ISERP

Working Paper 07-06

Does Gaming the System Affect Students' Academic Achievement?

JENNIFER BOOHER-JENNINGS
Department of Sociology
Columbia University

ANDREW A. BEVERIDGE
Department of Sociology
Queens College and CUNY Graduate Center

Columbia University in the City of New York
May 2007

INSTITUTE FOR SOCIAL AND ECONOMIC RESEARCH AND POLICY
PIONEERING SOCIAL SCIENCE RESEARCH AND SHAPING PUBLIC POLICY

DOES GAMING THE SYSTEM AFFECT STUDENTS' ACADEMIC ACHIEVEMENT?

Jennifer Booher-Jennings
Department of Sociology
Columbia University

Andrew A. Beveridge
Department of Sociology
Queens College and CUNY Graduate Center

ISERP Working Paper 07-06

May 2007
12 Tables
5 Figures
~10,000 Words

An earlier version of this paper was presented at the 2007 Meetings of the American Educational Research Association. We thank Linda Darling-Hammond, Tom DiPrete, David Rindskopf, Chris Weiss, and Julian Vasquez Heilig for their helpful comments and suggestions. Direct all correspondence to: Jennifer Booher-Jennings, Columbia University Department of Sociology, 1180 Amsterdam Avenue, 413 Fayerweather Hall, New York, NY, 10027. Email: jlj2102@columbia.edu.

Abstract

A growing body of evidence suggests that schools use test exemption to game educational accountability systems. However, it is not known whether test exemption affects students' academic progress. Analyzing data from an urban school district in Texas, we find that special education students make larger achievement gains when they are tested. Using our most conservative estimates, the effect of being tested is approximately .40 standard deviations in reading and .28 standard deviations in math for grades 3-8. Because special education students are more likely to be minority and poor students and these students are more likely to be exempted than their white and non-poor special education counterparts, the exemption of special education students contributes to the growth of black-white, Hispanic-white, and high-low socioeconomic status achievement gaps.

| 1 | INTRODUCTION

For the last fifty years, American education policy has attempted to close achievement gaps between advantaged and disadvantaged groups. Despite these efforts, significant gaps persist between the academic performance of white and African-American and Hispanic, poor and privileged, special education and general education, and English language learners and native speakers (Brooks-Gunn, Klebanov, & Duncan, 1995; Cook & Evans, 2000; Fryer & Levitt, 2004; Phillips, Crouse, & Ralph, 1998). Most recently, the No Child Left Behind Act of 2001 sought to narrow these gaps by requiring states to test students in grades 3-8 annually and to disaggregate scores by race, ethnicity, poverty, special education, and English proficiency status. Schools are held accountable for improving the performance of students in each of these subgroups so that all students reach proficiency in math and reading by 2014. The theory underlying this law is that educators will channel more time, resources, and attention to minority students, poor students, and students with special educational needs when their performance is made public and schools are held accountable for results.

However, many recent studies have documented that schools "game the system" to artificially inflate their passing rates (Cullen & Reback, 2006; Deere & Strayer, 2001; Figlio, 2005; Figlio & Getzler, 2002; Jacob, 2005; Jacob & Levitt, 2003; Haney, 2000; McNeil, 2000, 2005; McNeil & Valenzuela, 2001; Vasquez Heilig, 2006). At least two types of gaming are possible. In one scenario, educators do not alter the way they allocate instructional resources. Instead, they adopt a series of creative accounting practices such as exempting potentially low-scoring students from state tests in order to artificially increase passing rates (Cullen & Reback, 2006; Figlio & Getzler, 2002; Figlio, 2005; Jacob, 2005). In a second scenario, educators fundamentally change the way that they distribute educational resources by diverting resources to students close to passing the test and thus most likely to improve the passing rate, or shifting resources away from exempted students who will not sit for state tests (Gillborn & Youdell, 2000; Reback, 2006). These two mechanisms have very different implications for the No Child Left Behind Act's potential to attenuate achievement gaps. While the first scenario might leave real achievement gaps unaltered while reporting that such gaps have narrowed, the second has the potential to further exacerbate these gaps.

In this study, we take advantage of a unique feature of Texas' accountability system to test the hypothesis that educators game the accountability system by exempting low-scoring students from taking high-stakes tests and subsequently allocate resources strategically to tested students. Texas allows educators to exempt special education students and English Language Learners (ELLs) from taking mainstream high-stakes state tests. While the rules regarding ELL exemptions are more stringent, only one-third of special education students in our study take mainstream state tests. In place of the high-stakes exam, exempted special education students take an alternate assessment for which each student's Individual Education Plan committee sets the passing rate individually. As a result of this feature—that is, some special education students "count" towards the school's accountability rating, while others do not—schools and teachers may recognize that the marginal investment in a student who counts towards the accountability rating will yield greater returns than an investment in a student who does not "count." By analyzing this setting, we can gain insight into two key policy questions: first, how might schools use exemptions if the exemption provisions in the No Child Left Behind Act are extended, and second, in the context of a high-stakes testing system, do special education students gain more academically when they are tested than they would if they were not tested?

| 2 | LITERATURE REVIEW

Should NCLB hold schools accountable for all students, including special education students and English Language Learners (ELL)? Initially, NCLB allowed only one percent of students to take an alternate assessment while being counted towards the 95% participation requirement. In response to mounting pressure, the U.S. Department of Education amended this rule in March 2005, inviting states to apply for a variance from the one percent cap and develop alternate assessments for an additional two percent of disabled students. Altogether, the regulations now allow schools to test up to three percent of students on an alternate assessment and count these scores toward adequate yearly progress requirements (U.S. Department of Education, 2007). However, as NCLB approaches reauthorization, participation requirements have again resurfaced as a key policy concern.

Two competing goals of accountability systems—to improve the performance of all students and to fairly measure the performance of schools—suggest different remedies to the participation question. Proponents of holding all students to the same standard (i.e., testing all students on the same grade-level tests) contend that exemptions or alternate assessments for these students perpetuate the educational neglect that NCLB is intended to correct. If students are exempted or held to a different standard, they argue, schools will have little incentive to focus time and attention on these students. As a result, they are unlikely to ever reach proficiency. The consequence of excluding some students, according to this view, is their loss of access to scarce educational resources. Avoiding such negative consequences would entail testing and holding schools accountable for all students, though measurement accuracy would be sacrificed and schools serving the most vulnerable students would be unfairly penalized.

Opponents of NCLB's current participation requirements hold that grade-level English-only tests are inappropriate measures for some special education and ELL students. Requiring students to take these tests, it is argued, is detrimental to the students themselves and punishes schools serving large numbers of these students. If accuracy of measurement is privileged, these students should be excluded from accountability calculations. Adherents to this view believe that there are valid reasons, from both an educational and a measurement perspective, for excluding categories of students from schools' scores.

A growing body of literature suggests that when schools are given the opportunity to exempt students, exemptions are used strategically. Figlio and Getzler (2002) analyzed data from Florida to determine whether low-achieving students were more likely to be classified as special education following the introduction of high-stakes tests. In their most conservative estimates (estimating student fixed effects models and controlling for separate classification time trends for low and high income students), they found that the introduction of high-stakes testing increased non-low-income students' risk of classification by .9 percentage points. The effects on low-income students were larger; the introduction of high-stakes testing was associated with a 3.9 percent increase in students' risk of classification. Figlio and Getzler further determined that low-scoring students were more likely to be classified than high-achieving students, and that this effect was strongest in high-poverty schools.

Jacob (2005) also investigated the relationship between high-stakes testing and special education classification. Analyzing data from Chicago, his study incorporated a rich set of controls (student demographic characteristics and prior achievement, as well as school and neighborhood characteristics) and found that the introduction of high stakes testing was associated with a one percent increase in special education classification. However, these effects varied by the school's achievement; special education classification for low-achieving students (those in the bottom quartile of the national distribution) attending a school in the bottom one-third of Chicago schools increased 3.6 percent, while the increase in middle-achieving schools was 2.1 percent. High-stakes testing had no effect on special education classification for low-achieving students in the top one-third of schools. Jacob found additional manipulation of score reporting as Chicago maintained a policy where students could be tested and their scores withheld. Low-achieving students in low-achieving schools were 1.6 percent less likely to have their scores reported, while low-achieving students in middle-achieving schools were 2.2 percent less likely to have their scores reported. Again, Jacob found that high-stakes testing had no effect on score reporting in high-achieving schools.

Finally, Cullen and Reback (2006) analyzed data from Texas and found that schools use exemption strategically. In a particularly comprehensive treatment of individual schools' incentives to exempt students, they determined that proximity to the next accountability rating increased the percentage of students exempted by 11 percent. Moreover, they also found that when the performance of Hispanic and African-American students would keep schools from achieving a higher rating, exemption rates increased for these groups by 7 and 14 percent, respectively.

These studies provide strong evidence that schools use exemption provisions to game the system. However, the total impact of exemption on school passing rates and achievement gaps, as well on the impact of exemption of student achievement, has not been addressed in the current literature. These omissions result both from data availability as well as the structure of many exemption policies. Few states and districts require students to sit for multiple assessments; if a student is exempted from a high-stakes test, there is no supplementary record of the student's achievement. Furthermore, estimating the "treatment effect" of testing on students requires that a substantial number of students with similar propensities to take a high-stakes exam are assigned to either a "testing" or "exemption" condition. Yet as the studies above document, strong selection pressures drive higher achieving students into the high-stakes testing treatment. Consequently, it is difficult to construct accurate comparison groups for tested students.

In what follows, we use a unique student-level dataset from the Houston Independent School District to fill this gap in the literature. In line with the studies reviewed above, we first assess the extent to which educators utilize exemption strategically. If educators use exemption as a tool to increase their schools' scores, we expect to find that low-scoring students have higher odds of test exemption. We then evaluate the impact of these exemptions on schools' passing rates and determine how test exemption changes schools' passing rates as well as the estimated size of achievement gaps. Low-scoring students are not equally distributed across demographic groups, and if low-scoring students are more likely to be exempted, their exclusion will alter the estimated achievement gaps of some groups more than others. Next, we estimate the causal effect of taking Texas' high-stakes test on special education students' year-to-year academic growth and determine the extent to which the treatment effect of testing varies by students' propensity to take Texas' high-stakes test. Finally, we test the robustness of our estimates by simulating the impact of an unobserved variable on the magnitude of treatment effects. Taken together, these analyses inform current policy debates over the issue of test exemption in high-stakes accountability systems.

| 3 | DATA AND METHODS

In this study, we analyze a longitudinal dataset of all students tested in the Houston Independent School District (HISD) for the six school years ending in 2003-2004. HISD is the seventh largest school district in the country and the largest in the state of Texas. Fifty-six percent of HISD students are Hispanic, while 31 percent are African American. Seventy-four percent of students qualify for free or reduced-price lunch. Twenty-four percent of students are classified as Limited English Proficient (LEP), and 11 percent are classified as special education. The composition of HISD thus makes it a useful test case in evaluating NCLB's potential impact on poor students, students of color, and students with special needs.

