9 objectives were “either partially or totally matched” with the TAAS objectives. These estimates also suggest that the SAT-9 has somewhat broader content domain in grade-4 mathematics than the TAAS.

Secondly, the TAAS and the SAT-9 were administered about one month apart in the spring of the respective school years (HISD, 2001, nd.).

Thirdly, although both the TAAS and the SAT-9 score-scales were created using the same scaling model (the Rasch model), the TAAS scores were not vertically linked across grades to allow an interpretation of a student’s difference in scores from one grade to the next as growth in achievement (TEA, 2011). In contrast, the SAT-9 scores were vertically scaled to allow such interpretations (Jorgensen, 2004; NCS Pearson Inc, 2004).

Finally, Texas allowed state-test exemptions for eligible SPED and LEP students (TEA, 2007b, 2012), but HISD mandated that “all student groups including ESL [English as a Second Language], most of Special Education, and other special populations” (HISD, 2007, p.1) take the SAT or the alternative Spanish-version, Aprenda, for comparability with the national sample.[2]

[2] LEP students in grades 3 to 5 are allowed to take the Spanish-language version of the state tests or the Aprenda.

Data

The HISD dataset that we used contained student-level TAAS and SAT-9 results in reading and mathematics of two cohorts of students who were in grade 5 in 2000 and 2001.

Measures

In Appendix A, we display the principal variables that we used.

Achievement in Target Year. Students' academic achievement in reading and mathematics in the target year (*Y*=2000, 2001) was measured by scale scores on the TAAS and the SAT-9, separately for each subject. We denote each of these generically by $SCORE(Y)$.

Achievement in Previous Year. Students' achievement in reading and mathematics in the year immediately prior to the target year was measured by scale scores on the two tests, separately for each subject. We denote each of these by $SCORE(Y-1)$.

Prior Achievement at Grade 3. Students' achievement in reading and mathematics when they were in grade 3 was also available, measured by both TAAS and SAT-9 scale scores. Students were required to demonstrate adequate reading proficiency on the third-grade TAAS test in order to be promoted to fourth grade (TEA, 2006). We denote these collectively by the vector ***PA***. Following the approach taken by Lockwood, McCaffrey, Hamilton, Stecher, Le, and Martinez (2007), ***PA***, unlike $SCORE(Y-1)$, included scores in *both* reading and mathematics.

As the scores for the different tests (TAAS versus SAT-9) and language versions (English versus Spanish) were on different scales, we standardized each scale with reference to the performance of a selected anchor cohort of students, separately for each combination of test, subject, grade, year, and language version. We used the 1998 grade-3 cohort, 1998 being the earliest year that we have access to in the dataset and grade 3 being the earliest grade whose data we used in the study. That is, we used the following means and SDs for standardization:

- For Grade 3: mean and SD of grade-3 scores of test administered in 1998
- For Grade 4: mean and SD of grade-4 scores of test administered in 1999
- For Grade 5: mean and SD of grade-5 scores of test administered in 2000

The resulting standardized scores are thus interpreted with reference to the performance of the 1998 grade-3 cohort as it progressed through the different grades. For example, a grade-5 student with a standardized scale score of 1unit on the TAAS reading test in 2001 was 1standard deviation (SD) above the average score of the 1998 grade-3 cohort when that cohort took the grade-5 TAAS reading test in 2000. This is akin to using a norming sample in the creation of a scale with normative interpretations (Kolen, 2006).

However, for gain-score measures (see later), we needed to preserve the vertical-scale properties of the SAT-9. Therefore, we could not standardize the SAT-9 scores separately for each grade. Instead, we standardized both grade-4 and grade-5 SAT-9 scores using the 2000 grade-5 results (i.e., the reference mean and SD used to standardize all grade-5 scores). This results in standardized scales at both grades which retain the vertical-scale properties of the SAT-9 scores, thereby preserving this particular difference between the TAAS and the SAT-9.

Student Background. Three sets of dichotomously coded covariates recorded selected student-background characteristics: (1) gender, family-economic status, and several race/ethnicity categories; (2) LEP, SPED, and disability statuses[3]; and (3) the language-medium of the state-tests/SATs the student took. We denote these covariates collectively by the vector ***B***.

[3] Students were classified with SPED status if they were eligible for special education services. In the dataset that we used, about 75% of the students with disability status had SPED status.

School-level Aggregate Variables. We also derived school-level measures by averaging the corresponding student-level variables for students within each school. We denote these collectively by the vector ***S***.

Construction of Analytic Samples

Our analytic samples were subsets of those for a larger study that involved two grades and two different pairs of state tests and SATs in four years.[4] The analytic samples for the larger study were constructed as follows.

First, we retained only student-level observations that were non-missing on at least one state-test or SAT score, and in all the student-background variables.

Next, as the test effect under investigation was in part associated with differences in tested-student populations between the TAAS and SAT-9, we preserved any such difference by creating separate analytic samples for the two tests. We did so by retaining only student-level observations that were non-missing on scores for the particular test in the particular year. Among students with non-missing data in all student-background variables, considerably more students did not have a TAAS score in at least one subject (9.9%) compared to those who did not have a SAT-9 score in at least one subject (0.3%).

In addition, consistent with the state-policy provisions for exemptions and alternative assessments applicable to students with LEP, SPED or disability statuses, far more students in these subgroups than others had at least one missing state-test score. 63.2% of the SPED students had at least one missing TAAS score compared to 3.6% of non-SPED students; the

[4] For simplicity, we have chosen to report our findings for grade 5 on the TAAS and SAT-9; results for grade 4 are largely similar and are available upon request.

corresponding percentages for the other two subgroups were 12.8% (LEP) versus 8.9% (non-LEP), and 59.4% (disabled) versus 2.9% (non-disabled). However, we were unable to discern whether (if at all) over-classification of students had occurred in any of the exemption categories.

Finally, we retained only schools that contained students at both grades who satisfied the student-level inclusion criteria for both subjects in all target years. This created a common set of schools for computing the school-performance estimates for all tests, subjects, grades, and years in the larger study. This is essential because normative school-performance measures depend on the particular schools that are included in the estimation sample.

The resulting analytic samples used in this article comprised 20,921 and 17,992 students for the SAT-9 and TAAS respectively, both in the same 164 schools. Both analytic samples were comparable to the total sample with regard to student-background characteristics (Appendix B). The only exceptions were with regard to SPED and disability statuses in the analytic sample for the TAAS. Consistent with the observations about missingness on the TAAS scores, the analytic sample for the TAAS comprised considerably fewer SPED (3%) and disabled (4%) students than the corresponding total sample (11% and 12% respectively).

In addition, across subjects and tests, students in either analytic sample were higher performing on average than those in the total sample. For SAT-9, the average performance of students in the analytic sample was between .02 and .08 SD higher than in the total sample; for TAAS, it was between .06 and .21 SD.

Methods

Creating School-Performance Measures

We defined a set of school-performance measures using two common types of “value-added” measures that differ in how the previous year’s achievement is incorporated into the estimation model: (a) as a covariate in the prediction of current achievement, so that the pretest-adjusted post-test score is used as a measure of "value" (i.e., covariate-adjustment); or (b) by simply subtracting the previous year's achievement from the target year's achievement and using the resulting difference as a measure of "value" (i.e., gain score). For each type of measure, we fitted six types of models defined by the control predictors included (Table 1), giving a total of 12 different school-performance measures comprising 6 covariate-adjustment measures (CA1-CA6) and 6 gain-score measures (GS1-GS6).

For each combination of test, subject, year, type of measures, and set of covariates, we generated the school-performance estimates by fitting this 2-level random-intercepts multilevel model:

$$(1)\ SCORE(Y)_{is} = \mu + \alpha SCORE(Y-1)_{is} + \beta' B_{is} + \pi' PA_{is} + \gamma' S_s + \psi_s + \varepsilon_{is}$$

for student *i* in school *s*, where $\alpha = -1$ for the gain-score measures. All variables are grand-mean-centered. For each combination of test, subject, year, type of measures, and set of covariates:

- ε_{is}, the residual error term for student *i* in school *s*, is assumed to be independent and normally distributed with mean zero and variance σ_{ε}^2, for all *i* and *s*.
- ψ_s, the estimate of the performance of school *s*, is the empirical Bayes residual, i.e., the shrunken deviation of the school's mean performance from its performance predicted by the model specified in equation (1) (Raudenbush & Bryk, 2002). We assumed these school-performance estimates to be independent of ε_{is} for all *i*, and *s*, and that they were drawn from a normal distribution with mean zero and variance σ_{ψ}^2.

Part I: Estimating the Impact on Schools' Performance Ratings

We estimated the impact of using different tests on two common school-performance measures: (1) rank-ordered lists of schools; and (2) classification of schools into broad performance bands.

Impact on Schools' Ranks. We used two indices to quantify test effect on schools' ranks. First, we computed the Spearman's *rho* (rank correlations) between school-performance estimates obtained from the two tests:

$$\textbf{(2)}\ r^{S}(\mathrm{BT}) = Corr\left(Rank(\hat{\psi}_s \,|_{\mathrm{TAAS}}), Rank(\hat{\psi}_s \,|_{\mathrm{SAT\text{-}9}}) \right)$$

where $Rank(\hat{\psi}_s \,|_{\mathrm{TAAS}})$ and $Rank(\hat{\psi}_s \,|_{\mathrm{SAT\text{-}9}})$ denote school ranks on the school-performance estimates derived from the TAAS and SAT-9 respectively. Secondly, we computed the mean absolute difference between the two ranked variables:

$$\textbf{(3)}\ MAD = \frac{1}{N}\sum_{s=1}^{N}\left| Rank\left(\hat{\psi}_s \,|_{\mathrm{TAAS}}\right) - Rank\left(\hat{\psi}_s \,|_{\mathrm{SAT\text{-}9}}\right) \right|$$

This index represents the average shift in school ranks in either direction when one test was replaced with the other.

