NATIONAL CENTER for ANALYSIS of LONGITUDINAL DATA in EDUCATION RESEARCH

TRACKING EVERY STUDENT'S LEARNING EVERY YEAR

A program of research by the American Institutes for Research with Duke University, Northwestern University, Stanford University, University of Missouri-Columbia, University of Texas at Dallas, and University of Washington

CALDER

AIR

Do the Cheated Ever Prosper? The Long-Run Effects of Test-Score Manipulation by Teachers on Student Outcomes

Jarod Apperson

Carycruz Bueno

Tim R. Sass

WORKING PAPER 155 • March 2016

Do the Cheated Ever Prosper? The Long-Run Effects of Test-Score Manipulation by Teachers on Student Outcomes

Jarod Apperson
Georgia State University

Carycruz Bueno
Georgia State University

Tim R. Sass
Georgia State University

Contents

Acknowledgements

This research was made possible in part by generous support from the National Center for the Analysis of Longitudinal Data in Education Research (CALDER) and an anonymous foundation. Research supported by CALDER is funded through Grant R305C120008 to the American Institutes for Research from the Institute of Education Sciences, U.S. Department of Education.

We wish to thank session participants at the 2015 AEFP annual meetings and seminar participants at the University of Virginia for helpful comments.

CALDER • American Institutes for Research
1000 Thomas Jefferson Street N.W., Washington, D.C. 20007
202-403-5796 • www.caldercenter.org

Do the Cheated Ever Prosper? The Long-Run Effects of Test-Score Manipulation by Teachers on Student Outcomes

Jarod Apperson, Carycruz Bueno, Tim R. Sass

CALDER Working Paper No. 155
March 2016

Abstract

One of the concerns over high-stakes testing is the incentive for teachers to alter the scores of their students. We investigate the effects of teacher cheating on subsequent student achievement, attendance, behavior and educational attainment. We find that test scores drop below expected levels in the first year post-cheating year. These effects persist for reading and ELA, but not for math. The drop in later test scores appears to be due in part to a reduction in access to remediation services. We also find some evidence that cheated middle-school students may be more likely to drop out of high school.

I. Introduction

There are increasing calls to scale back testing of students and to lower the stakes attached to test results in the United States. One of the concerns over high-stakes testing is the incentive for teachers to alter test scores by providing answers to students before or during an exam or by correcting wrong answers after the test is taken. Indeed a number of instances of such test-score manipulation have been uncovered in recent years (Judd, 2012; Perry, et al., 2012). While extant research has developed methods for detecting test-score manipulation by teachers or administrators, little is known about how the falsification of student test scores has impacted students.

In this paper we seek to determine the long-run impacts of test-score manipulation on student outcomes. While illicit behavior by teachers and administrators (henceforth "teacher cheating") obviously boosts student test scores in the short term, it is unknown whether this cheating impacts subsequent student achievement and other outcomes such as attendance, discipline and educational attainment. To address the issue we draw upon a 10-year panel of student-level data from an anonymous urban school district where documented teacher cheating occurred. Not only do our data include reported scores on mandatory high-stakes tests and non-achievement outcomes for students, they also include individual-level data on test answer erasures on the high-stakes exams and individual item-level responses on a set of low-stakes exams. We employ the data on erasures to identify the incidence of teacher cheating and information from the low-stakes exam to control for student ability.

We find robust and persistent negative effects of being cheated on later performance on reading and English-Language Arts (ELA) exams; effects on later math achievement appear to last only one year and then fade away. The estimated achievement effects are at least as great as having a rookie teacher, rather than a teacher with five or more years of experience. In contrast, we find little or no evidence that teacher cheating has deleterious effects on student attendance or student behavior.

However, we do find some evidence that students who are cheated in middle school may be more likely than observationally equivalent non-cheated peers to drop out of high school.

The paper proceeds as follows: Section 2 is a literature review, section 3 provides background information and our data, section 4 explains our research design and econometric models, section 5 presents our results and the final section concludes.

II. Related Literature

The existing research related to test-score manipulation is quite limited and has focused on methods to identify and deter cheating. There is no extant research that specifically considers the possible impacts of altered test scores on subsequent student performance.

Jacob and Levitt (2003a, 2003b) use data from the Chicago Public Schools to develop two types of cheating indicators. One is based on unexpected test-score fluctuations, where students experience large increases in test scores in one year followed by modest increases or even declines in the following year. The other approach relies on identifying blocks of identical answers within a classroom, unusual correlations of responses across students in a classroom or unlikely patterns of answers within an individual student's exam. van der Linden and Jeon (2012) offer improvements on the Jacob and Levitt approach, exploiting information about the number and patterns of erasures. A recent book by Kingston and Clark (2014) summarizes the current research on cheating detection.

Bertoni, Brunello and Rocco (2013) consider the effect of external monitors on student exam scores. They exploit a policy experiment whereby external monitors were assigned to administer an exam to second and fifth grade students in Italy and collect the results. The external examiners were randomly assigned to schools and to classes within schools. The authors find that the presence of an external monitor lowers the proportion of correct answers by 5.5 percent. About 80 percent of the reduction in test scores is attributable to having an external monitor present in a classroom; the remaining 20 percent is due to the indirect effect of having a monitor in another classroom within the same school. Consistent with the idea that cheating behavior by students or teachers in un-monitored classes should lead to more homogenous test scores, the within-classroom variance of test scores is lower in classes without an external examiner. Likewise a Herfindahl-like measure of the heterogeneity of student answers to individual questions is greater in classes with external monitors.

III. Background and Data

Our source of data is an urban school district where documented cheating occurred on a large scale. Allegations of widespread cheating in the district on state-mandated criterion-referenced exams (CREs) first became public in 2009, and in early 2010 a state agency conducted a statewide analysis of erasures on the CREs that had been administered to students in grades 1-8 in spring 2009. Classes were "flagged" based on high numbers of wrong-to-right (WTR) erasures and schools were categorized based on the proportion of flagged classrooms in the school.[1] For the district in question, over half of elementary and middle schools were in the category of greatest concern and nearly 60 percent of elementary and middle schools were identified as having 20 percent or more of their classrooms flagged.

Following the initial erasure analysis of 2009 test scores, state monitors were dispatched to oversee testing in the district in 2010. Analysis of erasures on the 2010 exam yielded a much smaller proportion of classes flagged for high rates of wrong-to-right erasures. Thus, it appears that cheating had been greatly reduced, though not entirely eliminated in 2010. Erasure analyses were subsequently conducted in each of the years 2011-2013, but they revealed a virtual absence of any unusual levels of WTR erasures.

[1] Classrooms were flagged when the number of WTR erasures was greater than three standard deviations above the state mean. An adjustment was made for class size by dividing the standard deviation by the square root of the class size. The state investigation refers to "flagged classrooms," though they were in fact groups of students who were administered a given test by a single proctor. The test score administrator was not necessarily the classroom teacher for the tested subject.

An investigation of alleged cheating in the district was launched by state authorities in the summer of 2010. The 2009 erasure analysis was used to select schools for detailed investigation, which included interviews with school personnel. Just over 60 percent of the district's elementary and middle schools received a detailed investigation. In over half of these schools educators confessed to cheating and investigators concluded that systemic misconduct occurred in over three-fourths of the schools that were investigated in detail. The investigation also revealed that cheating had been going on for some time, perhaps as far back as 2001 in some schools. Figure 1 illustrates the magnitude of cheating in the district; students in flagged classrooms within investigated schools in 2009 had substantially higher normalized scores (on the order of 0.5 standard deviations) than did students in non-flagged classes in 2009 and 2010 or students in formerly investigated schools in 2010.