Our data include test scores for students in grades one to 11 (the Texas Assessment of Knowledge and Skills (the TAKS), the Stanford 10 Achievement Test[1], and the Spanish-language version of this test (the Aprenda) both of which are nationally norm-referenced exams), demographic data, and classification status data for approximately 160,000 students per year for each of the six years. Approximately 130,000 of these students are enrolled in TAKS-tested grades, grades 3-11. Because our dataset includes two measures of student achievement for each school year, it represents a significant improvement upon previously available student-level administrative data.

We present three sets of analyses to address our questions of interest:

1) Do schools strategically exempt low-scoring students, and how does the use of exemption vary across schools?

2) How does student exemption impact schools' passing rates and the estimated size of achievement gaps?

3) What is the causal effect of testing on special education students' academic growth? In other words, do tested special education students gain more academically than they would if they were not tested?

We first estimate students' odds of exemption using a multilevel logistic regression model, where students (level 1) are nested within-schools (level 2). The dependent variable in this analysis is whether the student took the TAKS test in math or reading. As discussed previously, an off-grade level test, the SDAA, is also offered for special education students, and Texas counts these students as passing if they meet the criteria set by their special education committee. The state mandates that special education students whose individual education plans specify instruction in grade-level state standards should take the TAKS. Accordingly, we include in our analysis only those students who also took the Stanford or Aprenda tests, which are on-grade level tests. Based on the state's regulations, any student sitting for the Stanford or Aprenda exams should also sit for the TAKS exam.

Because the relationship between TAKS test taking and the two versions of the low-stakes test (Stanford and Aprenda) are potentially different, we estimate separate models for the English and Spanish TAKS tests. In the first model, taking the English TAKS is the dependent variable; all students in this analysis took the Stanford 10. The second model includes students who took the Spanish Aprenda test (only offered in grades 3-5), and the dependent variable is taking the Spanish TAKS exam. Each of these models takes the form:

$$\text{logit}\{\Pr y_{ij} = 1 \mid (\mathbf{X}_{ij}, \zeta_j)\} = \beta_1 + \beta_2\text{LOW-STAKES SCORE}_i + \beta_3\text{GRADE LEVEL}_i + \beta_4\text{RACE}_i + \beta_5\text{POOR}_i + \beta_6\text{LEP}_i + \beta_7\text{SPECIAL EDUCATION}_i + \zeta_j + \varepsilon_{ij}$$

where y_{ij} is the test taking status of student i in school j, $\mathbf{X}_{ij}$ are characteristics of the student i displayed on the right hand side of the equation, and ζ_j is a random intercept that varies across schools. Both GRADE and RACE are vectors of dummy variables, where the reference categories are 11th grade and Asian, respectively. In our model of Spanish TAKS test taking, the reference grade is 5th grade; race is not included as all students are Hispanic. Because we expected the effects of the low-stakes test on high-stakes testing to be non-linear, the low-stakes score is represented as five dummy variables coded as 1 if the student is in the 0-9th, 10-19th, 20-29th, 30-39th, and 40-49th percentile of the national distribution. The reference category is a score in the 50th percentile or above.

How would we know if schools used exemption to "game the system?" After all, many educators feel that they are acting in the best interests of the students themselves when they exempt students from taking a test that is inappropriate for them. In order to rule out this alternate explanation for educators' behavior, we exclude from our analyses those students with cognitive disabilities so severe that they *did not* also take the low-stakes Stanford 10 or Aprenda exam (1.5% of students in Houston). The Stanford 10 and the Aprenda are on-grade level tests; students who are able to take this on-grade level low-stakes test should, from an educational standpoint, also be able to take the high-stakes test. If low-scoring students are more likely to be exempted, this would suggest that exemption is used strategically. Our models predicting English TAKS test taking include 117,593 students for Reading and 117,635 students for Math. Our models predicting Spanish TAKS test taking include 9,412 students for Reading and 9,397 students for Math. Altogether, our models for this section of the paper include 127,005 and 127,032 students for reading and math, respectively. In this paper, we estimate these models only for the 2003-2004 school year.

In the second part of our study, we estimate the impact of student exemption on schools' scores and the perceived size of the achievement gap. Because our data include a second measure of achievement, we use multiple imputation to estimate high-stakes scores for each of the students excluded from the test (Rubin, 1987). Traditional imputation replaces each missing value with a single prediction. Using Markov Chain Monte Carlo methods, multiple imputation takes into account the uncertainty about the correct value, and instead identifies a set of possible values for each missing data point. Students' Stanford/Aprenda scores, in addition to student demographics, program status, and grade level were all used in the imputation. Five imputations were performed, and the differences between the imputations were not statistically significant. We then re-estimated schools' passing rates and simulated how schools would have performed if all students were tested and counted in schools' pass rates.

Finally, we ask whether an "educational treatment"—that is, being tested on the high-stakes test—affects special education students' academic growth between 2003 and 2004. Before describing the analytic strategy we use to estimate these effects, it is important to draw a distinction between the way that the term "treatment" is generally used and the way it is deployed in this paper. For example, in the context of a medical study of the effects of a drug on blood pressure, the "treatment" is the drug, and the difference in blood pressure at the end of the study between patients randomly assigned to control and treatment groups is the "treatment effect." In this case, the test itself is not the treatment. Rather, the treatment is the educational process set into motion if a student will be tested on a high-stakes test. Two mechanisms are possible. First, teachers may invest more time and resources in a special education student who will be tested than one who will not; as a result, this student will exhibit more academic growth. Second, if students are notified early in the school year that they will be tested and they are more motivated by the TAKS test than the special education SDAA test, they may exert more effort, which may in turn lead to larger achievement gains.

Ideally, in order to test the hypothesis that special education students exhibit greater achievement gains when they are tested, we would conduct an experiment and randomly assign students to tested and untested groups. Since conducting such an experiment is not feasible, we adopt a counterfactual framework of causality (Rubin, 1977; Rosenbaum & Rubin 1983a, 1984, 1985; Rosenbaum, 2002). Here a causal effect is defined as the difference between the academic growth a special education student in a world where he is tested compared to the growth the same student would have exhibited had he not been tested. Of course, no student can, in a given year, be both tested and not tested, a problem which Holland (1986) coined "the fundamental problem of causal inference."

Establishing the causal effect of being tested on the high-stakes test on students' academic achievement is not straightforward because students are not randomly assigned to "treatment" and "control" conditions. On many observable characteristics, tested special education students differ from those who are not tested. Most tested students have very low probabilities of being exempted, while most exempted students have very low probabilities of being tested. To address this selection bias and accurately estimate the causal effect of testing on students' year-to-year academic growth, we used propensity score matching, which allows us to simulate treatment and control groups from observational data.

With unlimited data, we could match students perfectly on all observable characteristics to which we have access. In the absence of such data, we can construct a one number summary of students' probability of being tested: the propensity score. Untested students with similar propensities to be tested serve as the counterfactual for tested students. If matched treated and control students are balanced on students' observed pre-treatment characteristics—that is, treatment and control groups have statistically indistinguishable differences in their observable characteristics—the differences in achievement gains for students can be understood as the effect of being tested (Rosenbaum & Rubin, 1983a).

To predict each student's propensity score, we estimated a logistic regression where the dependent variable was taking the high-stakes test. We included in our model a set of 53 pre-treatment student-level covariates from our dataset and teacher and school-level covariates drawn from Texas' Academic Excellence Indicator System. These variables included students' race, free lunch status, limited English proficiency status, students' Stanford scale score at the end of the 2002-2003 school year, grade level, school size, per-pupil expenditures, teacher characteristics such as experience, salary, and race, school passing rates in the prior year, and school-level demographic composition variables including the percent of students who are bilingual, gifted, African-American, Hispanic, and special education. (A full list of these variables, as well as means for tested and untested students, is available in Appendix Table A1.) Achieving optimal covariate balance between matched treatment and control groups, rather than statistical significance, was the key criterion for keeping a covariate in our model predicting students' propensity scores. We employed within caliper (.001) nearest neighbor matching with replacement. For example, if a tested student's predicted probability of taking the high-stakes test was .532, and no untested student had a predicted probability of .532 ± .001, this case was not matched and thus is not included in our estimates of treatment effects. While there is a tradeoff in using such a tight matching criterion, as matching fewer cases means that our treatment effect estimates have larger standard errors, we preferred this outcome to matching dissimilar students and potentially upwardly biasing our treatment effect estimates. We further constrained our matches in two ways. We required students to match with their peers in the same grade level, and required LEP students to match only with their LEP counterparts.

After matching tested and untested students with similar propensities to be tested, we tested the simulated treatment and control groups to ensure that matches were balanced on all 53 pre-treatment covariates. Our sample included the 9,566 special education students enrolled in grades 3-11 who took the reading component of the Stanford Achievement Test in 2002-2003 and 2003-2004, and the 9,563 special education students who took the math component of this test in the same years.

Our first objective was to establish the "average treatment effect on the treated." Here, this refers to the effect of taking the high-stakes test on the average academic gain of the children in the "treatment" group compared to the outcomes these children would have had they been in the "control," or untested, group, i.e.:

$$E[Y_i(1)-Y_i(0) \mid Z_i=1] = E[Y_i(1) \mid Z_i=1] — E[Y_i(0) \mid Z_i=1]$$

where Z_i is the treatment experienced by student *i* and Y_i is an outcome affected by the treatment.[2] Because we are interested not only in the average treatment effect, but in how this effect varies across the propensity score distribution, we then divided students into four quartiles based on their propensity to be tested and estimated an average treatment effect within each quartile.

In these analyses, academic growth is operationalized as the change in students' Stanford National Percentile score between the 2003 and 2004 administrations of the test (these national norms remained constant across the two years and were not affected by HISD students' performance).[3] We chose this metric as it is familiar to most researchers and educators, and is easily interpretable. Students take the Stanford exam in the spring, approximately one month before they take the TAKS exam. However, Stanford results *are* *not* available to teachers or parents until well after the TAKS tests are administered, and thus are not being used in real time to select students to take the TAKS.

The observable characteristics of students, teachers, and schools to which we have access go a long way towards controlling for differences in the school environments that students experience, but we recognize that there are also unobserved elements of schools that affect both students' probabilities of treatment as well as students' academic growth. We thus used two different matching approaches. First, we constrained our matches to students in the same school with similar propensities to be tested and estimated these effects separately for grades 3-8 and grades 9-11. Next, we allowed tested students to match with untested students in any HISD school and compared these estimates with our within-school matching estimates. Upon finding these results to be qualitatively similar, we took advantage of the larger sample allowed by the between-school matching approach to examine both treatment heterogeneity more closely across levels of schooling as well as across the propensity score distribution.