Impact on Schools' Assignment to Performance Bands. The impact of changing tests 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). 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 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 average percentage agreement in schools’ ranks between the two tests for each scheme, separately for each combination of subject, year, and model specification. We compared these percentages to the percentage of chance agreement (that is, 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.

Part II: Evaluating the Size of the Test Effects

We compared the size of the test effects to the amounts of inconsistency associated with using different (1) subjects; (2) years of measurement; (3) types of measures; and (4) types of covariates. We used these factors as benchmarks because past research has shown that the generalizability of schools' performance tends to be moderate across different subjects, years, or model specifications (e.g., Darmawan & Keeves, 2006; Doolaard, 2002; Keeves, Hungi, & Afrassa, 2005; Ma, 2001; Tekwe, Carter, & Ma, 2004).

We decomposed the total variation in the set of school-performance estimates into components associated with each of the five factors—test, subject, year, type of measures, and types of covariates—and the interactions among them. In this analysis of variance, we treated the school-performance estimates as the outcome and the levels of each factor as fixed. Then we compared the total variation associated with the test factor and all interaction terms involving it with the corresponding total variation associated with each of the other factors.

Part III: Examining Difference in Stakes as an Explanation for Observed Test Effect

We conducted two sets of analyses to investigate the contribution high stakes to the test effects: (1) we examined the relationships between students' TAAS versus SAT-9 score gaps and selected characteristics associated with differential exposure to inappropriate responses to high stakes; and (2) we compared the constructs of school performance in each subject by which the TAAS and SAT-9 had ranked schools, using a multi-trait-multi-method (MTMM) framework, with the "trait" being the construct of school performance in this instance.

Students' TAAS versus SAT-9 Score Gaps and Selected Characteristics. High stakes for schools can induce inappropriate responses that could inflate students' scores on the high-stakes test. Research noted earlier found that minority students, poor students, and students in schools with high proportions of minority or poor students tended to be exposed to more such inappropriate responses in their schools than others. Insofar as these responses inflate scores on the higher-stakes test, we expect such students to show larger gaps between their scores on the two tests than other students.

Therefore, using restricted samples comprising students with scores on both the TAAS and SAT-9 in the same subject, we investigated the relationships between students' TAAS versus SAT-9 score gaps and (1) race/ethnicity; (2) economically-disadvantaged status; (3) school's proportion of non-white students; and (4) school's proportion of economically-disadvantaged students. We estimated the hypothesized relationships by fitting the following generic 3-level random-intercepts multilevel model, separately for reading and mathematics:

$$\textbf{(4)}\ \left(TAAS_{isy} - SAT9_{isy}\right) = \alpha_{000} + \alpha_{100} NONWHITE_{isy} + \alpha_{200} ECONDIS_{isy} \\ + \alpha_{010} SM\_NONWHITE_{sy} + \alpha_{020} SM\_ECONDIS_{sy} \\ + \alpha'_{300}(\mathbf{Controls}_{isy}) + \left(e_{isy} + u_{0sy} + v_{00y}\right)$$

for student *i* in school *s* in target year *y*. The set of control variables comprised students' TAAS results in the same subject in the previous year and its quadratic term, gender, LEP status, SPED status, and disability status. All variables were grand-mean-centered. For each subject, we assumed that e_{isy}, the student-level residual for student *i* in school *s* in year *y*, was normally distributed with mean zero and variance σ_e^2, for all *i*, *s*, and *y*. Similarly, we assumed that u_{0sy}, the school-level residual for school *s* in year *y*, was drawn from a normal distribution with mean zero and variance σ_u^2, for all *y*. Finally, we assumed that v_{00y}, the year-level residual for year *y*, was drawn from a normal distribution with mean zero and variance σ_v^2. We also assumed that e_{isy}, u_{0sy} and v_{00y} were mutually independent, for all *i*, *s*, and *y*.

The parameters of interest are α_{100}, α_{200}, α_{010}, and α_{020}:

- α_{100} and α_{200} represent the population difference in average score gap between non-white and white students, and between economically-disadvantaged and non-economically-disadvantaged students, respectively, controlling for everything else. Based on past research, we expect both parameters α_{100} and α_{200} to be positive, indicating that on average, everything else being equal, non-white and economically-disadvantaged students have larger score gaps than white and non-economically-disadvantaged students respectively.
- α_{010}, and α_{020} represent the relationship between schools' average score gap and their proportions of non-white and economically-disadvantaged students respectively, controlling for everything else, in the population. Based on past research, we also expect both parameter α_{010} and α_{020} to be positive, indicating that, on average, everything else being equal, students in schools with higher proportions of non-white or economically-disadvantaged students have larger score gaps than their peers in schools with lower proportions of such students.

MTMM Analyses. The MTMM approach evaluates validity by comparing measures of at least two distinct traits obtained using at least two different methods. If inferences about the traits are valid, intended measures of the same trait should produce more similar estimates than intended measures of different traits, and the method of measurement should be irrelevant. Therefore, the MTMM approach compares the correlations between different measures of the same trait to correlations between measures of different traits obtained using the same method. The term "trait effect" is often used to refer to the desired influence of the intended trait on scores—i.e., the correlation within trait across measurement methods. "Method effect" refers to the influence of measurement method on scores—i.e., the correlation within measurement method across traits. Valid inference requires minimally that the first set of correlations exceed the second (Campbell & Fiske, 1959), i.e., that the trait effect exceed the method effect.

In this study, the constructs measured were school performance in reading and mathematics, and the two methods of measurement were the TAAS and SAT-9. We compared the between-test (BT), within-subject Spearman’s *rho* (i.e., trait effect) to the within-test, between-subject (BS) Spearman’s *rho* (i.e., method effect). For each combination of subject, year, and model specification, the BT Spearman’s *rho* was given by $r^S(\mathrm{BT})$ defined in equation (2) earlier. In an analogous way, for each combination of test, year, and model specification, we computed the BS Spearman’s *rho* given by:

$$\textbf{(5)}\ r^S(\mathrm{BS}) = Corr\left(Rank(\hat{\psi}_s\,|_{\mathrm{Reading}}), Rank(\hat{\psi}_s\,|_{\mathrm{Math}})\right)$$

where $Rank(\hat{\psi}_s\,|_{\mathrm{Reading}})$ and $Rank(\hat{\psi}_s\,|_{\mathrm{Math}})$ denote variables created by ranking schools based on school-performance estimates derived from the reading and mathematics tests respectively.

A larger method effect on the TAAS than the trait effect for each subject (i.e., $r^S(\mathrm{BS})|_{\mathrm{TAAS}} > r^S(\mathrm{BT})|_U$ for *U* = reading or mathematics) indicates that schools’ rankings in the two subjects on the TAAS are more similar than their rankings in either subject on the two tests. This is evidence of the lack of construct validity for inferences about schools’ rankings in either subject using a particular test. It would also be consistent with the view that the high stakes for schools, a school-level characteristic that the two subjects on the TAAS shared, has contributed to the similarity of the constructs in the two subjects on the TAAS.

There are two alternative explanations of larger method than trait effects—i.e., a larger influence of the test than the subject on scores. The first is that the tests were intended to measure different constructs. This does not seem plausible, as it is difficult to posit intended inferences that would justify larger correlations between subjects within a test than within subjects between tests. However, it is entirely plausible that differences in the intended inferences would weaken the correlation within subjects between tests. A second alternative, which appears more plausible, is that one or both of the measures is faulty, independent of the effects of high stakes on test preparation.

Many of the responses to high stakes that can generate score inflation are implemented by teachers and administrators, which should generate classroom- or school-level effects. For example, Klein, Hamilton, McCaffrey, & Stecher (2000) found that the correlations between test scores and background variables were very different at the student and school levels for a high-stakes test but not a lower-stakes test, consistent with greater score inflation in disadvantaged schools. Koretz & Hamilton (2006) suggested that in the general case, such differences in correlational structure between levels of aggregation could be an indicator of score inflation. By the same token, classroom- and school-level score inflation might be expected to undermine MTMM results at the school level but to have less effect on similar analyses at the student level.

Therefore, we estimated the corresponding BT and BS Pearson correlations, representing the method and trait effects respectively, at the student level. To account for potential differences in the estimated student-level relationships between schools, we estimated the *within*-school student-level correlations by fitting the following set of four 2-level random-intercepts and random-slopes multilevel models, separately for each year:

$$\textbf{(6.1)}\ \text{BT(R)}:\quad R_{is}^{\text{TAAS}} = \gamma_{0,\text{R}} + \gamma_{1,\text{R}} R_{is}^{\text{SAT-9}} + \left(u_{s,\text{R}} + v_{s,\text{R}} R_{is}^{\text{SAT-9}} + e_{is}^{\text{R}}\right)$$

$$\textbf{(6.2)}\ \text{BT(M)}:\quad M_{is}^{\text{TAAS}} = \gamma_{0,\text{M}} + \gamma_{1,\text{M}} M_{is}^{\text{SAT-9}} + \left(u_{s,\text{M}} + v_{s,\text{M}} M_{is}^{\text{SAT-9}} + e_{is}^{\text{M}}\right)$$

$$\textbf{(6.3)}\ \text{BS(SAT-9)}:\ M_{is}^{\text{SAT-9}} = \gamma_{0,\text{SAT-9}} + \gamma_{1,\text{SAT-9}} R_{is}^{\text{SAT-9}} + \left(u_{s,\text{SAT-9}} + v_{s,\text{SAT-9}} R_{is}^{\text{SAT-9}} + e_{is}^{\text{SAT-9}}\right)$$

$$\textbf{(6.4)}\ \text{BS(TAAS)}:\ M_{is}^{\text{TAAS}} = \gamma_{0,\text{TAAS}} + \gamma_{1,\text{TAAS}} R_{is}^{\text{TAAS}} + \left(u_{s,\text{TAAS}} + v_{s,\text{TAAS}} R_{is}^{\text{TAAS}} + e_{is}^{\text{TAAS}}\right)$$

for student *i* in school *s*, where *R* and *M* represent the standardized (within each combination of test, subject, and year), school-mean-centered scale scores in reading and mathematics respectively, on the particular test indicated by the superscripts. In each equation, the e_{is} represents the student-level residual for student *i* in school *s*, which we assumed to be independently and normally distributed with mean zero and variance σ_e^2, for all *i*, and *s*. Similarly, in each equation, u_s and v_s represent school *s*'s deviations from the average intercept and average slope respectively. We assumed them to be independently drawn from normal distributions with means zero and variances σ_u^2 and σ_v^2 respectively. Finally, we assumed that the e_{is}'s, u_s's and v_s's from the different equations were mutually independent among themselves, both within and between equations.