Using data provided by the district, we constructed a longitudinal dataset covering all students enrolled from the 2004/05 school year through the start of the 2014/2015 school year. The district files included individual-level information on enrollment, attendance, disciplinary incidents, withdrawal from school, high school diploma receipt, student demographics and program participation. In addition we obtained test results from state-mandated CREs administered annually in grades 1-8 in five different subjects.[2] Finally, we received scores from a national normed-reference test (NRT) that covered the same five subjects and was administered by the district in grades 3, 5 and 8 in each of the years 2006-2008.

The testing data include individual-level erasure data for the CRE in school years 2008/09 through 2012/13.[3] The erasure data provided to us for 2008/09 and 2009/10 only cover schools that were investigated by the state (henceforth "investigated schools").[4] Thus we have erasure data for

[2] Scores on a set of high school end-of-course tests (EOCTs) were also collected. However, the number of students with pre-cheating ability measures who later took EOCTs was too small to conduct an analysis of student achievement in high school.
[3] The criterion referenced exam was administered in grades 1-8 through the 2009/10 school year. In later years, the test was administered only in grades 3-8.
[4] Our 2009 erasure data cover all schools initially targeted for investigation, including two schools that did not

nearly two-thirds of the district's elementary and middle schools in 2008/09, about one-fifth of elementary and middle schools in 2009/10 and then complete district-wide erasure data in 2010/11-2012/13. The erasure data provided to us contain the raw number of correct answers and the number of WTR erasures.

The NRT data include not only scaled scores and national percentile scores; they also include individual item-level responses on each subject-area exam. Although the NRT was not directly used for any teacher or school accountability purposes, there was at least anecdotal evidence there may have been some manipulation of NRT scores. It is possible that some teachers who cheated on the high-stakes CRE may have altered scores on the low-stakes NRT in order to avoid discrepancies that could have signaled cheating on the CRE. As described below, we exploit the item-level responses in order to eliminate NRT scores that may have been tainted by teacher cheating.

IV. Research Design and Methods

A. Identifying Cheating

A key element in the analysis of the effects of cheating is identifying which students had their test scores manipulated. There are three types of teacher cheating possible. First, teachers with advance knowledge of the exam questions and answers could have used actual test questions in their lessons and communicated the answers prior to the exam. We refer to this type of score manipulation as "ex-ante cheating." Second, teachers could have guided students to the correct answer during the exam or given students the correct answers outright during the exam. We call this "contemporaneous cheating." Third, teachers could have corrected wrong answers after students turned in their exams. We dub this "ex-post cheating."

---

receive a detailed investigation because initial inquiries uncovered no evidence of improprieties.

Interviews conducted during the state investigation uncovered evidence that all three types of cheating occurred in at least some schools within the district. Teachers and other school personnel admitted they had employed a variety of methods to manipulate test scores. These included reviewing the test questions prior to test administration and prepping student responses (ex-ante cheating); positioning low and high-ability students next to each other and allowing students to copy answers from one another during the exam plus signaling the correct answers to students during the test (contemporaneous cheating); and filling in empty answers with correct responses or changing students answers from wrong to right after the exam (ex-post cheating).

There are three ways to identify teacher cheating after the fact. One method, developed by Jacob and Levitt (2003b), is to look for unusual patterns of item-level responses, either within a single student's answer sheet or across student answer sheets. We employ this approach to ferret out potentially tainted scores on the NRT exam. However, we do not have access to item-level responses on the CRE and thus this strategy cannot be used to identify cheated students in the main analysis of high-stakes exam scores.

A second approach employed by Jacob and Levitt is based on unusual inter-temporal changes in test scores. Teacher cheating of any sort should lead to increases in test scores. Correspondingly, once cheating stops, test scores should drop to reflect students' true achievement levels. Jacob and Levitt identified students as being cheated if they experienced large increases in test scores in one year followed by modest increases or even declines in the following year. Given that cheating allegedly occurred over several years in the district being analyzed, the run-up in test scores associated with cheating will not be observed for students whose test scores were always manipulated prior to the elimination of cheating. For example, if cheating was pervasive in a school and it began before a student entered first grade, then only manipulated scores would be observed prior to the end of cheating. Consequently, identifying teacher cheating based on unusual increases in test scores is problematic in our context. While we should observe a drop in test scores for all students who were subjected to some form of teacher cheating, using test scores drops to identifying cheating also creates problems for our analysis. Our goal is not simply to identify cheating, but rather to determine the effects of teacher cheating on both subsequent student achievement and non-achievement outcomes. Using test score drops to identify cheating would prohibit the analysis of impacts on test scores in the first post-cheating year. We therefore do not employ test score drops to identify teacher cheating.

Rather than unusual answer patterns or large inter-temporal changes in test scores, we rely on the third method of identifying cheating by teachers, counts of wrong-to-right erasures, to determine which students had their CRE scores manipulated. In the absence of cheating, erasures of any kind should be relatively infrequent. Also, if erasures are the result of student uncertainty between two possible answers we would expect wrong-to-right and right-to-wrong erasures to be about equally likely. One advantage of erasure analysis is that, in contrast to inter-temporal changes in test scores, high levels of wrong-to-right erasures would not result from students who are becoming sick, external disruptions on the test day or other random events that are unrelated to cheating.[5] The major disadvantage of erasure analysis, however, is that it can only identify ex-post cheating. To the extent that ex-ante cheating or contemporaneous cheating occurred it would tend to reduce the number of WTR erasures and lead to under-identification of cheating based on erasure counts.

One way to gauge the extent of ex-ante and contemporaneous cheating (and hence the potential degree of under-identification of cheating when WTR erasures are used to identify cheating) is to observe changes over time in initial test answers (i.e. answers given prior to erasures). If ex-ante and contemporaneous cheating occurred during the 2009 exam, then when state monitors were present in district schools during the 2010 CRE administration we would expect a reduction in the number of initially correct answers (prior to any erasures), relative to 2009. We approximate the initial number of

[5] One notable exception is cases where a student initially marks their answers in the wrong column (e.g. bubbles answer B rather A, C rather than B, etc.), realizes their mistake, and erases their answers to correct the mistake.

correct answers by subtracting the number of WTR erasures from the total number of correct answers. We refer to this measure as the "initial right."[6] The initial-right scores could have risen in 2010 in the absence of ex-ante and contemporaneous cheating if the exam simply became easier. We can distinguish between these hypotheses by taking into account how the CREs were administered. In grades 1 and 2 the exam questions and possible answers were read to students while in grades 3-8 the students read the questions and possible answers independently. Having teachers reading the questions and answers in the lower grades would make it easier for teachers to engage in contemporaneous cheating in a number of ways. One method is simply changing the inflexion of their voice when reading the correct answer.

Figure 2 provides evidence that ex-ante cheating did occur and was concentrated in grades 1 and 2. The graphs show the distribution of "initial right" scores on the math CRE by grade for elementary students (grades 1-5) for both 2009 and 2010. The sample is limited to schools that had a significant proportion of their classrooms flagged for high WTR erasure counts in both 2009 and 2010 (since we have erasure data for only such schools in those years). Thus the sample includes slightly less than 20 percent of district schools. Consistent with contemporaneous cheating being easier to implement in grades 1 and 2, we see that the number of initially correct answers fell sharply in 2010 for grades 1 and 2 whereas the test score distributions in 2009 and 2010 are roughly similar for grades 3 through 5.[7] Thus, although we cannot reject the possibility that some ex-ante or contemporaneous

[6] The actual number of initial correct answers will equal the total right after erasures, minus WTR erasures, plus right-to-wrong erasures. Unfortunately, we do not possess information on the number of right-to-wrong erasures in 2009. We assume that the number of right-to-wrong erasures, while not mean zero, is randomly distributed across students. Thus using the number correct less the number of WTR erasures will serve as a reasonable proxy for the number of initially correct answers for our purposes.