The primary threat to our treatment effect estimates is the potential presence of an omitted variable that both affects students' odds of being tested and students' academic growth (Rosenbaum & Rubin 1983b). For example, if tested students had significantly larger levels of school engagement than did exempted students, these students might be more likely to be tested and to gain more academically. In other words, we want to know how large the effect of an omitted variable would have to be to diminish our treatment effect estimates. To address this concern, we generated variables to mimic the effects of two of our strongest predictors of test taking and academic gain in our model: classification in special education for at least two years (2003 and 2004) and being an African-American student. We then re-estimate our treatment effects separately under each of these cases to demonstrate how robust our results are to the presence of unobserved variables.

| 4 | RESULTS

Descriptive Results

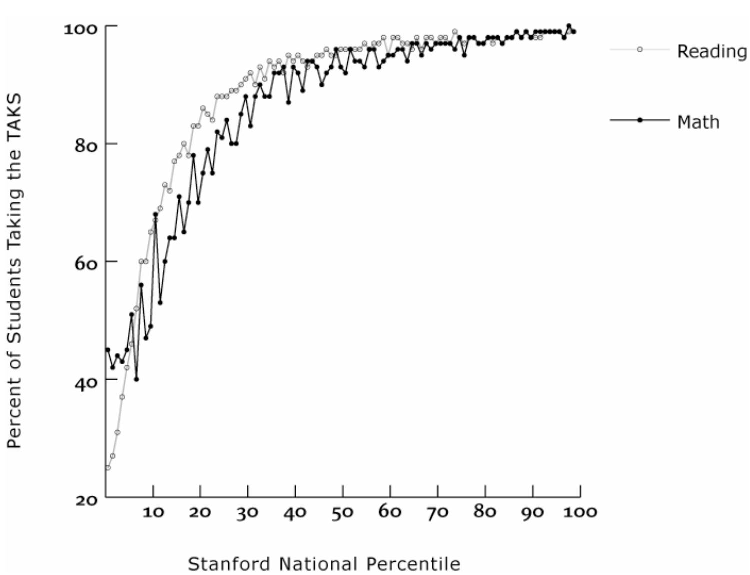
Table 1 presents the percent of students taking the reading and math TAKS by demographic group and by grade. Overall, 87.6 percent of students take the TAKS reading test, and 88 percent take the math test. Test taking, however, varies significantly by demographic group and by grade. Among racial and ethnic groups, white students are the most likely to take the TAKS (reading=92.8%, math=92.9%), while African-American students are the least likely (reading=85.4%, math=86.3%). Limited English Proficient and special education students are much less likely to take the TAKS than their general education counterparts. Only 31.3% and 36.9% of special education students take the reading and math tests, respectively. In addition, TAKS test taking declines as grade level increases. While 91.4% of third graders take the reading test, 86.1% of eleventh graders do.

Table 1. Percent of Students Taking TAKS by Race, Ethnicity, Grade Level, and Program Status, 2004

Subgroup	Reading (Percent)	Math (Percent)
White	92.79	92.88
Asian	91.96	92.19
Hispanic	87.63	87.83
Poor	86.22	86.73
African-American	85.37	86.25
LEP	79.47	79.76
Special Education	31.25	36.85
Grade 3	91.36	91.94
Grade 4	88.71	90.00
Grade 5	87.93	89.13
Grade 6	87.50	88.28
Grade 7	86.46	87.17
Grade 8	86.90	87.26
Grade 9	84.21	83.02
Grade 10	88.12	87.55
Grade 11	86.11	86.02
Total	87.58	87.97
n	129,989	

As expected, TAKS test taking is strongly associated with students' academic performance. Figure 1 presents the percentage of students taking the TAKS test by their national percentile on the Stanford 10 test. The percentage of students taking the TAKS increases logarithmically as percentile on the Stanford exam increases.

Figure 1. The Relationship Between TAKS Test Taking and Stanford National Percentile



Do schools strategically exempt low-scoring students, and how do these exemptions vary by school?

Table 2 provides descriptive statistics for the variables used in this analysis, while Tables 3a and 3b present a series of hierarchical logistic regression models predicting English and Spanish TAKS test taking, respectively. The coefficients of interest are those on the dummy variables for our five categories of low Stanford and Aprenda scores: the 1-9, 10-19, 20-29, 30-39, and 40-49 percentiles.

Table 2. Descriptive Statistics for Variables Used in Hierarchical Logistic Regression Models

	English TAKS		Spanish TAKS		Minimum and Maximum	
Variable	n	Mean	n	Mean		
Reading Analyses						
Take Reading TAKS	117593	0.878	9412	0.914	0	1
Stanford/Aprenda National Percentile Rank (NPR) 1-9	117593	0.104	9412	0.044	0	1
Stanford/Aprenda NPR 10-19	117593	0.097	9412	0.050	0	1
Stanford/Aprenda NPR 20-29	117593	0.121	9412	0.045	0	1
Stanford/Aprenda NPR 30-39	117593	0.113	9412	0.069	0	1
Stanford/Aprenda NPR 40-49	117593	0.106	9412	0.085	0	1
Special education	117593	0.117	9412	0.058	0	1
Poor	117593	0.718	9412	0.967	0	1
Limited English Proficient	117593	0.180	9412	0.984	0	1
Hispanic	117593	0.530	—	—	0	1
African-American	117593	0.331	—	—	0	1
White	117593	0.105	—	—	0	1
Grade 3	117593	0.096	9412	0.584	0	1
Grade 4	117593	0.115	9412	0.364	0	1
Grade 5	117593	0.132	9412	0.052	0	1
Grade 6	117593	0.125	—	—	0	1
Grade 7	117593	0.123	—	—	0	1
Grade 8	117593	0.110	—	—	0	1
Grade 9	117593	0.135	—	—	0	1
Grade 10	117593	0.094	—	—	0	1
Grade 11	117593	0.071	—	—	0	1
Math Analyses						
Take Math TAKS	117635	0.882	9397	0.909	0	1
Stanford/Aprenda National Percentile Rank (NPR) 1-9	117635	0.049	9397	0.050	0	1
Stanford/Aprenda NPR 10-19	117635	0.087	9397	0.057	0	1
Stanford/Aprenda NPR 20-29	117635	0.114	9397	0.087	0	1
Stanford/Aprenda NPR 30-39	117635	0.108	9397	0.090	0	1

Note: Because fixed student characteristics (i.e. grade level and race) do not vary in a meaningful way across reading and math models, this table reports student characteristic descriptive statistics for students included in our reading models; the Stanford/Aprenda national percentile rank is reported separately for both reading and math models.

Table 3a. Logit Coefficients from Hierarchical Logistic Regression of Taking the English TAKS on Student Characteristics

	Math		Reading	
Variable	(1)	(2)	(1)	(2)
Intercept	2.176*** (0.052)	3.101*** (0.124)	2.076*** (0.051)	2.956*** (0.131)
Stanford NPR 1-9		-2.351*** (0.045)		-2.970*** (0.047)
Stanford NPR 10-19		-1.829*** (0.039)		-1.824*** (0.048)
Stanford NPR 20-29		-1.336*** (0.038)		-1.185*** (0.049)
Stanford NPR 30-39		-0.804*** (0.043)		-0.824*** (0.054)
Stanford NPR 40-49		-0.636*** (0.048)		-0.640*** (0.059)
Grade 3		1.159*** (0.117)		1.763*** (0.122)
Grade 4		1.052*** (0.115)		1.423*** (0.119)
Grade 5		1.029*** (0.114)		1.371*** (0.118)
Grade 6		0.971*** (0.113)		1.400*** (0.118)
Grade 7		0.662*** (0.114)		1.006*** (0.118)
Grade 8		0.569*** (0.114)		0.926*** (0.119)
Grade 9		-0.233*** (0.052)		0.615*** (0.058)
Grade 10		0.443*** (0.058)		0.940*** (0.065)
Poor		-0.072* (0.034)		0.010 (0.037)
Limited English Proficient		-1.546*** (0.033)		-1.108*** (0.037)
Special Education		-3.277*** (0.029)		-3.585*** (0.031)
Hispanic		0.695*** (0.077)		0.629*** (0.080)
White		0.468*** (0.088)		0.396*** (0.094)
Black		0.446*** (0.080)		0.350*** (0.085)

^ p≤.10; * p≤.05; ** p≤.01; *** p≤.001

Table 3b. Logit Coefficients from Hierarchical Logistic Regression of Taking the Spanish TAKS on Student Characteristics

	Math		Reading	
	(1)	(2)	(1)	(2)
Intercept	2.340*** (0.080)	1.290*** (0.294)	2.417*** (0.083)	1.043*** (0.283)
Stanford NPR 1-9		-2.852*** (0.156)		-2.019*** (0.162)
Stanford NPR 10-19		-1.921*** (0.158)		-1.089*** (0.170)
Stanford NPR 20-29		-1.309*** (0.147)		-0.753*** (0.189)
Stanford NPR 30-39		-0.729*** (0.166)		-0.934*** (0.158)
Stanford NPR 40-49		-0.505** (0.172)		-0.453** (0.162)
Grade 3		2.603*** (0.158)		2.738*** (0.152)
Grade 4		1.788*** (0.159)		1.879*** (0.152)
Poor		0.298 (0.250)		0.172 (0.245)
Special Education		-3.304*** (0.125)		-3.041*** (0.121)

^ p≤.10; * p≤.05; ** p≤.01; *** p≤.001

Students with higher Stanford national percentile scores are more likely to take the TAKS, and these effects are non-linear. Relative to students at or above the 50th percentile, students' odds of taking the TAKS are 90% lower for math (Odds Ratio=.095) and 95% lower for reading (Odds Ratio=.051) if they score in the 1-9th percentile. While students in the 40-49th percentile are still much less likely to take the TAKS than students above the 50th percentile, their odds of taking the test are much higher (Odds Ratio for Math and Reading=.53). Even net of their lower test scores, students classified as LEP or special education have much lower odds of taking the TAKS than do general education students. This suggests that the opportunity structure for exemption matters. The models predicting Spanish TAKS test taking presented in Table 3b show a similar pattern; low-scoring students are much less likely to take the high-stakes test. Taken together, these results provide strong evidence that schools use exemption strategically.

School effects on taking the high-stakes test are substantial in size. One way of thinking about the size of school effects is to consider the impact of moving a student from a school in the 25th percentile of the school effects distribution (the distribution of school intercepts in our hierarchical model) to the 75th percentile. Net of the individual characteristics of students included in our model, we find that moving a student from a 25th to a 75th percentile school increases her probability of taking the English TAKS test by .165 (25th percentile=.420, 75th percentile=.585). School effects are of similar size for the math test; the difference in a student's probability of taking the TAKS for a school in the 25th percentile of the distribution versus the 75th is .171 (25th percentile=.422, 75th percentile=.593).

How does student exemption impact schools' passing rates and the estimated size of achievement gaps?

Our data allow us to predict the scores of students who were not tested, and then recalculate schools' passing rates to demonstrate the impact of exemption. Table 4 shows the overall results of our imputation. We provide passing rates for two different ways of measuring schools' performance. First, we display the official passing rates released by the Texas Education Agency. Using this measure, an average of 67 percent and 80 percent of students in HISD passed the math and reading tests, respectively. Second, we calculate the percent passing if all students were tested and counted. In this case, the passing rate falls to 57.5 and 69.0 percent for math and reading. In short, exemption produces an increase of 9.5 percentage points in the passing rate for math and an 11 percent increase for reading.