If the estimated BT correlation in either subject (i.e., $\hat{\gamma}_{1,\text{R}}$ or $\hat{\gamma}_{1,\text{M}}$) were smaller than the estimated BS correlations (i.e., $\hat{\gamma}_{1,\text{SAT-9}}$ or $\hat{\gamma}_{1,\text{TAAS}}$), then there is evidence that the two tests were not measuring the same student-level trait in the particular subject.

Results

Part I: Impact of Test Choice on Schools' Ratings

Results show that inferences about schools' relative performance, based on lists of either schools' ranks or performance bands, depend substantially on the test used.

Impact on Schools' Ranks. Across all combinations of subject, year, and model specification, the use of TAAS versus SAT-9 led to substantial inconsistencies in schools' ranks, with larger test effects in reading than in mathematics in general. The BT Spearman's *rho* ranged from .27 to .63 across years and model specifications, with a median of .40 and .58 in reading and mathematics respectively (Table 2). These correlations corresponded to average shifts in ranks in either direction of between 45 and 31 rank positions in reading, and between 35 and 30 rank positions in mathematics, shifts that are considerable for a rank-ordered list of 164 schools. The substantial amounts of inconsistency are also evident from the scatter-plots of schools' ranks on the two tests for the cases with the minimum and maximum BT correlations (Figure 1).

Impact on Classification of Schools in Broad Performance Bands. The test used also has considerable impact when we classify schools in broad performance bands. In fact, for a majority of combinations of subject, year, and model specification, the observed amount of inconsistency in schools' assigned performance bands associated with a switch in test was close to that arising from ignoring schools' performance estimates entirely and randomly assigning them to the bands on both tests. This 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 72%, across subjects, years, and model specifications (Table 3). There is little difference in percentage agreement either between the two subjects or between the two types of measures. This agreement rate appears high, but it was little better than chance. The percentages of observed agreement above chance agreement rates ranged from a mere 5% to 18% (average = 12%; Cohen's *kappa* ranging from .04 to .48) across subjects, years, and model specifications.

Consistent with our expectations, for each combination of subject, year and model specification, the test 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 for Scheme 1 was 64%, the corresponding *maximum* observed percentages agreement were 41% and 59% for Schemes 2 and 3 respectively.

Between the two particular 5-band schemes that we used, for each subject, the average test effect for the equal-proportion scheme (Scheme 2) was consistently larger than that for the unequal proportion and asymmetric scheme (Scheme 3). This applies to both (a) across all model specifications; and (b) within each type of measures. For example, in reading, the average observed percentage agreement across all model specifications was 31% and 40% for Schemes 2 and 3 respectively. Similarly, the average observed percentage agreement for covariate-adjustment measures were 31% (Scheme 2) and 40% (Scheme 3), and for gain-score measures, 30% (Scheme 2) and 39% (Scheme 3).

Part II: Size of the Test Effect

We compared the estimated test effect to the effect associated with other factors (Table 4)

Comparison with Subject Effect. The test used in deriving the school-performance estimates accounted for more inconsistency in schools' ratings than the subject tested. 44% of the total variation in the set of school-performance estimates was associated with the test effect and any interaction term involving it (column labeled "*t* "). This is larger than the corresponding 31% associated with the subject effect of 31% (column labeled "*u*").

Substantively, this indicates that schools' relative performance are more dependent on the *method* (i.e., the test) used to measure student achievement, than they are dependent on the *trait* (i.e., subject) defining the achievement being measured. This implies a lack of construct validity for at least one of the tests in supporting inferences about schools' relative performance in a particular subject. We examine this contrast between the method and trait effects, and its implications, further using the MTMM framework later.

Comparison with Year Effect. In contrast, we found that schools' ratings were more dependent on the particular year than the test used. The year effect and any interaction term involving it accounted for 57% of the total variation (column labeled "*y*"), which is larger than the corresponding 44% associated with the test effect. The relatively smaller test effect compared to the year effect is unsurprising. This is because the year effect incorporated variation due to different cohorts of students, a source of variation that was absent in the test effect.

Comparison with Measure & Covariate Effects. Finally, we found that schools' ratings depended more on the particular test used than on either the type of model (gain score vs. covariate adjustment) or set of covariates used. These two effects and any interaction term involving them accounted for only 10% (7%) of the total variation (columns labeled "*m*" and "*c*"). These are both considerably smaller than the corresponding 44% associated with the test effect. Thus, the observed test effects are unlikely to be idiosyncratic to the specific model specifications that we have included in our study.

Part III: Differences in Stakes as an Explanation for Test Effects

In this final subsection, we discuss two pieces of evidence suggesting that the observed test effects were driven in part by the difference between the tests in stakes for schools.

Differential Score Gaps for Groups of Students. In all cases, the TAAS/SAT-9 score gaps showed relationships with student characteristics associated with differential exposures to inappropriate test preparation are consistent with what past research would predict. In Table 5, we display the taxonomies of the fitted relationships that we obtained, by subject. For each fitted relationship, we display the estimates of the four parameters of interest (i.e., α_{100}, α_{200}, α_{010}, and α_{020}), where applicable, in the first four rows.

Relationships with Student-Level Characteristics. On average, in the population, controlling for everything else in the respective models, non-white students and economically-disadvantaged students have larger TAAS/SAT-9 score gaps than their white or non-economically-disadvantaged counterparts. This applies to both subjects. For example, in reading, controlling for previous year's state-test achievement in reading and all other control variables, the score gap of a non-white student is .25 SD larger than that of a white student (p < .001) (column 2 in panel A). Similarly, on average, in the population, controlling for everything else (including the student's race/ethnicity), the corresponding average score gap difference between an economically-disadvantaged student and non-economically-disadvantage student is .14 SD (p < .001) (column 3 in panel A). Although the parameter estimates of these two student-level characteristics became smaller when the two school-level characteristics were included in the model, there was no change in their sign or significance (p < .001 in all cases) (columns 4 through 6).

Relationships with School-Level Characteristics. Everything else (including the students' own race/ethnicity and economic-disadvantage status) being equal, students in schools with higher proportions of non-white students or economically-disadvantaged students have larger TAAS/SAT-9 score gaps than their counterparts in schools with lower proportions of such students ($p < .001$) (columns 4 and 5). This applies to both subjects, and is consistent with what past research would predict. For example, in mathematics, on average, in the population, everything else (including the students' own race/ethnicity) being equal, the score gap of a student in a school with 1% more non-white students is .47 SD larger than that of a student in a school with 1% fewer non-white students ($p < .001$) (column 4 in panel B). Similarly, on average, in the population, everything else (including the students' own economically-disadvantaged status) being equal, the score gap of a student in a school with 1% more economically-disadvantaged students is .41 SD larger than that of a student in a school with 1% fewer such students ($p < .001$) (column 5 in panel B).

However, because these two school-level characteristics were highly negatively correlated (– .93), they do not constitute independent sources of evidence. Schools' proportions of economically-disadvantaged students appeared to exert a stronger impact for either subject. In reading, the proportion of non-white students was no longer related to average score gaps ($p >$ .05) when the proportion of economically-disadvantaged students was included in the model (column 6 in panel A). In mathematics, the relationship between average score gaps and schools' proportions of non-white students became marginally negative ($p < .05$) when schools' proportions of economically-disadvantaged students were included in the model (column 6 in panel B).

The school-level findings are particularly important, given that we hypothesize that inappropriate test preparation is to a considerable degree a school-level variable, that is, that there are large between-school differences in these behaviors.

Large test effects in the MTMM analysis. The MTMM analysis shows consistently larger method effects (correlations within tests between subjects) than trait effects (correlations within subjects between tests), which suggests that the TAAS and the SAT-9 in the same academic subject are not ranking schools based on the same construct. This applies to all combinations of year and model specification, and is consistent with the results from the variance decomposition discussed earlier. In panel A in Table 6, we display the estimated method and trait effects, by year, type of measures and covariates. The estimated trait effects for reading and mathematics are in columns labeled “BT(R)” and “BT(M)” respectively. The estimated method effects for the TAAS are in columns labeled “BS(TAAS)”. For example, in 2000, for the covariate-adjustment measure without any covariate (first row, columns labeled “1. None”), the BT correlations for reading and mathematics were .43 and .58 respectively. These are both smaller than the corresponding BS correlation for the TAAS of .69.

It is not likely that one could obtain these results simply from intentional differences between the tests in the absence of inflation, because even two very different mathematics tests should usually share more performance variance than either of them would share with a reading test. However, we can test this alternative explanation. If it were correct, we would expect the estimated method effects to be larger than the estimated trait effects at the *student* level, indicating a discrepancy in the constructs measured by the two tests in either subject, just as we had observed at the school level.

This was *not* what we observed. Unlike the school-level results, there is no indication that the estimated method effects were larger than the estimated trait effects at the student level. In fact, the trait effects were larger in both years. For example, in 2000, the estimated average within-school BT correlations for reading and mathematics were .70 and .73 respectively, which are larger than the corresponding estimated BS correlation of .61 for the TAAS (Table 7). In addition, for each year, the estimated within-school BT correlations differed among schools, with estimated standard deviations ranging from .17 to .28 (p < .05 in all cases; in parentheses in Table 7).