[7] The middle school distributions are based on only a few schools and thus are much less precise.

cheating occurred in higher grades,[8] we can be more confident that WTR erasure counts are a good measure of the extent of teacher cheating in grades 3 and above.[9]

We base our erasure-count measure of cheating on the distribution of WTR erasures on the Spring 2013 CRE test. More specifically, we used WTR erasure counts from Spring 2013 (when all evidence suggests that cheating no longer existed) and calculated the 95th percentile of the WTR distribution. The value of the 95th percentile was then applied to WTR erasure counts in 2009. Students whose WTR erasure count in 2009 exceeds the 95th percentile of the WTR erasure distribution in 2013 are deemed to have been cheated.[10] Of course, among cheated students, the degree of test score manipulation may vary; some students may have more answers corrected than others. To account for heterogeneous "treatment" we also estimate models that divide cheated students into two groups: those with fewer than 10 WTR erasures and those with 10 or more WTR erasures.[11]

B. Econometric Models of Student Achievement

In order to estimate the relationship between teacher cheating and later student achievement, we begin with the cumulative model of student achievement developed by Boardman and Murnane (1979) and Todd and Wolpin (2003):

$$A_{it} = A_t\left[\mathbf{X}_i(t), \mathbf{F}_i(t), \mathbf{E}_i(t), \mu_{i0}, \varepsilon_{it}\right] \quad (1)$$

---

[8] If there was any ex-ante or contemporaneous cheating in grades 3 and higher it would inflate scores of students with few WTR erasures and attenuate our estimates of the effect of being cheated.

[9] Given the apparent lack of ex-ante or ex-post cheating in grades 3 and above it might appear that the number of "initial right" answers could be used as a measure a student's true performance. However, not all questions are of equal difficulty and we do not know which questions were changed from wrong to right. Consequently, the number of "initial right" answers would not correspond to the scale score in the absence of cheating.

[10] In 2013 the 95th percentile of WTR erasures for Math, Reading and ELA were 5, 4 and 4, respectively. Therefore students with 6 or more WTR erasures on their 2009 CRE in math and 5 or more WTR erasures in reading or ELA were counted as having been cheated. Over 86% of students identified as cheated by our measure were in classrooms flagged by state investigators. If we measure cheated as the combination of being in a flagged classroom and having an unusually high level of WTR erasures, our findings do not change in any significant way.

[11] For math, the split is near the median; 50.4 percent of the cheated students had nine or fewer erasures. For reading and ELA, 64 percent and 68 percent of cheated students, respectively, had nine or fewer erasures. Placing cheated students into more than two categories was not informative, as the density of students is quite high just above the cheated threshold. For example, approximately 20 percent of cheated students have the minimum number of erasures to be defined as being cheated.

where A_{it} is the achievement level for individual student i at the end of their t^{th} year of life, $\mathbf{X}_i(t)$, $\mathbf{F}_i(t)$ and $\mathbf{E}_i(t)$ represent the entire histories of individual, family and school-based educational inputs, respectively. The term $\square_{i0}$ is a composite variable representing time-invariant characteristics an individual is endowed with at birth, and $\square_{it}$ is an idiosyncratic error.

Assuming the cumulative achievement function is linear and additively separable, family inputs do not vary over time, the marginal effects of the endowment and family inputs are equal to each other in each period and the cumulative achievement function is grade-invariant, then, following Sass, Semykina and Harris (2014), we can express student achievement level at any grade *t* as:

$$A_{it} = \sum_{h=1}^{t} \left[\boldsymbol{\alpha}_h \mathbf{X}_{i(t+1-h)} + \boldsymbol{\beta}_h \mathbf{E}_{i(t+1-h)} \right] + \omega_t \chi_i + \varepsilon_{it} \qquad (2)$$

where χ_i represents the student's fixed achievement component, reflecting innate intellectual ability and any fixed contribution to achievement of family inputs. For shorthand we will refer to this effect simply as "ability." The potentially time-varying marginal effect of ability on achievement is represented by ω_t. In addition to ability, current achievement depends on current and all prior individual time-varying characteristics, $\mathbf{X}_{it}$, and all contemporaneous and past school-based inputs, $\mathbf{E}_{it}$.

In order to make estimation of the achievement function computationally tractable, we assume that the impact on current achievement of educational inputs applied at least two periods in the past (i.e., twice-lagged, third-lagged, etc.) is negligible.[12] We also limit the set of individual characteristics to contemporaneous values since there is little variation in observable student characteristics from one year to the next. Further, we add an indicator for students whose test score in the relevant subject was manipulated in 2009, $Cheated_{2009}$. The achievement function for years after 2009 becomes:

$$A_{it} = \boldsymbol{\alpha}_t \mathbf{X}_{it} + \boldsymbol{\beta}_t \mathbf{E}_{it} + \boldsymbol{\beta}_{t-1} \mathbf{E}_{i(t-1)} + \pi_t Cheated_{i2009} + \omega_t \chi_i + \varepsilon_{it} \qquad (3)$$

[12] Jacob, Lefgren and Sims (2010) find the impact of educational inputs, particularly teacher quality, erodes quickly. They estimate a one-year persistence rate is in the range of one-fifth to one-third, suggesting that twice and greater lagged inputs would have an impact of at most 0.11 times the impact of contemporaneous inputs.

The potentially time-varying impacts of student characteristics, educational inputs, student ability, and being cheated, $\square_{it}$, $\square_{it}$, $\square_{t}$ and $\square_{t}$, are taken into account by estimating the equation separately for each year t.

Of primary interest is the estimated impact of being cheated on subsequent achievement, $\square_{t}$. This estimated effect is subject to potential bias resulting from the standard selection problem. Students are not randomly assigned to "treatment" (having their test scores manipulated). If the factors that determine whether a student is cheated also affect student outcomes and those factors are not fully accounted for in our model, estimates of the effect of being cheated on subsequent student outcomes will be biased.

One obvious selection mechanism is that students are not randomly assigned to schools. Schools with high proportions of generally low-performing students will face greater pressure to improve student test scores and thus be more likely to cheat. Likewise, schools with poor leadership or low average faculty quality may find it more difficult to improve test scores by increasing true student achievement and be more inclined to manipulate the test scores of their students. Many schools in the district were not investigated by the state because there were no classrooms flagged for having high levels of WTR erasures within the school or there was only an isolated flagged classroom with no corroborating evidence of teacher cheating. As illustrated in Table 1, schools that were investigated due to there being classrooms with unusually high WTR erasure counts served very different student bodies, on average, than did non-investigated schools. Relative to non-investigated schools, investigated schools served a higher proportion of minority students, fewer gifted students and a larger fraction of students from low-income families (as measured by free and reduced-price eligibility). While there were a number of non-investigated schools serving student populations with similar demographics to those of investigated schools, the fact that no cheating was uncovered in these schools may be indicative of differences in unmeasured characteristics like school leadership, school culture, or average teacher quality. We therefore restrict our analysis of the effects of teacher cheating to students in investigated schools.[13] Even among investigated schools, schools may differ along unobserved dimensions that affect both the likelihood of teacher cheating and student performance. To guard against any bias associated with school assignments among investigated schools and to allow for differences in tests across grades we compare outcomes among students attending the same school and grade in 2009 by employing school-by-grade fixed effects.