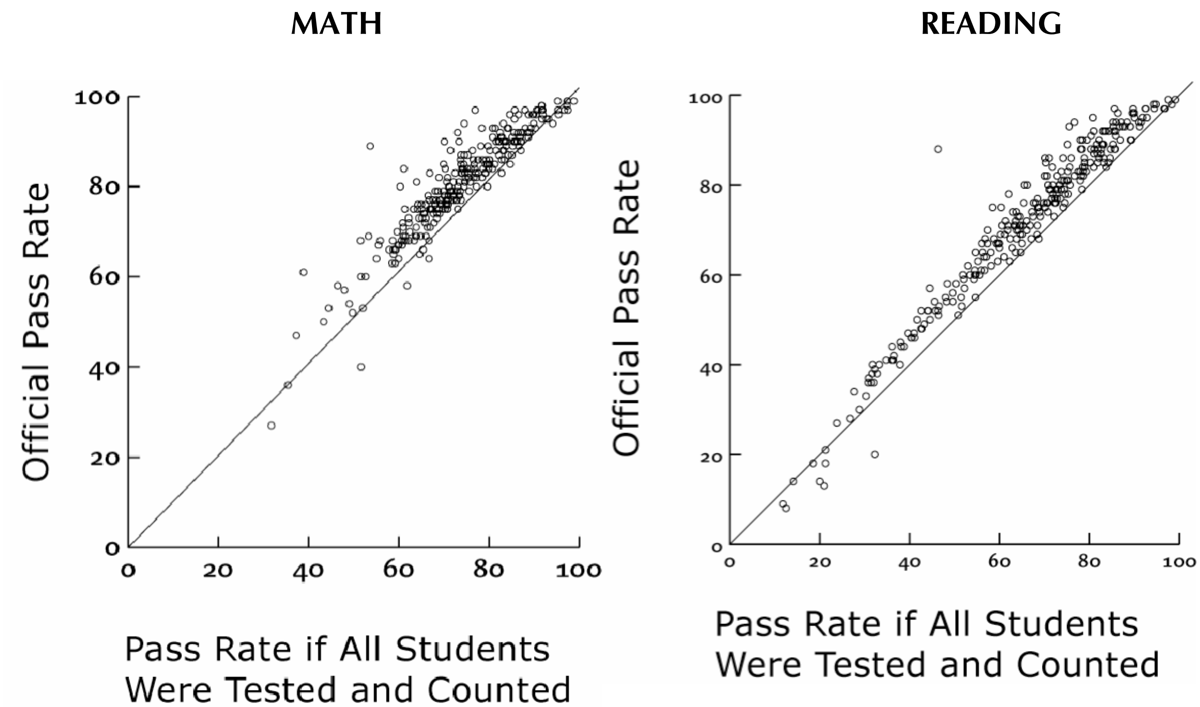
Table 4. The Impact of Exemption on District-Level Math and Reading Pass Rates by Subgroup

	All Students	African-American	Hispanic	White	Poor	Limited English Proficient	Special Education
Official Math Pass Rate	67.0	59.0	65.0	89.0	63.0	61.0	43.0
Math Pass Rate if All Students Were Tested and Counted	57.5	50.4	55.8	81.9	53.7	47.2	14.3
Official Reading Pass Rate	80.0	79.0	77.0	94.0	77.0	63.0	59.0
Reading Pass Rate if All Students Were Tested and Counted	69.0	66.5	66.6	86.5	65.3	49.8	17.0

The impact of exemption, however, is not uniform across schools. The graphs in Figure 2 plot the official passing rate against the passing rate if all tested students are counted. A school with no difference in the two passing rates would fall on the 45-degree line; data points above this line indicate that schools had a higher percentage of students passing because of this provision. Almost all schools fall above the 45-degree line, and some fall substantially above this line.

Table 4 further disaggregates the imputation results by subgroup, and shows that the effects of exemption on passing rates are not uniformly distributed across subgroups. The difference between the official pass rate for reading and the pass rate if all students were tested and counted is greatest for special education and LEP students. For math, the official special education pass rate is 28.7% higher than our imputed rate, while it is 13.8% higher for LEP, 9.3% higher for poor students, 8.6% higher for African-American students, and 9.2% higher for Hispanic students. For reading, the official special education pass rate is 42% higher than our imputed rate, while it is 13.2% higher for LEP, 12.5% higher for African-Americans, 11.7% higher for poor students, and 10.4% higher for Hispanic students.

Figure 2. The Impact of Test Exemption on Schools' Reading Pass Rates



Note: In these graphs, official reading and math pass rates are plotted against the imputed pass rate, which simulates a scenario in which all students are tested and counted. Points above the 45 degree line indicate that schools' scores were higher as a function of these exemptions.

Because of test exemption, the size of the achievement gap is underestimated. With exemption, the size of the African-American/white achievement gap is 30% for math, while the size of the Hispanic/white gap is 24%. If all students were tested and counted, these respective achievement gaps would grow to 31.5% and 26.1%. Exemption thus deflates these achievement gaps by 1.5 and 2.1 percent, respectively. These effects are larger for reading than for math. The African-American/white reading achievement gap is 5 percent larger without exemption, while the Hispanic/white gap is 2.9 percent larger. These are non-trivial rates of inflation.

Another way of thinking about the impact of these differences in passing rates is to consider what accountability rating Texas would assign each school if all students had been tested and counted. We calculate new ratings only for reading and math, and do so separately. In other words, while multiple subjects are used together to calculate ratings, we consider the rating parameter into which a school falls into based on its reading or math scores alone to analytically distinguish effects on math and reading scores. Schools in Texas can earn one of four ratings: Exemplary, Recognized, Acceptable, or Academically Unacceptable. In 2003-2004, to earn an exemplary rating, 90% of students must pass all subjects; to earn a recognized rating, 70% of students must pass all subjects, and to earn an acceptable rating, 70% of students must passing the reading test and 35% of students must pass the math test. Table 5 illustrates how schools' ratings would change if all students were tested and counted. The downward mobility that would result is significant:

- For reading, 106 schools, or 37.7% of all Houston schools, would fall into a lower rating category if all students were tested on the TAKS and counted.

- For math, 78 schools, or 27.7% of Houston schools, would fall into a lower rating category.
- For reading, 36.8% of Exemplary schools would decline to Recognized status. For math, 39.2% of Exemplary schools would decline to Recognized status.
- 15.2% of Recognized schools would decline to acceptable status for reading, while 29.4% of Recognized schools would do so for math. One Recognized school would decline to Academically Unacceptable status.
- 13.3% of Acceptable schools would decline to Academically Unacceptable status for reading, while 9.9% of Acceptable schools would do so for math.

In short, the performance of Houston schools would be dramatically lower if all students took the TAKS and were counted towards each school's scores.

Table 5. Official Accountability Ratings and Ratings if All Students Were Tested and Counted

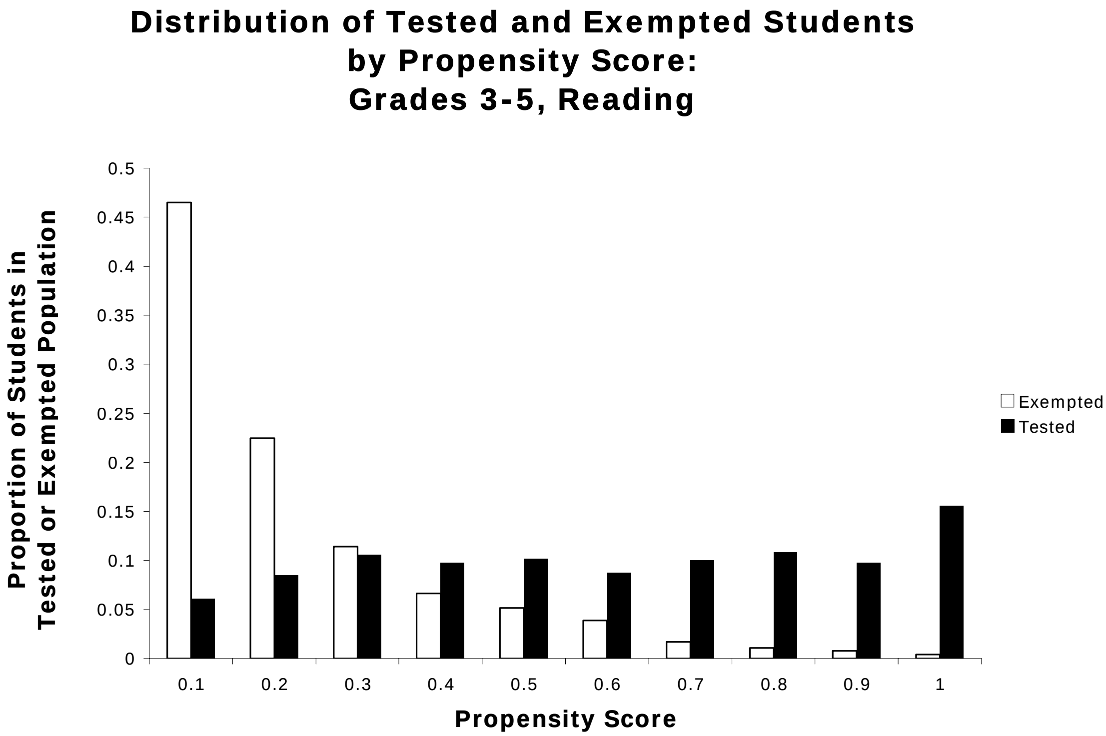
	Official Reading Accountability Rating	Rating If All Students Tested and Counted	Official Math Accountability Rating	Rating If All Students Tested and Counted
Exemplary	24.20% (68)	8.90% (25)	18.09% (51)	7.09% (20)
Recognized	58.36% (164)	53.38% (150)	44.68% (126)	42.20% (119)
Acceptable	16.01% (45)	34.16% (96)	32.27% (91)	42.20% (119)
Academically Unacceptable	1.42% (4)	3.20% (9)	4.96% (14)	8.51% (24)

What is the causal effect of testing on special education students' academic growth?

We hypothesized that students who took the high-stakes test in 2003-2004—that is, the students for whom schools and teachers were held accountable—would exhibit a greater one year gain in their national percentile rank on the Stanford Achievement Test than the same students would if they were not tested. If this is true, one possible explanation is the strategic allocation of school and teacher instructional resources. We suggest that schools and teachers spend more time and attention on those students who "count" and divert attention away from those students who do not "count," those students who are exempted. A second possibility, however, is that students' level of effort increases if they find out that they are taking the TAKS rather than the SDAA tests. Given the nature of our data, it is unclear if students in Houston know which test they are taking, and this likely varies across schools and teachers. It is also unknown how much students' effort would increase if they were taking the TAKS rather than the SDAA; in both cases, students have to sit for a multi-hour standardized exam and ultimately receive a score on the exam.