These results suggest that the observed discrepancy in constructs of school performance as measured by TAAS and SAT-9 in a particular subject is a school-level, rather than student-level, property. As noted earlier, this pattern is consistent with score inflation. Among the five mechanisms through which the use of two tests could lead to different schools' ratings that we discussed earlier, four—test-curricula alignment, students' test-taking motivational levels, test-administration procedures, and tested-student populations—could be induced by stakes. The fifth, test timing, is not likely to have a large impact in this case since the TAAS and SAT-9 were administered within just a month of each other.

Although we were unable to isolate any particular mechanism as a key contributor, we were able to ascertain that the observed discrepancies in the constructs measured by the TAAS and SAT-9 at the school level were not merely driven by differences in tested-student populations between the two tests. Using restricted samples comprising only students with both TAAS and SAT-9 scores, we re-estimated the school-level method and trait effects for all the measures. If the earlier results based on the full analytic samples—which we had constructed to preserve the differences in tested-student populations between the two tests—were driven solely by differences in the tested-student populations, then the results from the earlier MTMM analyses should disappear when we use the restricted samples.

This did not happen. Instead, all the earlier results from the MTMM analyses remained intact when the restricted samples were used: schools' ranks in the two subjects on the TAAS continued to be more similar than their ranks in the same subject on the two tests (panel B in Table 6). For example, in 2000, for the covariate-adjustment measure without any covariate, the estimated BT correlations for reading and mathematics were .48 and .54 respectively, which are still smaller than the corresponding estimated BS correlation for the TAAS of .68.

In sum, all these results suggest that the TAAS and SAT-9 in the same subject are not ranking schools based on the same construct of school performance. To the extent that the stakes-induced differences between the two tests were at play, these results further suggest that the schools' ranks in either subject, based on the TAAS results, reflect in part their relative levels of engagement, and effectiveness, in the stakes-induced behaviors that we had described earlier.

Discussion

We found that the test used to derive school-performance estimates matters substantially for inferences about schools' relative performance in a specific subject, grade, and year, and it matters more than the subject or model specification used but less than the year used. We could not fully isolate the contribution of the stakes-induced differences between the two tests from other differences—e.g., content differences and occasion of testing—using the existing data in our study. Nonetheless, two findings suggest that differences in stakes played some role in driving the observed test effects. The most important are results from the MTMM analyses showing that the two tests, which differed distinctly in stakes for schools, are not ranking schools based on the same construct of school performance in the same subject. This is even though it is reasonable to assume that the content domains of both tests were aligned to the same set of valued academic outcomes in HISD as the District mandated the administration of both tests. A standard approach for evaluating the validity of test-based inferences is a restricted form of MTMM commonly called ‘convergent/discriminant evidence’: scores on tests that purport to measure the same constructs (e.g., two mathematics tests) should correlate more highly than scores on tests designed to measure different constructs (e.g., a mathematics test and a reading test). The TAAS results reported here fail to meet that standard.

These results call for paying greater attention to the particular test adopted to derive school-performance estimates used to evaluate schools’ relative performance. This is because the inconsistencies in schools’ ratings threaten the validity of inferences about schools’ relative performance based on a particular test. This is particularly important under high-stakes testing conditions. When the inconsistencies are associated with differences in stakes between two tests, they imply that schools’ ratings on the high-stakes test in part reflect their relative degrees of engagement in inappropriate responses which distort the initial construct that the high-stakes test was designed to measure. Such distortions result in biased school-performance measures because of the disjuncture in the construct assumed by the target inferences about schools’ actual ratings and that actually measured by the high-stakes test.

To the extent that the differences in performance are attributable to educators’ behaviors, there are at least two consequences associated with the use of such biased measures. A direct consequence is that schools will be rewarded and sanctioned wrongly with regard to true student learning. An even more important consequence lies in possibly setting up a vicious cycle that facilitates the propagation of potentially questionable instructional practice throughout the system. This could happen via the three mechanisms of “institutional isomorphic change”

(DiMaggio & Powell, 1983). First, external authorities could seek to systematize the “best practices” deemed characteristic of high-performing schools (i.e., coercive isomorphism). Secondly, schools judged to be less successful could observe what their more successful counterparts did to get rewarded, and emulate their practices (i.e., mimetic isomorphism), some of which are questionable for achieving students’ long-term learning. Finally, when used as outcome measures for research purposes, such distorted school-performance measures could misinform the relationships between instructional practice and student learning. This could inadvertently lend unwarranted credence to potentially questionable “best practices”, thereby cementing the latter’s place in the norms of the teaching profession (i.e., normative isomorphism).

It is therefore critical that policymakers and researchers take the necessary steps to prevent, detect, and correct such potential biases in the measures, especially under high-stakes conditions. However, such biases in schools’ ratings are not easily addressed. For example, they cannot be addressed by averaging over multiple years’ of results from the same high-stakes test if the biases were present in all these test scores.

We sketch some possibilities related to the design of the testing and accountability systems, and the continual monitoring of the functioning of these systems.

First, a direct approach to preventing the distortions is to alter the incentives for educators to engage in distorting actions, or minimally, make it less easy for them to be successful in doing so. This includes first and foremost, building tests that contain fewer predictable recurrences of content or item features that encourage within-subject reallocation or facilitate coaching (Holcombe, Jennings, & Koretz, 2013; Koretz & Beguin, 2010). Another way to alter the incentives is to shift the focus away from standardized tests in the handful of academic subjects by including alternative, non-test-based measures of these academic outcomes, as well as measures of other academic and non-academic educational outcomes, in the accountability system. This entails developing such alternative measures *and* using them to design effective incentive systems that would support improvement in actual student learning, both being areas that existing research does not support adequately (Hout & Elliot, 2011; Ladwig, 2010; National Research Council, 2011).

Secondly, a bias-detection system involves, minimally, the continual validation of each use of the test scores. While this applies to any score-based inference, it is critical under high-stakes conditions. This is because the potential distortions brought about by educators' inappropriate responses to high-stakes pressures (Koretz, 2002, 2010), make any validity evidence obtained during the initial test-development phase, prior to the actual and continued high-stakes use of the test results, inadequate (Koretz & Hamilton, 2006; Koretz, McCaffrey, & Hamilton, 2001). Therefore, under high-stakes conditions, the validation of each use of the test scores must continue beyond the initial test-development phase. When used for inferences about educators' or institutions' performance, this includes the validation of such uses of aggregate scores separately from uses of student-level scores (Forer & Zumbo, 2011; Linn, 2006).

References

Amrein-Beardsley, A., Berliner, D. C., & Rideau, S. (2010). Cheating in the first, second, and third degree: Educators' responses to high-stakes testing. *Education Policy Analysis Archives, 18*(14), March 13, 2012.

Ballou, D. (2009). Test scaling and value-added measurement. *Education Finance and Policy, 4*(4), 351-383.

Barry, C. L., Horst, S. J., Finney, S. J., Brown, A. R., & Kopp, J. P. (2010). Do examinees have similar test-taking effort? A high-stakes question for low-stakes testing. *International Journal of Testing, 10*(4), 342-363.

Braun, H. I. (2005). *Using student progress to evaluate teachers: A primer on value-added models*. Princeton, NJ: Educational Testing Service.

Briggs, D. C., & Weeks, J. P. (2009). The sensitivity of value-added modeling to the creation of a vertical score scale. *Education Finance and Policy, 4*(4), 384-414.

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin, 56*, 81-105.

Castellano, K. E., & Ho, A. D. (2012). *Simple Choices among Aggregate-Level Conditional Status Metrics: From Median Student Growth Percentiles to Value-Added Models*. Manuscript submitted for publication

Corcoran, S. P., Jennings, J. L., & Beveridge, A. A. (2011, March). *Teacher effectiveness on high- and low-stakes tests.* Paper presented at the Society for Research on Educational Effectiveness Spring 2011 Conference, Washington, D.C.

Crone, L. J., Lang, M. H., & Teddlie, C. (1995). Achievement measures of school effectiveness: Comparison of model stability across years. *Applied Measurement in Education, 8*(4), 353-365.

Cruse, K. L., & Twing, J. S. (2000). The history of statewide achievement testing in Texas. *Applied Measurement in Education, 13*(4), 327-331.

Cullen, J. B., & Rebeck, R. (2006). *Tinkering towards accolades: School gaming under a performance accountability system* (NBER working paper 12286). Cambridge, MA: National Bureau of Economic Research.

D'Agostino, J. V., Welsh, M. E., & Corson, N. M. (2007). Instructional sensitivity of a state's standards-based assessment. *Educational Assessment, 12*(1), 1-22.

Darmawan, I. G., & Keeves, J. P. (2006). Accountability of teachers and schools: A value-added approach. *International Education Journal, 7*(2), 174-188.

DiMaggio, P. J., & Powell, W. W. (1983). The iron cage revisited: Institutional isomorphism and collective rationality in organizational fields. *American Sociological Review, 48*(2), 147-160.

Doolaard, S. (2002). Stability and change in results of schooling. *British Educational Research Journal, 28*(6), 773-787.

Doran, H. C. (2003). *Value-added analysis: A review of related issues*. Annual Meeting of the American Educational Research Association (April 21-25), Chicago, IL.

Figlio, D. (2002). *Accountability, ability and disability: Gaming the system*? Unpublished manuscript.

Figlio, D. (2006). Testing, crime and punishment. *Journal of Public Economics, 90*(4-5), 837-851.

Figlio, D., & Loeb, S. (2010). Chapter 8 - school accountability. In E. A. Hanushek, S. Machin & L. Woessmann (Eds.), *Handbook of the economics of education* (pp. 383-421) Elsevier.

Forer, B., & Zumbo, B. (2011). Validation of multilevel constructs: Validation methods and empirical findings for the EDI. *Social Indicators Research, 103*(2), 231-265.