Even within a school and grade, teachers may differ in their ability. If relatively ineffective teachers are more likely to cheat and teacher effectiveness is not fully taken into account, this would impart a negative bias on the relationship between a student being cheated and their subsequent performance. To combat such possible bias, we also estimate models that compare students within the same classroom by incorporating classroom fixed effects. Within-classroom comparisons may be less desirable, however, if teachers consciously select who to cheat.[14]

The extent of teacher cheating can vary across students within classrooms, i.e. the impact of a cheating teacher may not be the same for all students in the class. For example, given the time costs involved, cheating teachers could choose to only correct answers for their weaker students, who would likely have more initial wrong answers and would thus produce the greatest gains in test scores when their answers are corrected.[15] Similarly, even if a teacher reviewed the answers of all students, more able students would have fewer wrong answers to begin with and thus any manipulation of answers ex-post would have a smaller impact on the student's score. If the students selected by a teacher for

---

[13] Students who were in classrooms flagged for high levels of WTR erasures in 2010 are also eliminated from the analysis sample.

[14] Another downside of the classroom-fixed-effects approach is that it would not capture classroom-wide effects of teacher cheating. For example, if a teacher reduces her instructional effort in anticipation of being able to alter low test scores after the fact, all students may experience a reduction in achievement, whether or not their own exam is manipulated. We have estimated models that include an indicator for classrooms that were flagged by state investigators for high levels of WTR erasures in all of our analyses. We fail to find any negative effects of being in a flagged classroom, but not having been cheated.

[15] Analogously, there is evidence that when faced with accountability pressure, teachers devote more effort toward teaching students near the proficiency threshold used for accountability. See Neal and Schanzenbach (2010).

cheating were weaker students to begin with, then even a within-classroom analysis which failed to account for student ability could yield biased estimates of the effect of teacher cheating on subsequent student outcomes.

We find evidence that is consistent with selective cheating within classrooms. Table 2 shows the mean and standard deviation of WTR erasures by quintile of student ability (as measured by prior scores on the low-stakes NRT exam – more on this below) in the year of documented widespread cheating, 2009, and a year in which cheating did not occur, 2013, for Math, Reading and ELA. Across all ability quintiles there are obviously more WTR erasures in the cheating year (2009) than in the non-cheating year (2013). However, in 2009 we see that the most able students (fifth quintile) have fewer WTR erasures and the variation in the number of WTR erasures is smaller than for students in the lowest quintile of ability. In the non-cheating environment (2013), the patterns are quite different; there is no clear trend in the number of WTR erasures. For reading and ELA, the number of WTR erasures is not significantly different across ability quintiles. In math, the number of WTR erasures in 2013 does vary across student ability quintiles, though the magnitudes of the differences are quite small and it is the highest ability students who have the greatest number of erasures.

If teachers and administrators choose which exams to alter based on perceived student ability, the challenge imposed by potential selective cheating can be addressed by appropriately controlling for student ability. If cheating on the 2009 CRE were an isolated incident, we could rely on CRE scores in prior years as a measure of student ability. However, the district experienced widespread, long-term cheating which renders pre-2009 CRE results unreliable.[16] Fortunately, as described above, we also have access to scores on a low-stakes NRT that was administered throughout the district for three years prior to the year in which cheating was measured, 2006-2008. NRT exams were administered in grades 3, 5 and 8 in each of five subjects (math, reading, ELA, science and social studies). The fact that the NRT

[16] The fact that pre-2009 CRE scores cannot be trusted also precludes the use of student fixed effects models to control for time-invariant student characteristics, including innate ability.

was not used for accountability means it is much less likely that NRT scores would be tainted by teacher cheating. However, there exists the possibility that some teachers may have manipulated scores on the NRT in order to avoid test-score discrepancies between the NRT and CRE that could have signaled cheating on the CRE. Using the methodology developed by Jacob and Levitt (2003a), we utilize individual-level item-response information for the NRT to identify and eliminate any potentially tainted NRT scores.[17] In order to maximize sample size and increase precision, we compute the five-subject mean national percentile for each student, using all non-missing scores.[18] Since the NRT was only administered in grades 3, 5 and 8, our ability measure includes students in grades 4-11 in 2008/09.

In addition to using the cross-subject mean national percentile, we also control for a number of fixed or nearly time-invariant observable student/family characteristics in the **X** vector that may be correlated with student "ability." These include gender, race/ethnicity, free/reduced-price lunch status (a proxy for family income), gifted status, limited English proficiency and indicators for a variety of disability categories.

Lastly, there is the issue of how to account for educational inputs, $\mathbf{E}_{it}$ and $\mathbf{E}_{i(t-1)}$ in equation (3). If students are assigned to classrooms randomly (conditional on observable student characteristics and prior performance on nationally normed exams) then explicit measures of educational inputs could be omitted and the effects of peer and teacher quality would be subsumed into the error term, ε_{it}. If prior cheating has a causal impact on future inputs, then omission of explicit controls for inputs would

---

[17] Jacob and Levitt develop four measures of cheating. However, the first three all involve the use of prior and future test scores, which renders them impractical in our context since we only have NRT scores for three non-contiguous grade levels (3, 5 and 8) over three years. We therefore employ their last method, which involves comparing a student's answer for each question with the mean response for all students with the same total score and then looking for classrooms where the average deviation from mean responses is high. The underlying logic is that students who get a question right that was missed by most students with the same total score are likely to have been cheated. Details on the calculations are provided in the appendix to Jacob and Lefgren (2003a). We employ a 95th percentile standard and eliminate any individual NRT scores where Jacob and Lefgren's classroom-level cheating measure exceeded the 95th percentile, based on the distribution of all classrooms in the district.

[18] For all but one cohort, NRT scores are available for a single year. For students who are in 6th grade in 2008/09 we have their NRT scores from 2008 (5th grade) and 2006 (3rd grade). For those students we average over both subjects and years.

correctly place the impact of future inputs within the overall effect of being cheated on achievement, [illegible]$_t$. However, if future classroom assignments are correlated with having been cheated in the past, but the association is not causal, omission of current and lagged educational inputs would bias the effect of being cheated on subsequent performance. Given this range of possibilities, we estimate models both with and without controls for measurable school-based educational inputs. The measurable inputs include teacher experience (captured by a set of indicators for various experience levels 1-2 years, 3-4 years, 5-9 years, 10-14 years, 15-24 years and 25 or more years) and classroom-level student characteristics (fraction black, fraction Hispanic, fraction with a disability, fraction gifted, fraction limited English proficiency and fraction receiving free or reduced-price lunch). To the extent that there are unmeasurable educational inputs that influence achievement, they are assumed to be randomly distributed across students and their impact is incorporated into the error term.

C. Models of Drop-out/Graduation

To estimate educational attainment outcomes, we take an approach similar to that for estimating the relationship between being cheated and student achievement. We estimate the effect of being cheated on the likelihood of later dropping out of high school (or, conversely of graduating from high school).[19] In particular, we estimate a simple cross-sectional probit model of the following form:

$$Prob\left[D_{it}=1\right]=\Phi\left[\boldsymbol{\alpha}_t\mathbf{X}_{it}+\boldsymbol{\beta}_t\mathbf{E}_{it}+\boldsymbol{\beta}_{t-1}\mathbf{E}_{i(t-1)}+\pi_t Cheated_{i2009}+\omega_t\chi_i\right]+\varepsilon_{it} \tag{4}$$

where D_{it} = 1 if a student drops out. We estimate a similar equation for high school graduation, substituting an indicator for receipt of a regular high school diploma in place of the dependent variable

[19] Following the methodology prescribed by the state in which the study is conducted, we define dropouts as students with any of the following types of withdrawals: marriage, unknown, expelled, removed for lack of attendance, financial hardship/job, incarcerated/criminal justice authority, low grades/school failure, military, adult education/post-secondary, pregnant/parent, serious illness/accident. For students who transfer outside the district (public school in another district, home school, private school, out-of-state, juvenile justice school or Department of Defense school), drop out is coded as missing. Only students receiving a standard high school diploma are counted as graduated; students receiving a certificate of performance or a special education diploma are coded as not graduating. For students with no diploma receipt and who are not known to have dropped, graduation is assigned a value of missing.