Figure 3 shows the proportion of tested and untested students across the propensity score distribution for grades 3-5. Across the entire propensity score distribution, some students were tested, while others were not. Thus, sufficient overlap existed to establish an average treatment effect.

Figure 3



Why might it be the case that across the entire distribution, some students are tested and others are not? First, teachers attending an Individual Education Plan meeting are not econometricians with perfect information about a student's prior performance on standardized tests, which introduces some level of noise into test-taking decisions. Second, schools and teachers likely vary in their exemption practices; a tested student in one school might be exempted if he attended another school. Finally, parents sit on IEP committees and have some input in this decision. We imagine that some parents may want their children to be tested on high-stakes tests, while others may not. In sum, there is some element of randomness built into these processes that produces the distribution of test-taking and exemption displayed in Figure 3.

Before moving to our treatment effect estimates, we draw attention to the observable differences in the treated and untreated populations that our matching seeks to correct. In Table A1 (see Appendix A), we display the means of treated and untreated students in grades 3-8 for all of our covariates before matching. Tested students are much higher scoring, less likely to be of limited English proficiency, less likely to be African-American or Hispanic, less likely to be poor, and more likely to attend schools with lower proportions of poor and African-American students. However, as is demonstrated by the *t* tests for differences in means between matched students displayed in the far right hand column, our matching was able to eliminate any observable differences between tested and exempted students.

Descriptive statistics for treated and untreated groups' gain scores are available in Table 6. Table 7 presents the average treatment effects of taking the TAKS test. First, we discuss our most conservative estimates, where matches are constrained such that students can only match with another student in their school. Because of the small number of cases that we are able to match this way (the number of matched cases are displayed in the "Treated on Common Support" column of

Table 7), we pool students in grades 3-8 in one analysis and then examine grades 9-11 separately. In all of these analyses, however, students can only match with a student in the same grade level. We found that students in grades 3-8 who were tested gained more academically in one year than matched students who were not tested. The average treatment effect was a gain of 5.7 national percentile points for reading and 4.3 national percentile points for math. The treatment effect of taking the reading test for students in grades 9-11 is 2.9, but this result is only statistically significant at the .10 level (t=1.913). We find no treatment effect of taking the high school TAKS math test.

Table 6. Means and Standard Deviations for 2003-2004 Gain in Stanford National Percentile by Treatment

	Untreated		Treated	
	Mean	SD	Mean	SD
Reading				
3-5	.023	11.859	-.428	14.446
3-8	.933	10.921	1.484	14.237
6-8	1.632	10.089	2.865	13.929
9-11	1.398	10.775	3.029	13.376
Math				
3-5	.690	12.900	.562	16.19
3-8	1.478	12.186	.216	15.387
6-8	2.024	11.637	-.084	14.650
9-11	5.148	16.191	2.312	17.179

We now turn to the estimates of treatment effects where we allow matching across schools and compare them with our within-school estimates. The results for Reading and Math in grades 3-8 are slightly larger than the within-school estimates (6.0 and 5.7). The results for high school reading are also larger (5.3), and there continues to be no treatment effect associated with taking the high school math TAKS. These results are similar enough to the within-school estimates to justify using this method to examine how these effects vary across grade level and across the propensity score distribution, which we would not be able to do if we restricted our matches to students in the same school.

In the bottom panel of Table 7, we re-estimate average treatment effects, but now examine three sets of grade levels separately: 3-5, 6-8, and 9-11. Treatment effects are larger for grades 3-5 than they are for grades 6-8 and 9-11, with reading treatment effects of 7.4, 6.5, and 5.1, respectively. The parallel treatment effects for math are 7.0 for grades 3-5 and 5.2 for grades 6-8.

Table 7. Average Treatment Effects of Taking the High-Stakes Test on Special Education Students' Gain in Stanford National Percentile

PANEL A

	Treated	Controls	Difference	SE	*t* statistic	Untreated	Treated	Treated On Common Support	Total
Match Only Within-school									
Reading 3-8	4.671	-1.053	5.724	1.684	3.399***	4950	2048	170	6998
Reading 9-11	4.218	1.312	2.906	1.519	1.913^	1519	1049	170	2568
Math 3-8	3.457	-0.821	4.277	1.662	2.573***	4356	2642	173	6998
Math 9-11	2.615	3.926	-1.311	2.585	-0.507	1456	1109	135	2565
Match Overall									
Reading 3-8	4.387	-1.626	6.012	0.825	7.287***	4950	2048	827	6998
Reading 9-11	3.571	-1.746	5.318	0.854	6.227***	1519	1049	800	2568
Math 3-8	2.801	-2.864	5.665	0.759	7.464***	4356	2642	1100	6998
Math 9-11	4.341	2.353	1.988	1.853	1.073	1456	1109	252	2565

PANEL B

	Treated	Controls	Difference	SE	*t* statistic	Untreated	Treated	Treated On Common Support	Total
Reading									
Grades 3-5	3.542	-3.884	7.426	1.252	5.930***	2151	859	336	3010
Grades 6-8	5.111	-1.370	6.481	1.077	6.018***	2799	1189	530	3988
Grades 9-11	3.571	-1.746	5.318	0.854	6.227***	1519	1049	800	2568
Math									
Grades 3-5	4.510	-2.519	7.030	1.345	5.227***	1783	1227	437	3010
Grades 6-8	2.294	-2.919	5.214	0.940	5.547***	2573	1415	618	3988
Grades 9-11	4.341	2.353	1.988	1.853	1.073	1456	1109	252	2565

Table 8. Heterogeneity of Treatment Effects Across the Propensity Score Distribution

	Grades 3-5	Treated	Controls	Difference	SE	*t*	Untreated	Treated	Treated On Support	Total
Grades 3-5	Reading									
	Quartile 1	8.375	3.167	5.208	2.634	1.977*	729	24	24	753
	Quartile 2	7.351	-1.088	8.439	2.591	3.257**	673	79	57	752
	Quartile 3	4.753	-2.571	7.324	2.538	2.886**	518	235	105	753
	Quartile 4	-0.409	-10.272	9.864	5.540	1.781^	231	521	22	752
	Math									
	Quartile 1	10.020	4.392	5.627	2.594	2.169*	691	62	51	753
	Quartile 2	8.400	3.453	4.947	2.437	2.030*	571	181	95	752
	Quartile 3	3.705	-2.754	6.459	2.675	2.415*	379	374	122	753
	Quartile 4	-1.250	-12.125	10.875	4.364	2.492*	142	610	56	752
Grades 6-8	Reading									
	Quartile 1	6.737	1.816	4.921	2.058	2.391*	954	42	38	997
	Quartile 2	6.574	0.819	5.755	1.85	3.111**	885	112	94	997
	Quartile 3	4.083	-0.472	4.556	1.841	2.475*	681	316	144	997
	Quartile 4	3.497	-4.222	7.719	2.566	3.008**	279	718	167	997
	Math									
	Quartile 1	8.656	6.82	1.836	1.907	0.963	927	70	61	997
	Quartile 2	5.685	2.099	3.586	1.776	2.019*	842	155	111	997
	Quartile 3	0.261	-3.087	3.348	4.53	0.739	584	413	23	997
	Quartile 4	-2.237	-7.452	5.215	3.262	1.599	220	777	93	997
Grades 9-11	Reading									
	Quartile 1	5.859	4.366	1.493	1.885	0.792	569	73	71	642
	Quartile 2	6.509	2.376	4.133	1.344	3.075**	465	177	173	642
	Quartile 3	5.434	0.139	5.294	1.378	3.842***	333	309	265	642
	Quartile 4	-0.74	-6.693	5.953	2.249	2.647**	152	490	192	642
	Math									
	Quartile 1	10.237	10.368	-0.132	4.138	-0.032	565	77	38	642
	Quartile 2	4.88	3.205	1.675	3.128	0.535	457	184	83	641
	Quartile 3	2	1.671	0.329	3.774	0.087	305	336	73	641
	Quartile 4	2.655	-3.259	5.914	4.293	1.378	129	512	58	641

We were interested not only in the average treatment effect, but also in how this effect varied across the propensity score distribution. To examine this heterogeneity, we divided students into four strata based on their propensity to take the test and estimated average treatment effects within each of these strata. Results from this analysis can be found in Table 8. We found larger treatment effects for students with the highest propensity to be tested. For students in grades 3-5, those who are least likely to take the math TAKS have a treatment effect of 5.6, while those who are most likely to take the TAKS have a treatment effect of 10.9. This pattern holds for the grades 3-5, 6-8, and 9-11 reading tests, but not for math tests in grades 6-8 and 9-11.

Again, there are two potential explanations for this finding. Students' with higher propensities to be tested may be potential "conversions" from failing to passing students, and thus could improve the school's passing rate. That the treatment effects are largest for students in the strata that offer schools the largest investment/return ratio suggests that the mechanism explaining the treatment effect may be the diversion of instructional resources. A second possibility often discussed in the propensity score matching literature is that students may differ in the benefits that they can derive from a given intervention. It is possible that students in the top quartile of the propensity score distribution differentially benefit from this treatment. However, we cannot rule out either explanation here, and identifying mechanisms explaining the treatment effect of testing is an important area for future research.

Finally, in order to test the robustness of these results, we simulated the impact of two potential omitted variables in order to examine their impact on our treatment effect estimates. The goal of this exercise is to understand how large the effect of an omitted variable would have to be to eliminate the treatment effects that we observe. Here, we focus on our most conservative estimates for grades 3-8 (our within school estimates) as we did not find statistically significant treatment effects using within school matching estimates for grades 9-11. We first identify two variables that have large effects on test taking as well as effects on students' academic growth: placement in special education for at least two years and being African-American. In our model predicting taking the TAKS reading for grades 3-8, the odds ratios on these variables are .80 and .45 for reading and .81 and .59 for math, respectively. In two separate simulations, we generated an unobserved variable *U* with effects similar to these variables and re-estimate our treatment effects 100 times in the presence of this variable. In grades 3-8, an omitted variable would have to have an effect 34 times the size of being black and 42 times the size of being in special education for two years in order to wipe out the presence of a treatment effect of testing. For math in grades 3-8, an omitted variable would have to be 65 times the size of being in special education for two years in order to make these estimates statistically insignificant; an omitted variable mimicking being black actually slightly increases our estimates of math treatment effects because these variables decrease students' odds of being tested, but do not also have a large effect on their year-to-year academic gain. Given that the probability of such a variable existing is incredibly small, we conclude that our results are robust to the presence of an omitted variable.

| 5 | Discussion

To date, much of the participation debate has been grounded in rhetorical rather than empirical claims. As far as we are aware, this is the first study to estimate the impact of exemption on schools' passing rates and the size of achievement gaps. It is also the first to address the question of whether, in the context of a high-stakes accountability system, tested students exhibit more academic growth when they are tested than they would if they were not.