Firestone, W. A., Camilli, G., Yurecko, M., Monfils, L., & Mayrowetz, D. (2000). State standards & opportunity to learn in New Jersey. *Education Policy Analysis Archives, 8*(35), June 9, 2011.

Fuhrman, S. H. (2004). Introduction. In S. Fuhrman, & R. F. Elmore (Eds.), *Redesigning accountability systems for educatio*n (pp. 3-14). New York: Teachers College Press.

Fuhrman, S. H. (2010). Tying teacher evaluation to student achievement. *Education Week, 29*(28), 32-33.

Fuller, B., Gesicki, K., Kang, E., & Wright, J. (2006). *Is the No Child Left Behind Act working? The reliability of how states track achievement.* Working paper 06-1. Policy Analysis for California Education, PACE.

Haladyna, T. M., & Downing, S. M. (2004). Construct-irrelevant variance in high-stakes testing. *Educational Measurement: Issues and Practice, 23*(1), 17-27.

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

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

HISD Department of Research and Accountability. (2005). *Research brief: 2005 Houston Independent School District performance indicator accountability system*. Houston Independent School District.

Hoey, L., Campbell, P. B., & Perlman, L. (2001). *Where's the overlap? mapping the SAT9 and TAAS 4th grade test objectives.* Unpublished manuscript.

Holcombe, R., Jennings, J., & Koretz, D. (2013). The roots of score inflation: An examination of opportunities in two states' tests. In G. Sunderman (Ed.), *Charting reform, achieving equity in a diverse nation*. Greenwich, CT: Information Age Publishing.

Hollingsworth, S., & Sockett, H. (Eds.). (1994). *Teacher research and educational refor*m. Chicago: NSSE; Distributed by the University of Chicago Press.

Horne, L. V., & Garty, M. K. (1981, April). *What the test score really reflects: Observations of teacher behavior during standardized achievement test administration*. Paper presented at the annual meeting of the American Educational Research Association, Los Angeles, CA.

Houston Independent School District (nd.). *Introduction to Stanford/Aprenda testing*. Retrieved March 3, 2012, from http://www.houstonisd.org/portal/site/ResearchAccountability/?vgnextoid=d6f563055e644110VgnVCM10000028147fa6RCRD&vgnextfmt=alt8&epi_menuItemID=2de789f88f80e0c254177063e041f76a

Houston Independent School District. (2001). *Research brief: 2001 HISD accountability system*. Retrieved July 7, 2011, from http://www.houstonisd.org/ResearchAccountability/Home/Perform_Acount/AccountRpts/HISD_Briefs_Ratings/HISD_Briefs/2001HISDAccBrief.pdf

Houston Independent School District. (2007). *Stanford Achievement Test series, Tenth Edition results for grade 1-11 and Aprenda: La Prueba de Logros en Espanol, Tercera Edicion (Aprenda 3) results for grades 1-8, spring 2007*. Retrieved June 15, 2011, from http://www.houstonisd.org/ResearchAccountability/Home/Perform_Acount/StudPerf/Stanford%20Aprenda/Performance/PerfRpt%202007/Introduction_2007.pdf

Houston Independent School District. (2012). *2011-2012 facts and figures*. Retrieved January 19, 2013, from http://www.houstonisd.org/domain/7908

Houston Independent School District. (2011). *Three types of accountability reports*. Retrieved July 24, 2011, from http://www.houstonisd.org/portal/site/ResearchAccountability/menuitem.78ff2bf53f95920d20d1287fe041f76a/?vgnextoid=b78a1d3c1f9ef010VgnVCM10000028147fa6RCRD&vgnextfmt=default

Hout, M., & Elliot, S. W. (Eds.). (2011). *Incentives and test-based accountability in education*. Washington, D.C.: The National Academies Press.

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

Jacob, B. A. (2007). *Test-based accountability and student achievement: An investigation of differential performance on NAEP and state assessments.* NBER working paper no. 12817. National Bureau of Economic Research.

Jacob, B. A., & Levitt, S. D. (2003). Rotten apples: An investigation of the prevalence and predictors of teacher cheating. *The Quarterly Journal of Economics*, *118*(3), 843-877.

Jennings, J. L., & Beveridge, A. A. (2009). How does test exemption affect schools' and students' academic performance*? Educational Evaluation and Policy Analysis*, *31*(2), 153-175.

Jorgensen, M. A. (2004). *Assessment report: The value of the Stanford scale as a common metri*c. San Antonio, TX: Pearson Education.

Keeves, J. P., Hungi, N., & Afrassa, T. (2005). Measuring value added effects across schools: Should schools be compared in performance? *Studies in Educational Evaluation, 31*(2), 247-266.

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; http://www.rand.org/content/dam/rand/pubs/issue_papers/2006/IP202.pdf).

Kolen, M. J. (2006). Scaling and norming. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 155-186). Westport, CT: Praeger Publishers.

Koretz, D. M. (2002). Limitations in the use of achievement tests as measures of educators' productivity. *Journal of Human Resources*, *37*(4), 752-77.

Koretz, D. M. (2008). A measured approach: Maximizing the promise, and minimizing the pitfalls, of value-added models. *American Educator, Fall*, 18-27, 39.

Koretz, D. M. (2010). Implications of current policy for educational measurement. *Next Generation K–12 Assessment Systems: Exploratory Seminar Publications (December 6-8, 2009),* Princeton, NJ.

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

Koretz, D. M., McCaffrey, D. F., & Hamilton, L. S. (2001). *Toward a framework for validating gains under high-stakes conditions* (CSE technical report No. 551). Los Angeles, CA: National Center for Research on Evaluation, Standards, and Student Testing.

Ladwig, J. G. (2010). Beyond academic outcomes. *Review of Research in Education, 34*(1), 113-141.

Lai, E. R., & Waltman, K. (2008). Test preparation: Examining teacher perceptions and practices. *Educational Measurement: Issues & Practice*, *27*(2), 28-45.

Lindquist, E. F. (1951). Preliminary considerations in objective test construction. In E. F. Lindquist (Ed.), *Educational Measurement*, 119-158. Washington, D. C.: American Council on Education.

Linn, R. L. (2004). Accountability models. In S. Fuhrman, & R. F. Elmore (Eds.), *Redesigning accountability systems for education* (pp. 73-95). New York: Teachers College Press.

Linn, R. L. (2006). Following the standards: Is it time for another revision? *Educational Measurement: Issues and Practice, 25*(3), 54-56.

Lockwood, J. R., McCaffrey, D. F., Hamilton, L. S., Stecher, B., Le, V., & Martinez, J. F. (2007). The sensitivity of value-added teacher effect estimates to different mathematics achievement measures. *Journal of Educational Measurement, 44*(1), 47-67.

Lomax, R. G., West, M. M., Harmon, M. C., Viator, K. A., & Madaus, G. F. (1995). The impact of mandated standardized testing on minority students. *The Journal of Negro Education, 64*(2), 171-185.

Ma, X. (2001). Stability of school academic performance across subject areas. *Journal of Educational Measurement, 38*(1), pp. 1-18.

Marsh, C. J. (2009). *Key concepts for understanding curriculum* (4th ed.). London ;New York: Routledge.

Mathison, S. (2009). Serving the public interest through educational evaluation: Salvaging democracy by rejecting neoliberalism. In K. E. Ryan, & J. B. Cousins (Eds.), *The SAGE international handbook of educational evaluation* (pp. 525-537). Los Angeles: SAGE.

Mayston, D. (2006). *Educational value added and programme evaluation: DfES research report* (No. 847). London: Department for Education and Skills.

McCaffrey, D. F., Lockwood, J. R., Koretz, D. M., & Hamilton, L. S. (2003). *Evaluating value-added models for teacher accountability*. Santa Monica, CA: RAND Corporation.

McCaffrey, D. F., & Lockwood, J. R. (2008, November). *Value-added model: Analytic issues.* Paper prepared for a workshop held by the committee on value-added methodology for institutional improvement, program evaluation and educational accountability sponsored by the national research council and the national academy of education. Washington DC. Unpublished manuscript.

McGill-Franzen, A., & Allington, R. (2006). Contamination of current accountability systems. *Phi Delta Kappan, 87*(10), 762-766.

McMurrer, J. (2008). *Instructional time in elementary schools: A closer look at changes for specific subjects*. Washington, DC: Center on Education Policy.

McNeil, L., & Valenzuela, A. (2000). *The harmful impact of the TAAS system of testing in Texas: Beneath the accountability rhetoric* Reports-Research (ED443872).

Monsaas, J. A., & Engelhard Jr., G. (1994). Teachers' attitudes toward testing practices. *Journal of Psychology, 128*(4), 469.

National Research Council. (2011). *Assessing 21st century skills: Summary of a workshop.* Committee on the Assessment of 21st century skills. Board on Testing and Assessment, Division of Behavioral and Social Sciences and Education. Washington, DC: The National Academies Press.

NCS Pearson Inc. (2004). *Stanford achievement test series tenth edition: Technical data report*. USA: Pearson Education Inc.

New York City Department of Education. (2011a). *Educator guide: The New York City progress report Elementary/Middle/K-8 (2009-10)*. Retrieved March 9, 2012, from http://schools.nyc.gov/NR/rdonlyres/4015AD0E-85EE-4FDE-B244-129284A7C36C/0/EducatorGuide_EMS_2011_03_10.pdf

New York City Department of Education. (2011b). *Educator guide: The New York City progress report Elementary/Middle/K-8 (2010-11)*. Retrieved March 9, 2012, from http://schools.nyc.gov/NR/rdonlyres/A82481C5-A351-47BA-BF8C-9F353E9CFB22/0/EducatorGuide_EMS_2011_10_03.pdf

Nichols, S. L., & Berliner, D. C. (2005). *The inevitable corruption of indicators and educators through high-stakes testing*. Arizona State University, Education Policy Research Unit.

Nichols, S. L., & Berliner, D. C. (2007). *Collateral damage : How high-stakes testing corrupts America's school*s. Cambridge, MA: Harvard Education Press.