D_{it}. Given the non-linear form of equation (4), inclusion of fixed effects can produce biased results (Green, 2004). We therefore also estimate a linear probability model that includes school-by-grade or classroom fixed effects for year 2009.

V. Results

A. The Impacts of Teacher Cheating on Student Achievement

Estimates of our achievement model, both with and without controls for observable teacher and peer characteristics, are presented in Table 3. We estimate the model for each of three subjects and limit outcomes to years 2009/10-2011/12, since pre-2009 NRT scores and post-2009 CRE scores are available for only one cohort in 2012/13 (students who were in grade 4 in 2008/09).[20] The estimated models exclude controls for teacher and peer characteristics. Due to missing data for some teachers, including explicit controls for teacher experience and classroom peer characteristics reduces sample sizes, but has little effect on the estimated effects of being cheated. This suggests that a student's future classroom assignments are uncorrelated with being cheated in the past. Estimates from the model with explicit controls for contemporaneous and once-lagged teacher experience and classroom peer characteristics are provided in Appendix Table A1.

To illustrate the effects of controlling for possible selection into being cheated, we include estimates for models with no fixed effects in addition to estimates for the models that include either 2008/09 school-by-grade or 2008/09 classroom fixed effects. Consistent with non-random cheating across schools, for all three subjects we find that inclusion of school-by-grade fixed effects substantially reduces the estimated impact of being cheated on later achievement. In contrast, classroom fixed effects models yield estimates that are equal to, and frequently greater than, the estimates from models with school-by-grade effects. This is consistent with the idea that classroom fixed effects ameliorate any

[20] Pre-2009 NRT scores are available for students in grades 4-11 in 2008/09 and post-cheating CRE scores for years after 2008/09 are available for grades 3-8. Therefore the achievement analysis includes four cohorts of students, those in grades 4-7 in 2008/09.

bias associated with non-random cheating across classes, but may exacerbate selection bias due to selective cheating of students within a classroom.

For math, we find that having been cheated in 2008/09 is associated with a reduction in achievement in 2009/10 of 0.07 standard deviations, which is equivalent to the differential between having a rookie teacher rather than one with five years of experience (Clotfelter, et al., 2006). The estimates from the school-by-grade and classroom fixed-effects models are nearly identical, which could occur if the effects of being cheated are felt mostly by those students whose scores were manipulated, with little or no spillovers on classmates. Past 2009/10, we find no statistically significant effects of having been cheated on subsequent test scores.

The patterns of estimates for reading and ELA differ from those in math. In both reading and ELA, the effects of being cheated on later test scores persist past the first non-cheating year. In the first post-cheating year, we estimate that being cheated in the past lowers reading scores by 0.08 standard deviations, relative to non-cheated students in the same school and grade. In ELA the point estimate is a bit smaller, at 0.05, but still statistically significant at better than a 95 percent confidence level. In the second and third years after cheating, students whose tests were manipulated score .11 to .12 standard deviations lower in reading achievement than do schoolmates in the same grade with comparable observable characteristics and baseline test scores. In ELA, the differential is -0.05 to -0.10 standard deviations. When comparing students who were in the same classroom when cheating occurred, we observe somewhat larger differences in achievement two and three years after cheating took place. For reading, the within-classroom differential is -0.14 to -0.16 standard deviations; for ELA it is -0.07 to -0.12 standard deviations.

To account for the fact that some students' scores were manipulated more than others, we divide cheated students into two categories: those with fewer than 10 WTR erasures and those with 10 or more WTR erasures. Estimates of the effect of cheating on later student test scores with this alternative specification are presented in Table 4. For math, the effects on later test scores are concentrated among those students whose scores were manipulated the most. We find no significant drops in test scores for students with fewer than 10 WTR erasures, but large drops in test scores (9 to 14 percent of a standard deviation) for students with 10 or more WTR erasures. For reading, we observe negative effects for both groups of cheated students, though the impacts are at least twice as large for the group who had more answers changed from wrong to right. Similarly, in ELA the point estimates of the impact of cheating for the group of students whose scores were manipulated the most are at least three times that of the group of cheated students with fewer than 10 WTR erasures.

B. The Impacts of Teacher Cheating on Educational Attainment

We present estimates from the drop out and graduation equations in Table 5. Both probit and linear probability model estimates without fixed effects are presented, as well as linear probability model estimates with school-by-grade and classroom fixed effects. Both the non-linear and linear models without fixed effects produce similar results. Holding constant baseline test scores and observable student characteristics, both models estimate that middle school students who had their test scores manipulated in 2008/09 were six to seven percent more likely to have dropped out within five years after eighth grade than students who had not been cheated. The estimate is statistically significant at better than the 95 percent confidence in both models. Likewise, both models produce an estimate of the effect of cheating on graduation of -0.04, though the effects are not precisely estimated and we cannot reject the null of there being no effect on graduation. When we add school-by-grade or classroom fixed effects to the linear probability model, the estimated effects of being cheated drop by more than half in magnitude and are not statistically different from zero at conventional confidence levels.

C. Mechanisms by Which Cheating Affects Student Outcomes

Our primary interest is determining the extent to which teacher cheating impacts subsequent outcomes for students. However, our findings that cheating is associated with reductions in student achievement and possibly a greater probability of drop out do beg the question of what is the mechanism by which test-score manipulation translates into negative future outcomes. We consider four possible mechanisms and offer empirical evidence to help distinguish between each of the four possibilities.

1. Self-esteem

If a student is cheated and then later learns that their true performance is lower, it is possible this could damage their self-esteem. Prior studies have shown that self-esteem may have positive effects on both educational attainment and labor market outcomes (Waddell (2006), de Araujo and Lagos (2013)). If this self-esteem mechanism is at work, we would expect that the impact of being cheated would occur only after a student has learned that his prior scores were false. While it is possible that students knew that teacher cheating was occurring, they would not have known whether their own score had been manipulated until after they received the results from the first untainted test, which was administered in spring 2010. However, as discussed above, we find that student achievement dropped on the spring 2010 exam in all three subjects and remained constant (reading and ELA) or rose (math), relative to expectations based on measured ability and student characteristics. We would also expect that any reduction in self-esteem caused by the revelation that prior scores were false would impact student performance in non-achievement dimensions. For example, a reduction in self-esteem could make it less likely a student attends school or increase the probability a student misbehaves. In Tables 6 and 7 we provide evidence that this was not the case. Table 6 presents estimates from a model of attendance which is identical to the achievement model (equation (3)), except for a change in the dependent variable from normed test scores to the percentage of enrolled days in which a student attends school. Similarly, Table 7 presents estimates from a model that mimics the achievement model, but changes the dependent variable to the number of disciplinary fractions in a given year.[21] In contrast to achievement, we find no relationship between teacher cheating and subsequent student attendance or student behavior.

2. Grade Inflation

False test scores that result from manipulation by teachers and administrators are akin to so-called "grade inflation," where students receive grades that are greater that what might otherwise be justified by their academic performance. Babcock (2010) finds that increases in expected grades lead students to study less in college. If a similar relationship between expectations and student effort applied to standardized exams at the K-12 level, we would expect students to increase their effort once it is known that their true performance had been overstated by falsified exam scores.[22] In turn, this should lead to improvement in student achievement, all else equal. While the estimated relationship between prior teacher cheating and student achievement in math is consistent with such a scenario, the results for reading and ELA are not. In both of those subjects we do not observe an upturn in student test scores after they received their true achievement scores in 2010.