To recapitulate our key findings: first, we found that HISD schools took advantage of the exemption opportunities provided by the state to exempt low-scoring students, and, in doing so, increased their passing rates by 9.5 percent for math and 11 percent for reading. Thirty-eight percent of schools would drop to a lower accountability rating for reading in the absence of these exemptions, while 28% would do so for math.

Second, we established that exemption leads to the underestimation of achievement gaps.[4] Because Hispanic and African-American students are more likely to be classified in exemptable categories (LEP and special education) than are their white counterparts, these students are more likely to be exempted. Even within special education students, approximately half of white special education students take the TAKS, while only one-third of African-American and Hispanic students do. (See Appendix Tables A2 and A3.) As a result, White/Hispanic and White/African-American achievement gaps are larger than they appear to be when students are exempted.

Finally, we determined that special education students gain more academically when they are tested than the same students would have if they were not. These treatment effects are larger for students in grades 3-5 than they are for students in grades 6-8. Using our most conservative estimates (matching only within-school), the effect of being tested in reading is approximately .40 standard deviations in reading and .28 standard deviations in math for grades 3-8.[5] By any standard, these effects are large. However, using these within-school matching estimates, we found no treatment effect of testing for reading and math in grades 9-11.

Because of the two features discussed above—1) special education students are more likely to be minority and poor students than general education students, and 2) among special education students, students of color are less likely to take the TAKS—the exemption of special education students from state tests and the concomitant academic growth penalty contributes to the growth of black-white, Hispanic-white, and high-low socioeconomic status achievement gaps.

These results have substantial implications for current education policy. One key question at issue in high-stakes testing policies is whether special education students should take high-stakes tests. In this study, we have demonstrated that in the context of a high-stakes system, tested special education students in grades 3-8 exhibit more academic growth when they are tested than they would have if they were not. By the time students reach high school, however, there appear to be diminishing returns of testing special education students.

References

Brooks-Gunn, J., Klebanov, P.K., & Duncan, G.J. (1995). Ethnic differences in children's intelligence test scores: Role of economic deprivation, home environment, and maternal characteristics. *Child Development* 67(2): 396-408.

Cook, M., & Evans, W. Families or schools? Explaining the convergence in white and black academic performance. *Journal of Labor Economics* 18:4, 729-754.

Cullen, J.B. & Rebeck, R. (2006). Tinkering Towards Accolades: School Gaming Under a Performance Accountability System. National Bureau of Economic Research Working Paper 12286.

Deere, D. & Strayer, W. (2001). Putting schools to the test: school accountability, incentives, and behavior. Texas A&M Working Paper.

Figlio, D.N. (2005). Testing, crime, and punishment. National Bureau of Economic Research Working Paper 11194.

Figlio, D.N., & Getzler, L.S. (2002). Accountability, ability and disability: Gaming the system. National Bureau of Economic Research Working Paper 9307. Retrieved July 7, 2004 from http://www.nber.org/papers/w9307.

Gillborn, D., & Youdell, D. (2000). Rationing education: Policy, practice, reform, and equity. Milton Keynes: Open University Press.

Fryer, R., & Levitt, S.D. (2004). Understanding the black-white test score gap. *Review of Economics and Statistics*. 86(2): 447-464.

Haney, W. (2000). The myth of the Texas miracle in education. *Education Policy Analysis Archives, 8* (41) [Online]. Retrieved January 9, 2003 from: http://epaa.asu.edu/epaa/v8n41.

Holland, P.W. (1986). Statistics and causal inference. *Journal of the American Statistical Association* 81: 945-970.

Jacob, B.A. (2005). Accountability, incentives, and behavior: Evidence from school reform in Chicago. *Journal of Public Economics* 89(5-6): 761-796.

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

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

McNeil, L.M. (2005). Faking equity: High-stakes testing and the education of Latino youth. In A. Valenzuela (ed.), *Leaving children behind: How "Texas-style" accountability fails Latino youth* (pp. 57-112). Albany: SUNY Press.

McNeil, L.M. & Valenzuela, A. (2001) The harmful impact of TAAS testing in Texas: Beneath the accountability rhetoric. In G. Orfield & M.L. Kornhaber, *Raising Standards or Raising Barriers? Inequality and High-Stakes Testing in Public Education* (pp. 127-150). New York: Century Foundation.

Phillips, M., Crouse, J., & Ralph, J. (1998). Does the black-white test score gap widen after children enter school? In C. Jencks and M. Phillips (eds.), *The Black-White Test Score Gap* (pp. 229-272). Washington, DC: The Brookings Institution.

Reback, R. (2006). Teaching to the rating: School accountability and the distribution of student achievement. Barnard College Working Paper.

Rosenbaum, P.R., and Rubin, D.B. (1983a). The central role of the propensity score in observational studies for causal effects. *Biometrika* 70: 41-55.

Rosenbaum, P.R., and Rubin, D.B. (1983b). Assessing sensitivity to an unobserved covariate in an observational study with a binary outcome. *Journal of the Royal Statistical Society* 45: 212-218

Rosenbaum, P.R., and Rubin, D.B. (1984). Reducing bias in observational studies using subclassification on the propensity score. *Journal of the American Statistical Association* 79: 516-524.

Rosenbaum, P.R., and Rubin, D.B. (1985). Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *The American Statistician* 39: 33-38.

Rosenbaum, P.R. (2002). *Observational studies*. New York: Springer.

Rubin, D. B. (1977). Assignment to treatment group on the basis of a covariate. *Journal of Educational Statistics* 2: 1-26.

Rubin, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys*, New York: John Wiley & Sons.

United States Department of Education (2007). "Raising Achievement: Alternate Assessments for Students with Disabilities." Accessed April 1, 2007 at http://www.ed.gov/policy/elsec/guid/raising/alt-assess-long.html

Vasquez Heilig, J. (2006). Progress and Learning of Urban Minority Students in an Environment of Accountability. Unpublished doctoral dissertation, Stanford University.

Table A1. Comparison of Treated and Control Groups Before and After Matching for Within-school Matches, Grades 3-8

Variable	Sample	Mean Treated	Mean Control	%bias	%reduct bias	t-test t	t-test p>t
Stanford reading (*z*)	Unmatched	.80277	-.32921	124.5		50.24	0.000
	Matched	.00027	.012	-1.3	99.0	-0.16	0.876
Stanford math (*z*)	Unmatched	.62307	-.25475	93.4		36.42	0.000
	Matched	.11328	.05895	5.8	93.8	0.54	0.590
Limited English Proficient	Unmatched	.11328	.24828	-35.6		-12.79	0.000
	Matched	.1	.1	0.0	100.0	-0.00	1.000
Mobile student	Unmatched	.11719	.15434	-10.9		-4.04	0.000
	Matched	.11765	.14706	-8.6	20.8	-0.80	0.425
Poor student	Unmatched	.72363	.85879	-33.7		-13.54	0.000
	Matched	.8	.81176	-2.9	91.3	-0.27	0.785
African-American	Unmatched	.36328	.47333	-22.4		-8.48	0.000
	Matched	.58824	.59412	-1.2	94.7	-0.11	0.912
Hispanic	Unmatched	.44092	.46485	-4.8		-1.83	0.068
	Matched	.33529	.32941	1.2	75.4	0.11	0.909
Asian	Unmatched	.01221	.00808	4.1		1.64	0.102
	Matched	.00588	.00588	0.0	100.0	0.00	1.000
New to district in previous school year	Unmatched	.07373	.10061	-9.5		-3.53	0.000
	Matched	.07059	.06471	2.1	78.1	0.22	0.830
In special education for at least 2 years	Unmatched	.79688	.89939	-28.9		-11.69	0.000
	Matched	.85294	.84118	3.3	88.5	0.30	0.764
School attendance	Unmatched	95.929	95.74	12.5		4.71	0.000
	Matched	95.487	95.487	0.0	100.0	0.00	1.000
Size of mobility subset	Unmatched	7.4506	8.7888	-36.8		-14.04	0.000
	Matched	9.0306	9.0306	0.0	100.0	0.00	1.000
Per-pupil expenditures	Unmatched	5927.3	6054.5	-7.8		-3.23	0.001
	Matched	5941.9	5941.9	0.0	100.0	0.00	1.000
Percent minority staff	Unmatched	63.074	71.504	-40.9		-15.92	0.000
	Matched	73.848	73.848	0.0	100.0	0.00	1.000
Percent new teachers	Unmatched	8.9109	9.0471	-2.9		-1.09	0.275
	Matched	8.4582	8.4582	0.0	100.0	-0.00	1.000
Salary new teachers/100	Unmatched	339.55	341.2	-5.0		-1.92	0.055
	Matched	339.68	339.68	0.0	100.0	0.00	1.000
Percent of teachers with 1-5 years experience	Unmatched	33.689	34.085	-3.9		-1.47	0.141
	Matched	35.538	35.538	0.0	100.0	0.00	1.000
Salary 1-5 years experience/100	Unmatched	367.37	366.86	5.4		2.09	0.037
	Matched	367.03	367.03	0.0	100.0	0.00	1.000
Percent of teachers with 6-10 years experience	Unmatched	15.595	15.325	4.5		1.70	0.090
	Matched	15.811	15.811	0.0	100.0	-0.00	1.000
Salary 6-10 years experience/100	Unmatched	394.74	394.11	9.3		3.65	0.000
	Matched	393.39	393.39	0.0	100.0	0.00	1.000
Percent of teachers with11-20 years	Unmatched	19.881	19.926	-0.6		-0.23	0.819
	Matched	18.586	18.586	0.0	100.0	0.00	1.000