Organisation for Economic Cooperation and Development. (2008). *Measuring improvements in learning outcomes: Best practices to assess the value-added of school*s. Paris: OECD.

Papay, J. P. (2011). Different tests, different answers: The stability of teacher value-added estimates across outcome measures. *American Educational Research Journal, 48*(1), 163-193. doi:10.3102/0002831210362589

Pearson Education Inc. (2012). *Stanford achievement test series, ninth edition - complete battery*. Retrieved February 17, 2012, from http://www.pearsonassessments.com/HAIWEB/Cultures/en-us/Productdetail.htm?Pid=E132C

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*. National Board on Educational Testing and Public Policy. Retrieved from http://www.bc.edu/research/nbetpp/statements/nbr2.pdf

Porter, A. C., Polikoff, M. S., & Smithson, J. (2009). Is there a de facto national intended curriculum? Evidence from state content standards. *Educational Evaluation and Policy Analysis, 31*(3), 238-268.

Porter, A. C., Smithson, J., & Blank, R. (2007). Alignment as a teacher variable. *Applied Measurement in Education, 20*(1), 27-51.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (2nd ed.). Thousand Oaks, Calif.: Sage Publications.

Reardon, S. F., & Raudenbush, S. W. (2009). Assumptions of value-added models for estimating school effects. *Education Finance and Policy, 4*(4), 492-519.

Rothstein, J. (2008). *Teacher quality in educational production. tracking, decay, and student achievement.* (NBER Working Paper No. 14442). National Bureau of Economic Research.

Rothstein, J. (2009). Student sorting and bias in value-added estimation: Selection on observables and unobservables. *Education Finance and Policy, 4*(4), 537-571.

Rubin, D. B., Stuart, E. A., & Zanutto, E. L. (2004). A potential outcomes view of value-added assessment in education. *Journal of Educational and Behavioral Statistics, 29*(1), 103-116.

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

Sorenson, D. (2006). *2006 state education test security result*s. Midvale, Utah: Caveon Test Security.

Taylor, C., & White, K. R. (1982). The effect of reinforcement and training on group standardized test behavior. *Journal of Educational Measurement*, *19*(3), 199-209.

Tekwe, C. D., Carter, R. L., & Ma, C. (2004). An empirical comparison of statistical models for value-added assessment of school performance. *Journal of Educational and Behavioral Statistics*, *29*(1), 11-35.

Texas Education Agency. (2006). *Grade placement committee manual: For grade advancement requirements of the student success initiative (Phase two update for 2006-2007 school year)*. Retrieved April 28, 2011, from http://www.tea.state.tx.us/student.assessment/resources/ssi/gpcmanual107.pdf

Texas Education Agency. (2007a). Appendix C - comparison of state and federal systems *2007 accountability manual: The 2007 accountability rating system for Texas public schools and school district*s (pp. 155-159) Texas Education Agency.

Texas Education Agency. (2007b). *Revised ARD committee decision-making process for the Texas assessment progra*m. Texas: Texas Education Agency.

Texas Education Agency. (2008). *Technical digest 2006-2007*. Retrieved June 15, 2011, from http://www.tea.state.tx.us/student.assessment/techdigest/yr0607/

Texas Education Agency. (2011). *TAKS vertical scale*. Retrieved September 23, 2011, from http://www.tea.state.tx.us/index3.aspx?id=3818&menu_id3=793

Texas Education Agency. (2012). *Language proficiency assessment committee (LPAC) assessment resources*. Retrieved Feb 15, 2012, from http://www.tea.state.tx.us/student.assessment/ell/lpac/

Texas Education Agency Austin Division of Student Assessment. (2000). *Texas Essential Knowledge and Skills (TEKS) educator's guide to TEKS-based assessment. elementary level TAAS grades 3-6, spring 2000. Supplement.* Texas Education Agency.

Torrance, H. (2009). Pursuing the wrong indicators? The development and impact of test-based accountability. In K. E. Ryan, & J. B. Cousins (Eds.), *The SAGE international handbook of educational evaluatio*n (pp. 483-498). Los Angeles: SAGE.

West, M. R., & Peterson, P. E. (2005). *The efficacy of choice threats within school accountability systems: Results from legislatively induced experiments.* (PEPG 05-01). Program on Education Policy and Governance.

White, K. R., Taylor, C., Carcelli, L., & Eldred, N. (1981). *State refinements to the ESEA title I evaluation and reporting system: Utah 1979-80 project. Final repor*t. Salt Lake City: Utah State Office of Education.

Wilson, R. E., Bowers, M. J., & Hyde, R. L. (2011). *Investigative report by governor's special investigators appointed to probe allegations of test tampering and related matters in the Atlanta public school system (June 30*). Atlanta, GA: Office of the Governor.

Wise, S. L., & DeMars, C. E. (2005). Low examinee effort in low-stakes assessment: Problems and potential solutions. *Educational Assessment, 10*(1), 1-17.

Appendix A. Name, definition and coding of the principal variables

SN	Variable	Description	Coding / Remarks
I. Outcome Variables (Student-level)			
1.	*STATE_M(Y)*	Scaled score on state mathematics test in target year *t* (first attempt)	Integer
2.	*STATE_R(Y)*	Scaled score on state reading test in target year *t* (first attempt)	Integer
3.	*SAT_M(Y)*	Scaled score on Stanford mathematics test in target year *t*	Integer
4.	*SAT_R(Y)*	Scaled score on Stanford reading test in target year *t*	Integer
II. Previous Year's Achievement Variables (Student-level)			
5.	*STATE_M(Y-1)*	Scaled score on state mathematics test in immediate previous grade from grade in target year (first attempt)	Integer
6.	*STATE_R(Y-1)*	Scaled score on state reading test in immediate previous grade from grade in target year (first attempt)	Integer
7.	*SAT_M(Y-1)*	Scaled score on Stanford mathematics test in immediate previous grade from grade in target year	Integer
8.	*SAT_R(Y-1)*	Scaled score on Stanford reading test in immediate previous grade from grade in target year	Integer
III. Milestone Grade's Achievement Variables (Student-level)			
9.	*STATE_M(PA)*	Scaled score on state mathematics test in grade 3 (for grade-5 students only)	Integer
10.	*STATE_R(PA)*	Scaled score on state reading test in grade 3 (for grade-5 students only)	Integer
11.	*SAT_M(PA)*	Scaled score on Stanford mathematics test in grade 3 (for grade-5 students only)	Integer
12.	*SAT_R(PA)*	Scaled score on Stanford reading test in grade 3 (for grade-5 students only)	Integer

Continue on next page…

...continued

SN	Variable	Description	Coding / Remarks
IV. Background Variables (Student-level)			
13.	*SCHOOLID*	School student belongs to at point of taking state test	Integer
14.	*GRADE*	Grade level of state test taken in target year	Integer
15.	*FEMALE*	Binary variable coding for student's gender	1: female 0: male
16.	Race/Ethnicity	A set of binary variables coding for student's race/ethnicity	$BLACK = \begin{cases} 1\ , & \text{if Black, non-Hispanic} \\ 0\ , & \text{otherwise} \end{cases}$ $HISPANIC = \begin{cases} 1\ , & \text{if Hispanic} \\ 0\ , & \text{otherwise} \end{cases}$ $ASIAN = \begin{cases} 1\ , & \text{if Asian} \\ 0\ , & \text{otherwise} \end{cases}$ $INDIAN = \begin{cases} 1\ , & \text{if American Indian/Alaska Native} \\ 0\ , & \text{otherwise} \end{cases}$ Reference group: White non-Hispanic
17.	*ECONDIS*	Binary variable coding for economically-disadvantaged status	1: Economically disadvantaged, defined as students who (a) qualify for free or reduced-price lunch; (b) are members of families that qualify for AFDC; or (c) fall into the "other economic disadvantaged" category, which includes those who are in (a) but did not apply for free or reduced-price lunch 0: Not economically disadvantaged
18.	*LEP*	Binary variable coding for Limited English Proficiency status	1: has limited English proficiency 0: not limited in English proficiency
19.	*SPED*	Binary variable coding for special-education status	1: special education 0: not special education
20.	*DISABLED*	Binary variable coding for disability status	1: disabled 0: not disabled
21.	*SPED_STATE_M*	Binary variable coding for special-education status at the point of taking the state mathematics test	1: classified as special-education 0: not classified as special-education
22.	*SPED_STATE_R*	Binary variable coding for special-education status at point of taking the state reading test	1: classified as special-education 0: not classified as special-education
23.	*SPANISH_STATE_M*	Binary variable coding for taking the Spanish-language version of the state mathematics test	1: Spanish 0: English
24.	*SPANISH_STATE_R*	Binary variable coding for taking the Spanish-language version of the state reading test	1: Spanish 0: English
25.	*SPANISH_SAT*	Binary variable coding for taking the Spanish equivalent of the Stanford tests in mathematics/reading, Aprenda	1: Spanish (Aprenda) 0: English

Appendix B. Student-level descriptive statistics on selected variables for the total and constructed analytic samples for grade-5 students in 2000 and 2001