3. Rewards, Motivation and Teacher Effort

It is alleged that a primary cause of teacher cheating is external pressure to demonstrate positive school performance. Two ways to increase test scores are to improve instruction or to cheat by providing inappropriate assistance to students before or during the exam and/or correct wrong answers ex-post. The opportunity to boost scores via cheating should reduce the payoff from improved teacher effectiveness (since otherwise low test scores would be manipulated). This in turn should result in a

[21] In contrast to the achievement equation, where the determination of cheating is subject specific, we define being cheated in the attendance and discipline equations as having unusually high WTR erasures (greater than the 95th percentile of the 2013 distribution) in either math, reading or ELA.

[22] Any such effects on student performance may be muted, however, if the achievement exams were considered to be less consequential than course grades by students.

decrease in teacher effort and a reduction in true student learning. A key assumption for this to occur is that teachers adjust their effort to changes in incentives related to test scores. Indirect evidence on the relationship between teacher effort and incentives can be found in the literature on the effects of performance pay. While early experimental studies found little evidence that performance pay boosted teacher productivity (e.g. Springer, et al. 2010; Springer, et al. 2012; Fryer 2013), more recent analysis by Dee and Wyckoff (2015) of a district-wide scheme operated at scale (Washington DC's IMPACT teacher accountability system) provides strong evidence that existing teachers will in fact adjust their teaching performance in response to significant incentives. Thus another mechanism by which test score manipulation could have affected student outcomes would be through a reduction in teacher effectiveness. Once teacher cheating stopped, instructional quality should have returned to pre-manipulation norms. However, if teaching quality has persistent effects there could be negative consequences for later student outcomes.[23]

Our estimates of the relationship between teacher cheating and student achievement are not consistent with a general reduction in teacher effort, however. Recall that the estimated effect of teacher cheating on subsequent student performance was the same or greater in models with classroom fixed effects relative to models with school-by-grade fixed effects. If a teacher reduced her instructional effort in anticipation of manipulating test scores later, one would expect a reduction in achievement for all students in the class. Within-classroom comparisons, which occur when classroom fixed effects are employed, should therefore produce greatly reduced or negligible estimates of cheating compared to cross-classroom comparisons that occur in the school-by-grade fixed effect model.[24] The

[23] Research on student test scores suggests that teacher effects do not exhibit a high degree of persistence (Jacob, Lefgren and Sims 2010; Kinsler 2012). However, Chetty, Rockoff and Friedman (2014) do find that teacher value added is associated with positive later life outcomes.

[24] As discussed above, within-classroom estimates differ from school-by-grade estimates in two other ways. Since they account for non-random assignment of students to classrooms, within-classroom estimates of cheating impacts could be smaller due the reduction in selection bias. Also, within-classroom estimates may exacerbate bias due to selective cheating within classrooms. This would lead to larger estimated cheating effects in the classroom-fixed-effects model.

only way that a reduction in teacher effort would be consistent with the observed pattern of results is if that reduction in effort were selective, i.e. a cheating teacher reduced effort toward low-performing students (whose exams would later be altered) and increased effort to higher-performing students who would not later be cheated.

4. Interventions based on Student Achievement

One of the main concerns with teacher cheating is that students are not identified to receive remedial services, such as intervention programs, summer school or retention. To the extent that remediation improves student achievement, inflated test scores caused by teacher cheating could exclude deserving students from receiving services and consequently lower their achievement level. While results are somewhat mixed, existing research suggests that there are positive effects of remediation services, particularly at the elementary school level (Jacob and Lefgren (2004), Mariano and Martorell (2012), Matsudaira (2008), Winters and Greene (2012)).

In Table 8 we present estimates of the probability of being retained. The underlying model is the same as the drop-out model discussed above, save for a change in the dependent variable. Holding constant student ability (as measured by average national percentile scores on the NRT) and observable student characteristics, we find that students who had been cheated on the spring 2009 CRE were 0.2 to 0.5 percent less likely to be retained the following year (i.e. be enrolled in the same grade in 2009/10 as in 2008/09) than students whose scores had not been manipulated. This is a relatively substantial effect, given that the average retention rate is about two percent (see Table 1). Thus it does appear that a non-trivial number of students who would have otherwise been retained were instead advanced to the next grade based in part on falsified test scores.

In addition to retention, the district we study offered an "early intervention program" (EIP) to students who did not meet proficiency benchmarks on the high-stakes CRE in math or reading. With some exceptions, the EIP was limited to elementary school students (ie. students in grades 5 and below). Given that baseline scores on the low-stakes NRT are only available for grades 3, 5 and 8, we can only estimate the impact of cheating on receipt of EIP services for a single cohort, students who were in fourth grade in 2008/09. Estimates of the probability of receiving EIP services in 2009/10 based on this limited sample are presented in Table 9. Holding constant the student's measured ability, students who had their test score manipulated in 2008/09 were five percentage points less likely to receive EIP services in 2009/10 than observationally equivalent students. The estimated differential becomes smaller and statistically insignificant when school-by-grade or classroom fixed effects are employed. Given the relatively small total sample, within-school-and-grade and within-classroom analyses are going to have relatively few students in each comparison group, making it quite possible that the reduction in statistical significance is due to a lack of statistical power.

In order to gauge the extent to which loss of early intervention services resulting from inflated test scores may have contributed to reductions in later achievement, we created an indicator for cheated students who would have qualified for EIP services based on their "initial right" score (as defined above), but had post-manipulation scores that exceeded the stated threshold for intervention services as a result of wrong answers be changed to correct answers.[25] Estimates of the achievement equation which include this indicator are displayed in Table 10.[26] For math, the results are quite mixed and there is no significant effect of losing EIP eligibility on test scores in the first two post-cheating years. Results are much stronger for reading and particularly for ELA, however. In both of these subjects, cheated students who lost EIP eligibility as a result of WTR erasures generally had greater test

[25] We used the number of initially correct answers to predict students' scale scores in the absence of teacher cheating. According to state Department of Education criteria, a student whose CRE score exceeded the minimum proficiency standard could still qualify for EIP services if their "documented performance" indicated that they were performing below the standard for minimum proficiency. In practice, the test-score cutoff was "fuzzy"; some students who scored below the threshold did not receive EIP services and some who scored above the cutoff did receive services.

[26] We also re-estimated the achievement model that included indicators for cheated students with high and low numbers of erasures and obtained similar results. Estimates are reported in Appendix Table A2. Limiting the sample to the single cohort of students that could have qualified for EIP (i.e. fourth graders in 2008/09) also produced similar results.

score drops in future years than did cheated students who did not lose EIP eligibility. There is also some evidence that these effects diminished over time, as one might expect. Overall, these results suggest that loss of eligibility for early intervention services could have been a contributing factor in the negative long-run effects of teacher cheating on student achievement.

VI. Summary and Conclusion

One of the many concerns voiced by those who oppose high-stakes testing is the incentive for teachers and administrators to falsify test results. Most popular accounts of test manipulation have focused on illicit behavior by educators. Ours is the first attempt to determine what impact teacher cheating has on students. For all three subjects with requisite data, math, reading and ELA, we find that teacher cheating, as evidenced by high levels of wrong-to-right erasures, is associated with drops in student performance relative to what students would be expected to obtain based on their prior performance on untainted low-stakes exams. The estimated impacts are quantitatively significant, on the order of one to two times the achievement difference between having a rookie teacher and a teacher with five years of experience. For both reading and ELA, we also find negative effects on student test scores two and three years after cheating ended. In math, however, any negative consequences of teacher cheating appear to end after the first post-cheating year. We also find some weak evidence that middle school students who were cheated were more likely to drop out of high school, though any effects on ultimately earning a standard high school diploma are statistically insignificant.