experience							
Salary 11-20 years experience/100	Unmatched	463.51	462.6	7.2		2.74	0.006
	Matched	462.65	462.65	0.0	100.0	0.00	1.000
Percent of teachers with >20 years experience	Unmatched	21.918	21.609	3.5		1.35	0.176
	Matched	21.595	21.595	0.0	100.0	-0.00	1.000
Salary of teachers with >20 years experience	Unmatched	560.1	561.28	-7.2		-2.73	0.006
	Matched	562.02	562.02	0.0	100.0	0.00	1.000
Percent African-American teachers	Unmatched	41.637	49.772	-30.0		-11.28	0.000
	Matched	58.145	58.145	0.0	100.0	0.00	1.000
Average experience of teachers	Unmatched	11.243	11.157	3.4		1.31	0.189
	Matched	11.044	11.044	0.0	100.0	-0.00	1.000
Percent female teachers	Unmatched	75.691	74.867	7.2		2.72	0.006
	Matched	76.052	76.052	0.0	100.0	0.00	1.000
Percent Hispanic teachers	Unmatched	15.457	16.179	-4.8		-1.80	0.072
	Matched	10.986	10.986	0.0	100.0	-0.00	1.000
Average years of HISD teaching experience	Unmatched	9.139	9.217	-3.2		-1.22	0.222
	Matched	9.0118	9.0118	0.0	100.0	-0.00	1.000
Percent mobile students	Unmatched	19.669	21.541	-23.2		-9.32	0.000
	Matched	21.776	21.776	0.0	100.0	0.00	1.000
Ln(student enrollment)	Unmatched	6.7591	6.6819	18.3		6.91	0.000
	Matched	6.6849	6.6849	0.0	100.0	0.00	1.000
Percent bilingual students	Unmatched	20.864	22.713	-10.4		-3.92	0.000
	Matched	17.119	17.119	0.0	100.0	-0.00	1.000
Percent African-American students	Unmatched	30.553	36.949	-22.1		-8.23	0.000
	Matched	45.395	45.395	0.0	100.0	0.00	1.000
Percent poor students	Unmatched	80.083	88.775	-43.4		-17.97	0.000
	Matched	86.689	86.689	0.0	100.0	-0.00	1.000
Percent gifted students	Unmatched	11.398	7.3414	37.3		15.36	0.000
	Matched	8.2429	8.2429	0.0	100.0	0.00	1.000
Percent gifted students	Unmatched	54.863	55.459	-2.0		-0.76	0.446
	Matched	45.708	45.708	0.0	100.0	0.00	1.000
Percent LEP students	Unmatched	22.266	24.113	-10.0		-3.75	0.000
	Matched	18.174	18.174	0.0	100.0	-0.00	1.000
Percent Asian students	Unmatched	3.0126	1.9234	28.7		11.48	0.000
	Matched	1.8671	1.8671	0.0	100.0	0.00	1.000
Percent special education students	Unmatched	11.686	12.78	-17.2		-6.61	0.000
	Matched	14.373	14.373	0.0	100.0	-0.00	1.000
School passing rate for math, 02-03	Unmatched	71.419	68.433	18.1		6.91	0.000
	Matched	66.215	66.215	0.0	100.0	-0.00	1.000
School passing rate for reading, 02-03	Unmatched	82.772	80.362	29.0		11.30	0.000
	Matched	80.445	80.445	0.0	100.0	0.00	1.000
School passing rate for math for poor students, 02-03	Unmatched	68.83	67.127	10.6		4.05	0.000
	Matched	64.587	64.587	0.0	100.0	0.00	1.000
School passing rate for reading for poor students, 02-03	Unmatched	80.781	79.142	21.6		8.33	0.000
	Matched	78.926	78.926	0.0	100.0	0.00	1.000

School passing rate for math for girls, 02-03	Unmatched	70.924	68.203	16.4		6.28	0.000
	Matched	65.662	65.662	0.0	100.0	-0.00	1.000
School passing rate for reading for girls, 02-03	Unmatched	84.518	82.223	28.9		11.27	0.000
	Matched	82.065	82.065	0.0	100.0	0.00	1.000
School passing rate for math for boys, 02-03	Unmatched	71.875	68.639	19.2		7.34	0.000
	Matched	66.732	66.732	0.0	100.0	-0.00	1.000
School passing rate for reading for girls, 02-03	Unmatched	80.952	78.429	27.2		10.50	0.000
	Matched	78.706	78.706	0.0	100.0	0.00	1.000
Grade 3	Unmatched	.13867	.09899	12.3		4.82	0.000
	Matched	.10588	.10588	0.0	100.0	-0.00	1.000
Grade 4	Unmatched	.13037	.15838	-8.0		-2.99	0.003
	Matched	.11765	.11765	0.0	100.0	0.00	1.000
Grade 5	Unmatched	.15039	.17717	-7.2		-2.72	0.007
	Matched	.16471	.16471	0.0	100.0	-0.00	1.000
Grade 6	Unmatched	.17578	.21152	-9.1		-3.40	0.001
	Matched	.22353	.22353	0.0	100.0	-0.00	1.000
Grade 7	Unmatched	.22412	.19273	7.7		2.98	0.003
	Matched	.25294	.25294	0.0	100.0	0.00	1.000
Grade 8	Unmatched	.18066	.16121	5.2		1.99	0.047
	Matched	.13529	.13529	0.0	100.0	-0.00	1.000

Table A2. Percent of Students in Special Education by Race and Grade Level

	Grades 3-5	Grades 6-8	Grades 9-11
Asian	3.38	3.04	1.91
African-American	14.82	17.48	16.73
Hispanic	11.43	9.46	9.43
White	11.82	10.12	8.05
n	40,445	42,209	35,898

Table A3. Percent of Special Education Students Taking the TAKS, by Race and Grade

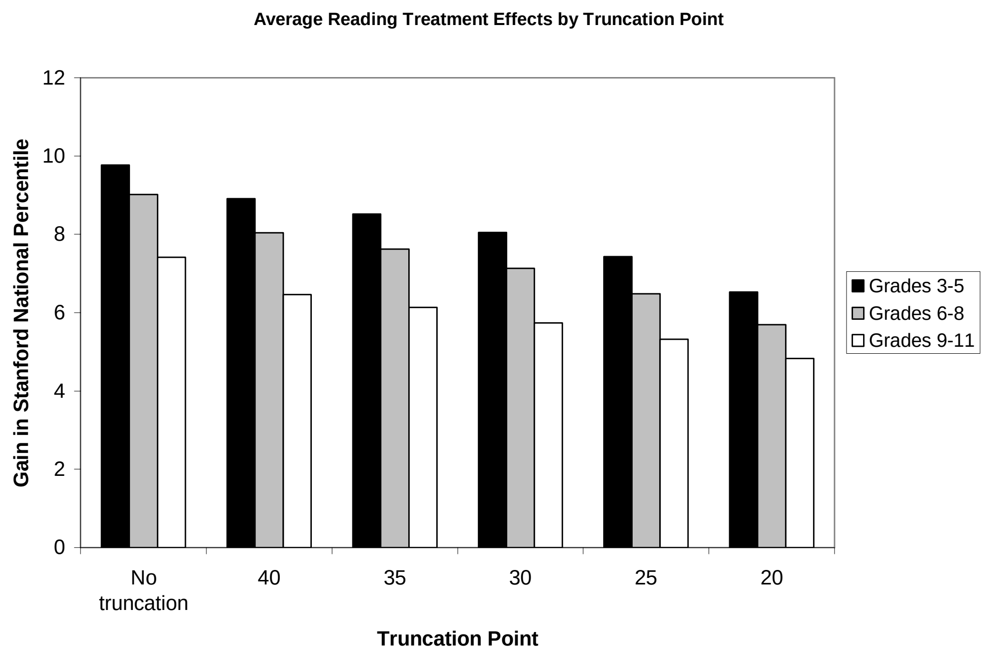
Panel A: Reading

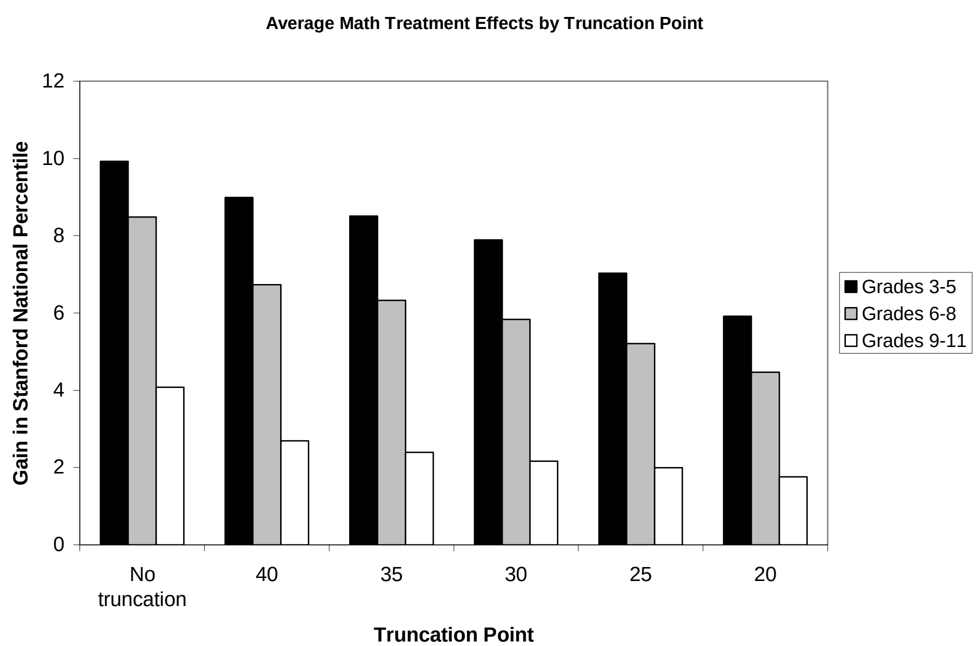
	Grades 3-5	Grades 6-8	Grades 9-11
Asian	31.91	50.00	53.85
African-American	22.63	22.71	30.55
Hispanic	24.68	28.84	41.98
White	52.17	56.82	57.31
n	5,027	4,998	4,043

Panel B: Math

	Grades 3-5	Grades 6-8	Grades 9-11
Asian	44.68	57.89	57.69
African-American	30.19	27.64	37.89
Hispanic	37.37	34.58	40.14
White	62.45	57.83	53.22
n	5,027	4,998	4,043

Appendix B: Average Treatment Effects by Truncation





Note: In our analyses, we truncate students' year-to-year academic gain at ±25 Stanford national percentile points. These graphs show how our average treatment effects change when we use different truncation points. All results are statistically significant with the exception of ninth grade math, which does not achieve significance irrespective of the truncation point that we use.

[1] The Houston Independent School District adopted the Stanford 10 exam in 2003-2004 school year and administered the Stanford 9 exam in the 2002-2003 school year. We used percentile equation tables provided to us by Harcourt Assessment to equate students' national percentile rank across years. The Aprenda test form and national norms remained constant over this time period.

[2] Because matched students are no longer independent of each other and matching with replacement allows students to be used as matches multiple times, standard errors may be downwardly biased if they are not adjusted. To address this issue, we bootstrapped the standard errors of our average treatment effects, using 1000 repetitions for each treatment effect to generate standard error estimates from the empirical sampling distribution.

[3] We took a second step in order to err on the side of conservative treatment effect estimates after examining the distribution of year-to-year change scores for tested and exempted students. As multiple scholars of testing have written, test scores can be highly volatile from year-to-year, as each observed score has two components: a "true" unobserved score and measurement error. If tested and exempted students had identical levels of volatility, this issue would not influence our measurement of testing effects. However, tested students have higher levels of volatility in the positive direction than do untested students. Since propensity score matching essentially tests for differences in means between matched cases, disproportionate positive outlying observations in one group have the potential to substantially inflate the size of treatment effects. To address this issue, we truncated change at ± 25 percentile points. That is, a student with an observed gain of 30 percentile points is coded as having a gain of 25. We have conducted these analyses with no truncation and with truncation at 40, 35, 30, 25, and 20 percentile points; please see Appendix B for estimates of treatment effects at each of these truncation points.

[4] We note that differences in rates of passing are imperfect measures of the achievement gap, as the achievement gap as measured by passing rates could be closing even as the continuous version of the achievement gap (e.g. based on students' scale scores) is increasing. Nonetheless, we employ this metric here as current policy debates are framed in terms of passing rates.