			Analytic Samples	
Variable		**Total Sample**	**SAT-9**	**TAAS**
N		186	164	164
n		28,923	20,921	17.992
% 2000		0.52	0.50	0.47
% 2001		0.48	0.50	0.53
I. Outcome Variables (Student-level)				
STATE_M(Y)	Mean	0.10	0.15	0.20
	SD	0.88	0.83	0.77
STATE_R(Y)	Mean	0.12	0.17	0.21
	SD	0.94	0.89	0.87
SAT_M(Y)	Mean	0.11	0.19	0.32
	SD	0.96	0.93	0.87
SAT_R(Y)	Mean	0.10	0.17	0.29
	SD	0.96	0.94	0.89
II. Previous Year's Achievement Variables (Student-level)				
STATE_M(Y-1)	Mean	0.19	0.24	0.31
	SD	0.94	0.91	0.86
STATE_R(Y-1)	Mean	0.20	0.25	0.31
	SD	0.91	0.88	0.83
SAT_M(Y-1)	Mean	– 0.38	– 0.31	– 0.17
	SD	1.02	1.00	0.93
SAT_R(Y-1)	Mean	– 0.27	– 0.22	-0.07
	SD	1.00	0.97	– 0.90
III. Milestone Grade's Achievement Variables (Student-level)				
STATE_M(PA)	Mean	0.05	0.07	0.12
	SD	0.92	0.91	0.86
STATE_R(PA)	Mean	0.16	0.18	0.22
	SD	0.90	0.89	0.84
SAT_M(PA)	Mean	0.55	0.59	0.73
	SD	1.10	1.09	1.04
SAT_R(PA)	Mean	0.44	0.46	0.62
	SD	1.03	1.02	0.95
IV. Background Variables (Student-level)				
FEMALE	Mean	0.51	0.52	0.53
Race/Ethnicity				
WHITE	Mean	0.11	0.11	0.11
BLACK	Mean	0.33	0.31	0.32
HISPANIC	Mean	0.54	0.55	0.54
ASIAN	Mean	0.03	0.03	0.03
INDIAN	Mean	0.00	0.00	0.00
ECONDIS	Mean	0.80	0.80	0.78
SPED	Mean	0.11	0.10	0.03
LEP	Mean	0.26	0.25	0.24
DISABLED	Mean	0.12	0.11	0.04
SPED_STATE_M	Mean	0.07	0.07	0.04
SPED_STATE_R	Mean	0.06	0.06	0.04
SPANISH_STATE_M	Mean	0.09	0.07	0.07
SPANISH_STATE_R	Mean	0.09	0.07	0.07
SPANISH_SAT	Mean	0.09	0.08	0.08

Table 1

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

	Type of Measures	
Type of Covariates	**Covariate-Adjustment**	**Gain-Score**
1. None	CA1	GS1
2. Student-level background only	CA2	GS2
3. Student- and school-level background	CA3	GS3
4. Student-level prior achievement in milestone grade only	CA4	GS4
5. Student-level prior achievement in milestone grade and background	CA5	GS5
6. Student- and school-level prior achievement in milestone grade and background	CA6	GS6

Table 2

Distribution of between-test Spearman's rank correlations and the corresponding mean absolute differences in rank (in parentheses), by year, type of measures, and subject

Measures (# of correlations)	**Reading**			**Mathematics**		
	Min	**Median**	**Max**	**Min**	**Median**	**Max**
A. 2000						
Overall (12)	**.27 (45)**	**.30 (43)**	**.43 (38)**	**.54 (35)**	**.58 (33)**	**.63 (30)**
CA (6)	.27 (44)	.30 (44)	.43 (38)	.54 (35)	.58 (33)	.61 (32)
GS (6)	.28 (45)	.30 (43)	.33 (42)	.57 (33)	.62 (32)	.63 (30)
B. 2001						
Overall (12)	**.39 (42)**	**.45 (38)**	**.63 (31)**	**.53 (35)**	**.56 (34)**	**.62 (32)**
CA (6)	.42 (40)	.47 (38)	.63 (31)	.53 (35)	.56 (34)	.62 (32)
GS (6)	.39 (42)	.42 (40)	.50 (36)	.53 (35)	.56 (34)	.57 (33)

Table 3

Distribution of between-test observed percentage agreement, percentage of chance agreement (in parentheses) and Cohen's kappa (in italics), by subject, type of measures, and classification scheme

	Classification Scheme								
Measures	**1 (±1 posterior SD)**			**2 (Quintiles)**			**3 (Unequal, Asymmetric)**		
(#)	**Min**	**Mean**	**Max**	**Min**	**Mean**	**Max**	**Min**	**Mean**	**Max**
A. Reading									
Overall (24)	**64 (57)**	**71**	**79 (45)**	**24 (20)**	**31**	**41 (20)**	**31 (28)**	**40**	**59 (28)**
	.14	***.26***	***.45***	***.05***	***.13***	***.27***	***.04***	***.16***	***.28***
CA (12)	66 (57)	71	74 (62)	24 (20)	31	41 (20)	34 (28)	40	47 (28)
	.14	*.25*	*.45*	*.05*	*.13*	*.25*	*.04*	*.16*	*.28*
GS (12)	64 (57)	71	79 (63)	24 (20)	30	40 (20)	31 (28)	39	48 (28)
	.21	*.28*	*.34*	*.05*	*.14*	*.27*	*.09*	*.17*	*.26*
B. Mathematics									
Overall (24)	**66 (53)**	**74**	**81 (64)**	**30 (20)**	**37**	**41 (20)**	**38 (28)**	**44**	**48 (28)**
	.22	***.36***	***.48***	***.13***	***.21***	***.27***	***.15***	***.22***	***.28***
CA (12)	66 (53)	73	81 (64)	32 (20)	37	41 (20)	39 (28)	44	48 (28)
	.29	*.37*	*.46*	*.13*	*.21*	*.25*	*.15*	*.23*	*.26*
GS (12)	68 (55)	74	80 (63)	30 (20)	37	40 (20)	38 (28)	44	47 (28)
	.22	*.34*	*.48*	*.15*	*.21*	*.27*	*.15*	*.22*	*.28*

Table 4

Estimated variance components associated with different sources of variation in the set of school-performance estimates

						% of Total Variance Associated with Factor:				
				Estimated	% of Total					
Source of Variation[6]	**SS**	**df**	**MS**	**VC**	**Variance**	***t***	***u***	***y***	***m***	***c***
School	287	163	1.7618	0.0184	17.30					
School × Test (t)	69	163	0.4264	0.0089	8.37	✓				
School × Measure (m)	20	163	0.1208	0.0025	2.37				✓	
School × Covariate (c)	26	815	0.0316	0.0020	1.85					✓
School × Subject (u)	42	163	0.2553	0.0053	5.01		✓			
School × Year (y)	143	163	0.8754	0.0182	17.19			✓		
School × t × m	4	163	0.0223	0.0009	0.87	✓			✓	
School × t × c	3	815	0.0037	0.0005	0.43	✓				✓
School × t × u	20	163	0.1203	0.0050	4.72	✓	✓			
School × t × y	59	163	0.3600	0.0150	14.14	✓		✓		
School × m × c	5	815	0.0062	0.0008	0.72				✓	✓
School × m × u	2	163	0.0105	0.0004	0.41		✓		✓	
School × m × y	8	163	0.0475	0.0020	1.86			✓	✓	
School × c × u	2	815	0.0024	0.0003	0.27		✓			✓
School × c × y	9	815	0.0105	0.0013	1.22			✓		✓
School × u × y	28	163	0.1723	0.0072	6.77		✓	✓		

[6] The main effects of “Test”, “Measure”, “Covariate”, “Subject”, and “Year” are zero by construction as the school-performance estimates were generated separately for each combination of subject, year, and model specification.

Source of Variation[6]	**SS**	**df**	**MS**	**Estimated VC**	**% of Total Variance**	**% of Total Variance Associated with Factor:**				
						t	***u***	***y***	***m***	***c***
School × t × m × c	1	815	0.0011	0.0003	0.25	✓			✓	✓
School × t × m × u	1	163	0.0043	0.0004	0.33	✓	✓		✓	
School × t × m × y	2	163	0.0117	0.0010	0.91	✓		✓	✓	
School × t × c × u	1	815	0.0018	0.0004	0.40	✓	✓			✓
School × t × c × y	3	815	0.0037	0.0009	0.85	✓		✓		✓
School × t × u × y	24	163	0.1446	0.0120	11.35	✓	✓	✓		
School × m × c × u	0	815	0.0004	0.0001	0.07		✓		✓	✓
School × m × c × y	2	815	0.0020	0.0005	0.44			✓	✓	✓
School × m × u × y	1	163	0.0076	0.0006	0.59		✓	✓	✓	
School × c × u × y	1	815	0.0009	0.0002	0.18		✓	✓		✓
School × t × m × c × u	0	815	0.0001	0.0000	0.02	✓	✓		✓	✓
School × t × m × c × y	1	815	0.0008	0.0004	0.33	✓		✓	✓	✓
School × t × m × u × y	0	163	0.0028	0.0004	0.42	✓	✓	✓	✓	
School × t × c × u × y	0	815	0.0005	0.0002	0.21	✓	✓	✓		✓
School × m × c × u × y	0	815	0.0002	0.0000	0.03		✓	✓	✓	✓
Residual	0	815	0.0001	0.0001	0.08					
Total	762	15743		0.1061	100	44	31	57	10	7

Key: SS = sum of squares; df = degree of freedom; MS = mean square; VC = variance component

Table 5

Taxonomy of fitted relationships between students' TAAS versus SAT-9 score-gaps and various student and school characteristics indicative of differential exposures to inappropriate behavioral responses under high-stakes conditions, by subject

	1	**2**	**3**	**4**	**5**	**6**
A. Reading (17,787 students in 164 schools)						
<u>*Fixed Effects*</u>						
NONWHITE ($\hat{\alpha}_{100}$)		0.245***	0.194***	0.178***	0.178***	0.181***
ECONDIS ($\hat{\alpha}_{200}$)			0.142***	0.128***	0.121***	0.121***
SM_NONWHITE ($\hat{\alpha}_{010}$)				0.524***		-0.373
SM_ECONDIS ($\hat{\alpha}_{020}$)					0.439***	0.676***
Intercept	-0.057	-0.061	-0.064	-0.063	-0.062	-0.062
TAAS_R(Y-1)	0.025***	0.032***	0.037***	0.038***	0.038***	0.038***
TAAS_R(Y-1)_sq	-0.055***	-0.054***	-0.054***	-0.053***	-0.053***	-0.053***
FEMALE	0.040***	0.038***	0.036***	0.036***	0.036***	0.036***
LEP	0.121***	0.103***	0.087***	0.086***	0.084***	0.084***
SPECED	0.174**	0.189***	0.187***	0.186***	0.186***	0.186***
DISABLED	-0.118*	-0.113*	-0.113*	-0.110*	-0.110*	-0.110*
<u>*Random Effects*</u>						
Between-year ($\hat{\sigma}_v^2$)	0.003***	0.002***	0.002***	0.002***	0.002***	0.003***
Between-school, within-year ($\hat{\sigma}_u^2$)	0.088***	0.077***	0.071***	0.064***	0.060***	0.059***
Within-school & within-year ($\hat{\sigma}_e^2$)	0.440***	0.437***	0.436***	0.436***	0.436***	0.436***
<u>*Goodness-of-Fit*</u>						

Δ(Deviance)
= Deviance(Null) –
Deviance(Model)
	263.556	421.074	494.924	525.222	540.436	543.102

Key: * p <.05; ** p < .01; *** p < .001. All variables were grand-mean-centered.