We posited four possible mechanisms by which teacher cheating could have led to lower student achievement. Both the student self-esteem and student effort mechanisms appear to be inconsistent with the pattern of results we observe. However, both selective re-allocation of teacher effort and denial of remedial services to low-achieving students may be contributing factors. In fact, we find evidence that cheated students were less likely to be retained and less likely to receive remedial services than similar students whose scores were not altered.

Our findings suggest that manipulation of test scores by teachers is not only morally objectionable, but can also have real impacts on student learning. Whether this substantially affects the cost-benefit calculus for high-stakes testing is another matter. Given the moderate negative effects we uncover, teacher cheating would have to be pervasive for it alone to outweigh the benefits of increased accountability associated with high-stakes testing.

References

Babcock, Philip (2010). "Real Costs of Nominal Grade Inflation? New Evidence from Student Course Evaluations," *Economic Inquiry,* 48(4):983-996.

Bertoni, Marco, Giorgio Brunello and Lorenzo Rocco (2013). "When the Cat is Near, the Mice won't Play: The Effect of External Examiners in Italian Schools," *Journal of Public Economics,* 104:65-77.

Boardman, Anthony E., and Richard J. Murnane (1979). "Using Panel Data to Improve Estimates of the Determinants of Educational Achievement," *Sociology of Education*, 52: 113-121.

Clotfelter, Charles T., Helen F. Ladd, and Jacob L. Vigdor (2006). "Teacher-Student Matching and the Assessment of Teacher Effectiveness," *Journal of Human Resources*, 41(4):778-820.

de Araujo, Pedro and Stephen Lagos (2013). "Self-Esteem, Education and Wages Revisited," *Journal of Economic Psychology,* 34:120-132.

Chetty, Raj, John Friedman and Jonah Rockoff (2014), "Measuring the Impacts of Teachers II: Teacher Value-Added and Student Outcomes in Adulthood," *American Economic Review,* 104(9):2633-2679.

Dee, Thomas S. and James Wyckoff (2015). "Incentives, Selection and Teacher Performance" Evidence from IMPACT," *Journal of Policy Analysis and Management*, 34(2):267-297.

Fryer, Roland G. (2013). "Teacher Incentives and Student Achievement: Evidence from New York City Public Schools," *Journal of Labor Economics*, 31(2): 373-407.

Greene, William (2004). "The Behaviour of the Maximum Likelihood Estimator of Limited Dependent Variable Models in the Presence of Fixed Effects," *Econometrics Journal*, 7:98:119.

Jacob, Brian A., Lars Lefgren and David P. Sims (2010). "The Persistence of Teacher-Induced Learning," *Journal of Human Resources,* 45(4):915-943.

Jacob, Brian A. and Steven D. Levitt (2003a). "Catching Cheating Teachers: The Results of an Unusual Experiment in Implementing Theory," *Brookings-Wharton Papers on Urban Affairs,* 2003:185-220.

Jacob, Brian A. and Steven D. Levitt (2003b). "Rotten Apples: An Investigation of the Prevalence and Predictors of Teacher Cheating," *Quarterly Journal of Economics,* 118(3):843-877.

Jacob, Brian A. and Lars Lefgren (2004). "Remedial Education and Student Achievement: A Regression-Discontinuity Analysis." *Review of Economics and Statistics*, 86(1): 226-244.

Judd, Alan (2012). "School Test Cheating Thrives While Investigations Languish," *Atlanta Journal-Constitution*, September 22, 2012. Available at http://www.myajc.com/news/news/education/school-test-cheating-thrives-while-investigations-/nSHwF/ .

Kingston, Neal M. and Amy K Clark (eds.). *Test Fraud: Statistical Detection and Methodology*. New York: Routledge, 2014.

Koedel, Cory and Julian R. Betts (2011). "Does Student Sorting Invalidate Value-Added Models of Teacher Effectiveness? An Extended Analysis of the Rothstein Critique," *Education Finance and Policy*, 6(1):18-42.

Mariano, Louis T. and Paco Martorell (2012). "The Academic Effects of Summer Instruction and Retention in New York City" *Educational Evaluation and Policy Analysis,* 35(1):96-117.

Matsudaira, Jordan D. (2008). "Mandatory Summer School and Student Achievement," *Journal of Econometrics*, 142(2):829–850.

Neal, Derek and Diane Whitmore Schanzenbach (2010). "Left Behind by Design: Proficiency Counts and Test-Based Accountability," *Review of Economics and Statistics*, 92(2):263-283.

Perry, John, Heather Vogell, Alan Judd and M. B. Pell (2012). "Cheating our Children: Suspicious School Test Scores Across the Nation," *Atlanta Journal-Constitution*, March 25, 2012. Available at http://www.myajc.com/news/news/cheating-our-children-suspicious-school-test-sco-1/nQSTS/ .

Sass, Tim R., Anastasia Semykina and Douglas N. Harris (2014). "Value-Added Models and the Measurement of Teacher Productivity," *Economics of Education Review*, 38:9-23.

Springer, Matthew, J.R. Lockwood, Dale Ballou, Daniel F. McCaffrey, Laura Hamilton, Matthew Pepper, Vi-Nhuan Le and Brian Stecher (2010). "Teacher Pay for Performance: Experimental Evidence from the Project on Incentives in Teaching." Nashville, TN: National Center on Performance Incentives.

Springer, Matthew G., John F. Pane, Vi-Nhuan Le, Daniel F. McCaffrey, Susan Freeman Burns, Laura S. Hamilton and Brian Stecher (2012). "Team Pay for Performance: Experimental Evidence From the Round Rock Pilot Project on Team Incentives," *Educational Evaluation and Policy Analysis*, 34(4):367-390.

Todd, Petra E. and Kenneth I. Wolpin (2003). "On the Specification and Estimation of the Production Function for Cognitive Achievement," *The Economic Journal*, 113: F3-F33.

van der Linden, Wim J. and Minjeong Jeon (2012). "Modeling Answer Changes on Test Items," *Journal of Educational and Behavioral Statistics*, 37(1):180-199.

Waddell, Glen R. (2006). "Labor-Market Consequences of Poor Attitude and Low Self-Esteem in Youth," *Economic Inquiry*, 44(1):69-97.

Winters, Marcus and Jay P. Greene (2012). "The Medium-Run Effects of Florida's Test-Based Promotion Policy," *Education Finance and Policy*, 7(3):305-330.

Figures

Figure 1: Distribution of Normalized Scores in 2009 and 2010 by Subject for Schools Investigated in 2009