[5] These estimates standardize our treatment effect estimates using the standard deviation of the gain score of the treated population, (5.724/14.24 for reading and 4.277/15.39 for math). If we instead use the standard deviation of whole population, the effect sizes for grades 3-8 are .48 for reading (5.724/11.99) and .32 for math (4.277/13.5).

ISERP Working Papers

2007

07-05: "Differential Effects of Graduating During a Recession across Race and Gender," Ayako Kondo, Economics, Graduate Fellow, ISERP, Columbia University

07-04: "PowerPoint Demonstrations: Digital Technologies of Persuasion," David Stark, Sociology, Columbia University, Verena Paravel, Center on Organizational Innovation, ISERP, Columbia University

07-03: "No Entiendo: The Effects of Bilingualism on Hispanic Earnings," Jeronimo Cortina, Political Science, Columbia University, Rodolfo de la Garza, Political Science and International Affairs and Public Affairs, Columbia University, Pablo Pinto, Political Science, Columbia University

07-02: "The Assessment of Poverty and Inequality through Parametric Estimation of Lorenz Curves," Camelia Minoiu, Economics, Columbia University, Sanjay Reddy, Barnard Economics

07-01: "Implementing Second-Best Environmental Policy under Adverse Selection," Glenn Sheriff, School of International and Public Affairs, Columbia University

2006

06-01: "The Impact of Parental Marital Disruption on Children's Performance in School," Christopher Weiss, ISERP, Columbia University, Kathleen Foley, University of Pennsylvania

06-02: "The Choice of Index Number: Part I, Valuation and Evaluation," Sanjay Reddy, Barnard Economics, Benjamin Plener, Yale University

06-03: "Real Income Stagnation of Countries, 1960-2001," Sanjay Reddy, Barnard Economics, Camelia Minoiu, Economics, Columbia University

06-04: "Chinese Poverty: Assessing the Impact of Alternative Assumptions," Sanjay Reddy, Barnard Economics, Camelia Minoiu, Economics, Columbia University

06-05: "Spaghetti Politics," Paolo Parigi, Sociology, Columbia University, Peter Bearman, Sociology, Columbia University

06-06: "Attention Felons: Evaluating Project Safe Neighborhoods in Chicago," Andrew Papachristos, University of Chicago, Tracey Meares, University of Chicago, Jeffrey Fagan, Law, Columbia University

06-07: "Dynamics of Political Polarization," Delia Baladassarri, Columbia University, Peter Bearman, Columbia University

06-08: "Why do Some Countries Produce So Much More Output per Worker than Others?" Emmanuel Pikoulakis, University of Hull Business School, Camelia Minoiu, Economics, Columbia University

06-09: "Trivers-Willard at Birth and One Year: Evidence from U.S. Natality Data 1983-2001," Douglas Almond, Economics, Columbia University, Lena Edlund, Economics, Columbia University

06-10: “Forecasting House Seats from General Congressional Polls,” Robert Erikson, Political Science, Columbia University

06-11: “From Drafts to Checks: The Evolution of Correspondent Banking Networks and the Formation of the Modern U.S. Payments System, 1850-1914,” John James, Economics, University of Virginia, David Weiman, Economics, Barnard College, and History, Columbia University

2005

05-01: “Social Construction of Flows: Price Profiles Across Producers Gear to Market Context Upstream, Downstream and Cross-Stream,” Harrison White, Sociology, Columbia University

05-02: “Temporality and Intervention Effects: Trajectory Analysis of a Homeless Mental Health Program,” Mary Clare Lennon, Public Health, Columbia University, William McAllister, ISERP, Li Kuang, Public Health, Columbia University, Daniel Herman, Public Health, Columbia University

05-03: “Do Parents Help More Their Less Well-off Children?: Evidence from a Sample of Migrants to France,” François-Charles Wolff, Université de Nantés, Seymour Spilerman, Sociology, Columbia University, and Claudine Attias-Donfut, Caisse Nationale d’Assurance Vieillesse

05-04: “Politics, Public Bads, and Private Information,” Glenn Sheriff, International and Public Affairs, Columbia University

05-05: “Determinants of Justification and Indulgence,” Ran Kivetz, School of Business, Columbia University, Yuhuang Zheng, School of Business, Columbia University

05-06: “Political Competition and Policy Adoption: Market Reforms in Latin American Public Utilities,” Victoria Murillo, International and Public Affairs, Columbia University, Cecilia Martinez-Gallardo, Centro de Investigación y Docéncia Económica

05-07: “In Search of Lost Memories: Domestic Spheres and Identities in Roman Amheida, Egypt,” Anna Lucille Boozer, Anthropology, ISERP Graduate Fellow, Columbia University

05-08: “Global Links, Local Roots: Varieties of Transnationalization and Forms of Civic Integration,” David Stark, Sociology, Columbia University, Balazs Vedres, Central European University, Laszlo Bruszt, European University Institute

05-09: “Socio-Technologies of Assembly: Sense-Making and Demonstration in Rebuilding Lower Manhattan,” Monique Girard, ISERP, Columbia University, David Stark, Sociology, Columbia University

2004

04-01: “Reducing Bias in Treatment Effect Estimation in Observational Studies Suffering from Missing Data,” Jennifer Hill, International and Public Affairs, Columbia University

04-02: “Production Markets Broker Upstream to Downstream, balancing their volume and quality sensitivities to firms through an oriented market profile of signals,” Harrison C. White, Sociology, Columbia University

04-03: “Measuring Economic Disadvantage During Childhood: A Group-Based Modeling Approach,” Robert L. Wagmiller, Jr., SUNY Buffalo, Mary Clare Lennon, Public Health, Columbia University, Philip M. Alberti, Public Health, Columbia University, and J. Lawrence Aber, New York University

04-04: “Policymaking and Caseload Dynamics: Homeless Shelters,” William McAllister, ISERP, and Gordon Berlin, Columbia University

04-05: “Fresh Starts: School Form and Student Outcomes,” Christopher Weiss, ISERP, Columbia University and Peter S. Bearman, Sociology, ISERP, Columbia University

04-06: “Parental Wealth Effects On Living Standards and Asset Holdings: Results From Chile,” Florencia Torche, Sociology, Queens College, Center for the Study of Wealth and Inequality, Columbia University and Seymour Spilerman, Sociology, Center for the Study of Wealth and Inequality, Columbia University

04-07: “Routes into Networks: The Structure of English Trade in the East Indies, 1601-1833,” Emily Erikson, Sociology, ISERP, Columbia University and Peter Bearman, Sociology, ISERP, Columbia University

<u>2003</u>

03-01: “The Plasticity of Participation: Evidence From a Participatory Governance Experiment,” Shubham Chaudhuri, Economics, Columbia University, and Patrick Heller, Sociology, Brown University

03-02: “Factional Politics and Credit Networks in Revolutionary Vermont,” Henning Hillmann, Sociology, Columbia University

03-03: “ ‘Active Patients’ in Rural African Health Care: Implications for Welfare, Policy and Privatization,” Kenneth L. Leonard, Economics, Columbia University

03-04: “Living at the Edge: America’s Low-Income Children and Families,” Hsien-Hen Lu, Public Health, Columbia University, Julian Palmer, Younghwan Song, Economics, Union College, Mary Clare Lennon, Public Health, Columbia University, Lawrence Aber, Public Health, Columbia University

<u>2002</u>

02-01: “Alternative Models of Dynamics in Binary Time-Series-Cross-Section Models: The Example of State Failure,” Nathaniel Beck, Political Science, UC San Diego, David Epstein, Political Science, Columbia, Simon Jackman, Political Science, Stanford and Sharyn O’Halloran, Political Science, Columbia

02-03: “Link, Search, Interact: The Co-Evolution of NGOs and Interactive Technology,” Jonathan Bach, Center on Organizational Innovation, Columbia University and David Stark, Center on Organizational Innovation, Columbia University

02-04: “Chains of Affection: The Structure of Adolescent Romantic and Sexual Networks,” Peter Bearman, Institute for Social and Economic Research and Policy, Columbia University, James Moody, Sociology, Ohio State, Katherine Stovel, Sociology, University of Washington

02-05: “Permanently Beta: Responsive Organization in the Internet Era,” Gina Neff, Center on Organizational Innovation (COI), Columbia University, and David Stark, Center on Organizational Innovation (COI), Columbia University

02-06: “Negotiating the End of Transition: A Network Approach to Political Discourse Dynamics, Hungary 1997,” Balázs Vedres, Columbia University, Péter Csigó, Ecole des Hautes Etudes en Sciences Sociales

02-07: “The Influence of Women and Racial Minorities Under Panel Decision-Making in the U.S. Court of Appeals,” Sean Farhang, Political Science, Columbia University, Gregory Wawro, Political Science, Columbia University

02-08: “The Role of Effort Advantage in Consumer Response to Loyalty Programs: The Idiosyncratic Fit Heuristic” Ran Kivetz, Business, Columbia University, Itamar Simonson, Business, Stanford University

2001

01-01: “Pathways of Property Transformation: Enterprise Network Careers in Hungary, 1988-2000 Outline of an Analytic Strategy,” David Stark, Sociology, Columbia and Balázs Vedres, Sociology, Columbia

01-02: “Policy Space and Voting Coalitions in Congress: the Bearing of Policy on Politics, 1930-1954,” Ira Katznelson, John Lapinski, and Rose Razaghian, Political Science, Columbia

01-03: “Doing Fractions: An Analysis of Partisan ship in Post-Socialist Russia,” Andrew D. Buck, Sociology, Columbia

01-04: “Opposite-Sex Twins and Adolescent Same-Sex Attraction,” Peter Bearman, Sociology/ISERP and Hannah Brückner, Sociology, Yale

01-05: “On the Uneven Evolution of Human Know-How,” Richard R. Nelson, Business/SIPA, Columbia

01-06: “Self-Control for the Righteous: Toward a Theory of Luxury Pre-Commitment,” Ran Kivetz, Business, Columbia and Itamar Simonson, Business, Stanford

01-07: “Distributing Intelligence and Organizing Diversity in New Media Projects,” Monique Girard, ISERP, Columbia and David Stark, Sociology, Columbia

01-08: “Agricultural Biotechnology’s Complementary Intellectual Assets,” Gregory D. Graff, Agricultural and Resource Economics, Berkeley, Gordon C. Rausser, Agricultural Economics, Berkeley and Arthur A. Small, SIPA/Earth Institute, Columbia

For copies of ISERP Working Papers
Visit http://www.iserp.columbia.edu/research/working_papers/
write to iserp@columbia.edu or call 212-854-3081

Institute for Social and Economic Research and Policy
Columbia University in the City of New York
420 West 118th Street
8th Floor, Mail Code 3355
New York, NY 10027
Tel: 212-854-3081
Fax: 212-854-8925
Email: iserp@columbia.edu
www.iserp.columbia.edu

MAY 2007