Continue on next page...

Table 5 (continued)

Taxonomy of fitted relationships between students' TAAS versus SAT-9 score-gaps and various student and school characteristics indicative of differential exposures to inappropriate behavioral responses under high-stakes conditions, by subject

	1	**2**	**3**	**4**	**5**	**6**
B. Mathematics (17,787 students in 164 schools)						
Fixed Effects						
NONWHITE ($\hat{\boldsymbol{\alpha}}_{\mathbf{100}}$)		0.200***	0.160***	0.147***	0.146***	0.149***
ECONDIS ($\hat{\boldsymbol{\alpha}}_{\mathbf{200}}$)			0.113***	0.101***	0.095***	0.095***
SM_NONWHITE ($\hat{\boldsymbol{\alpha}}_{\mathbf{010}}$)				0.470***		-0.459*
SM_ECONDIS ($\hat{\boldsymbol{\alpha}}_{\mathbf{020}}$)					0.407***	0.699***
Intercept	-0.094***	-0.097***	-0.099***	-0.098***	-0.098***	-0.098***
TAAS_M(Y-1)	-0.020***	-0.014*	-0.012*	-0.011	-0.011	-0.011
TAAS_M(Y-1)_sq	-0.101***	-0.101***	-0.101***	-0.101***	-0.101***	-0.101***
FEMALE	0.053***	0.053***	0.052***	0.052***	0.052***	0.052***
LEP	0.107***	0.092***	0.079***	0.078***	0.077***	0.076***
SPECED	0.069	0.081	0.080	0.079	0.079	0.079
DISABLED	-0.043	-0.039	-0.039	-0.037	-0.036	-0.037
Random Effects						
Between-year ($\hat{\boldsymbol{\sigma}}_{\boldsymbol{v}}^{2}$)	0.000	0.000	0.000	0.000	0.000*	0.000**
Between-school, within-year ($\hat{\boldsymbol{\sigma}}_{\boldsymbol{u}}^{2}$)	0.082***	0.074***	0.070***	0.064***	0.060***	0.059***
Within-school & within-year ($\hat{\boldsymbol{\sigma}}_{\boldsymbol{e}}^{2}$)	0.378***	0.377***	0.376***	0.376***	0.376***	0.376***
Goodness-of-Fit						

Δ(Deviance)
= Deviance(Null) –
Deviance(Model)
	730.932	853.384	907.144	932.444	947.846	951.970

Key: * p <.05; ** p < .01; *** p < .001. All variables were grand-mean-centered.

Table 6.

Between-test (BT) and between-subject (BS) Spearman's rank correlations for school-performance measures, by year, subject, and model specification

	Type of Covariates								
Type of Measures	**1. None**			**2. Student-level Background Only**			**3. Student- & School-level Background**		
	BT(R)	**BS(TAAS)**	**BT(M)**	**BT(R)**	**BS(TAAS)**	**BT(M)**	**BT(R)**	**BS(TAAS)**	**BT(M)**
A. Full Analytic Samples									
1. 2000									
CA	0.43	**0.69**	0.58	0.30	**0.72**	0.59	0.30	**0.72**	0.61
CA (with PA)	0.40	**0.67**	0.54	0.27	**0.71**	0.55	0.28	**0.72**	0.58
GS	0.30	**0.69**	0.58	0.30	**0.72**	0.62	0.33	**0.70**	0.63
GS (with PA)	0.30	**0.70**	0.57	0.28	**0.73**	0.62	0.31	**0.70**	0.61
2. 2001									
CA	0.63	**0.70**	0.61	0.48	**0.68**	0.54	0.42	**0.71**	0.53
CA (with PA)	0.59	**0.71**	0.62	0.46	**0.70**	0.58	0.45	**0.74**	0.55
GS	0.50	**0.66**	0.56	0.42	**0.68**	0.56	0.40	**0.69**	0.57
GS (with PA)	0.50	**0.68**	0.56	0.42	**0.69**	0.56	0.39	**0.68**	0.53

Continue on next page…

Key: BT – between-test; BS – between-subject; R – reading; M – mathematics; CA – covariate-adjustment measure; GS – gain-score measure; PA – prior achievement

Table 6 (continued)

	Type of Covariates								
Type of Measures	**1. None**			**2. Student-level Background Only**			**3. Student- & School-level Background**		
	BT(R)	**BS(TAAS)**	**BT(M)**	**BT(R)**	**BS(TAAS)**	**BT(M)**	**BT(R)**	**BS(TAAS)**	**BT(M)**
B. Restricted Samples									
1. 2000									
CA	0.48	**0.68**	0.54	0.31	**0.71**	0.57	0.30	**0.71**	0.59
CA (with PA)	0.43	**0.66**	0.50	0.27	**0.71**	0.52	0.22	**0.71**	0.55
GS	0.33	**0.68**	0.55	0.30	**0.72**	0.62	0.33	**0.69**	0.63
GS (with PA)	0.34	**0.69**	0.55	0.29	**0.72**	0.62	0.31	**0.69**	0.61
2. 2001									
CA	0.63	**0.69**	0.60	0.49	**0.67**	0.55	0.46	**0.71**	0.55
CA (with PA)	0.58	**0.71**	0.62	0.47	**0.70**	0.58	0.47	**0.74**	0.57
GS	0.48	**0.66**	0.56	0.42	**0.68**	0.58	0.41	**0.68**	0.59
GS (with PA)	0.48	**0.68**	0.56	0.42	**0.68**	0.56	0.40	**0.67**	0.53

Key: BT – between-test; BS – between-subject; R – reading; M – mathematics; CA – covariate-adjustment measure; GS – gain-score measure; PA – prior achievement

Table 7

Estimated average within-school student-level between-test (BT) and between-subject (BS) Pearson correlations for students with non-missing scores on both the TAAS and the SAT-9, by year, and subject

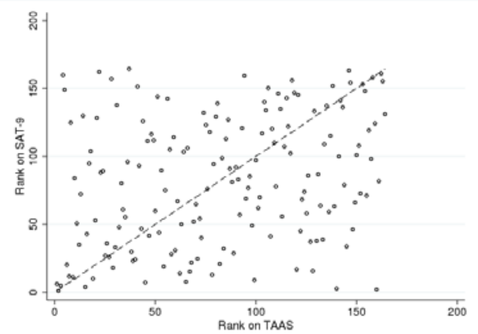
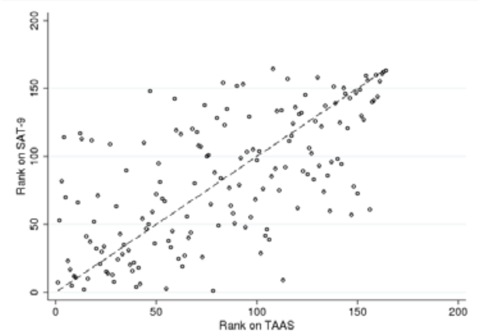
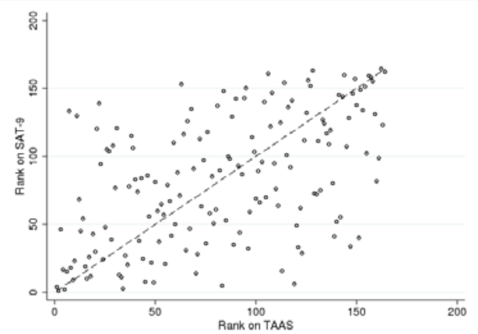
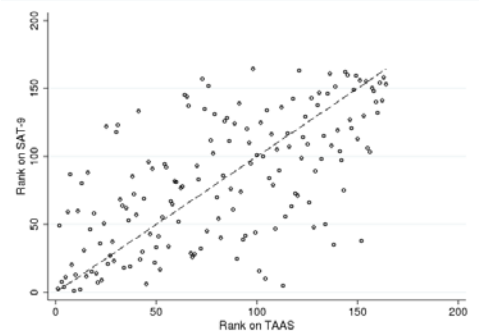
Year	**BT(R)**	**BS(TAAS)**	**BT(M)**
2000	**.70 (.17)**	.61 (.15)	**.73 (.28)**
2001	**.65 (.20)**	.56 (.16)	**.70 (.23)**

Key: R – reading; M – mathematics.

Note: Standard deviations of the distributions of estimated within-school correlation across schools (i.e., estimated random effects for the slope parameter) are in parentheses. All estimated random effects for the slope parameter are non-zero ($p < .05$).

Figure 1

Scatter-plots of schools' ranks on the SAT-9 and the TAAS, for the minimum and maximum estimated Spearman's rank correlations across all model specifications and years, by subject

	Min	**Max**
A. Reading	Correlation = .27 (2000, CA5)  	Correlation = .63 (2001, CA1) 
B. Mathematics	Correlation = .53 (2001, GS6)  	Correlation = .63 (2000, GS3) 

Note: On each scatter-plot, the dotted line is the identity line, i.e., rank on SAT = rank on state test.