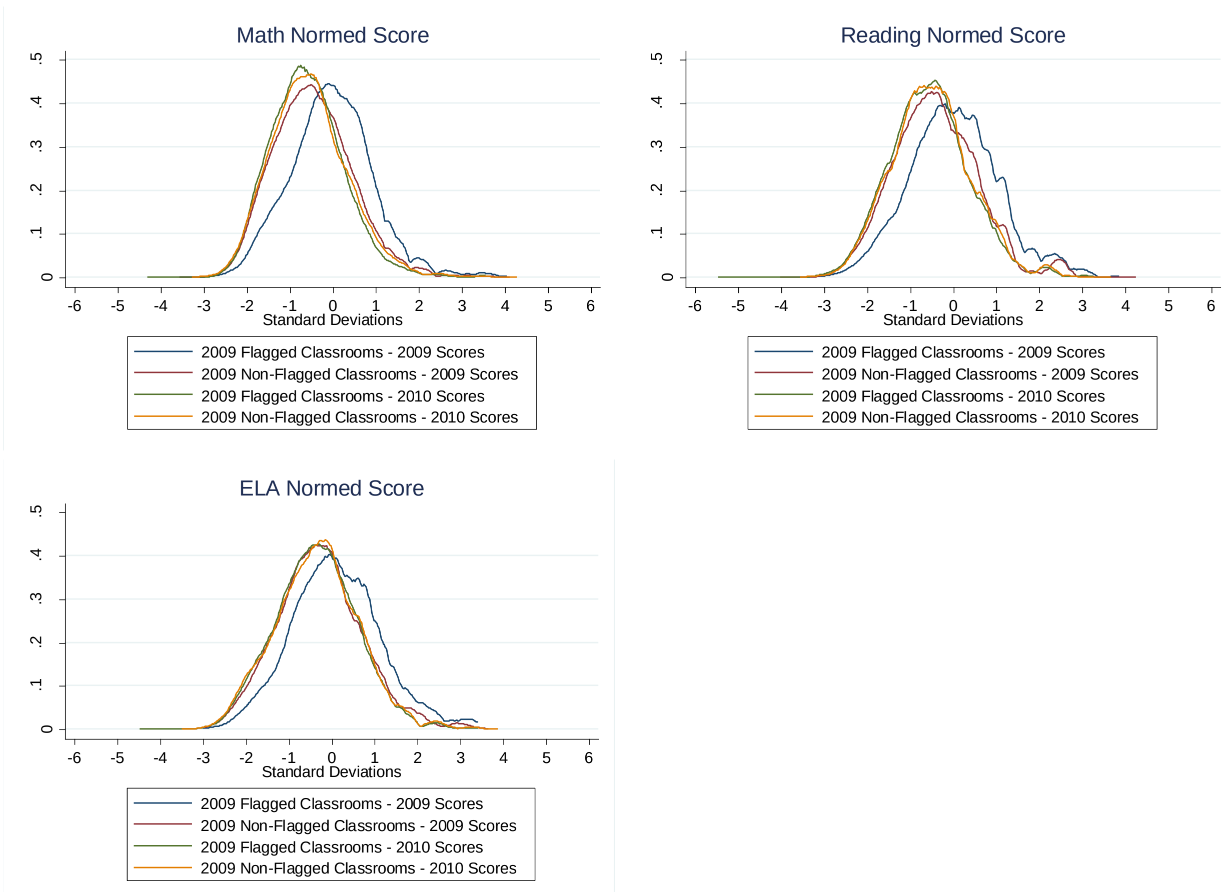


Figure 2: Distributions of the Number of "Initial Right" Answers on CRE Math Exam by Grade, Grades 1-5 (Schools which were Investigated in 2009 and in 2010)

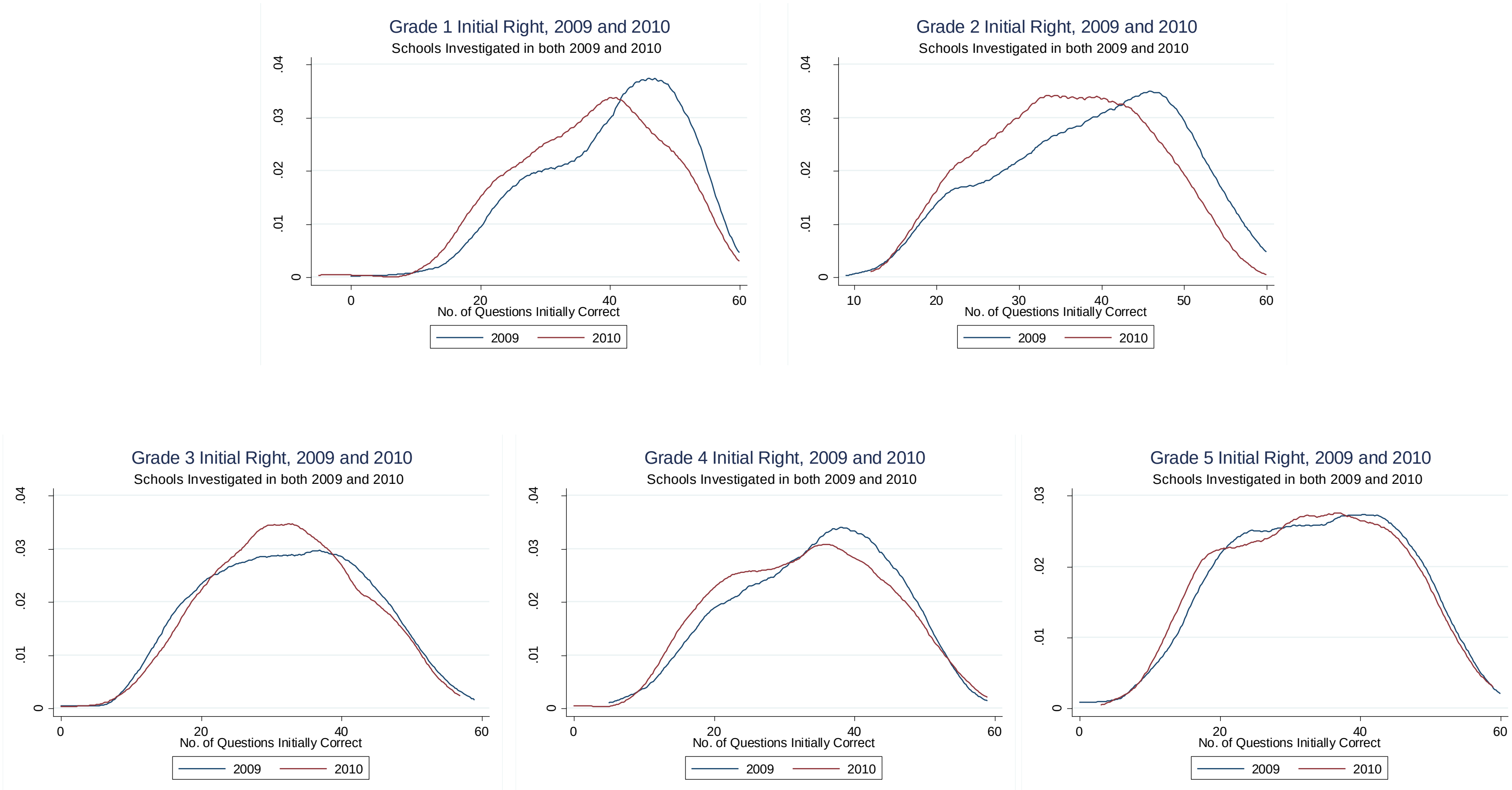
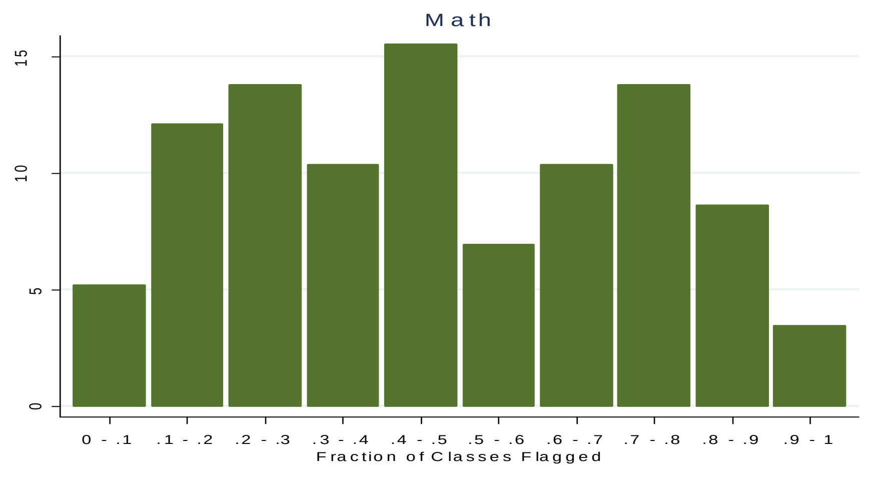
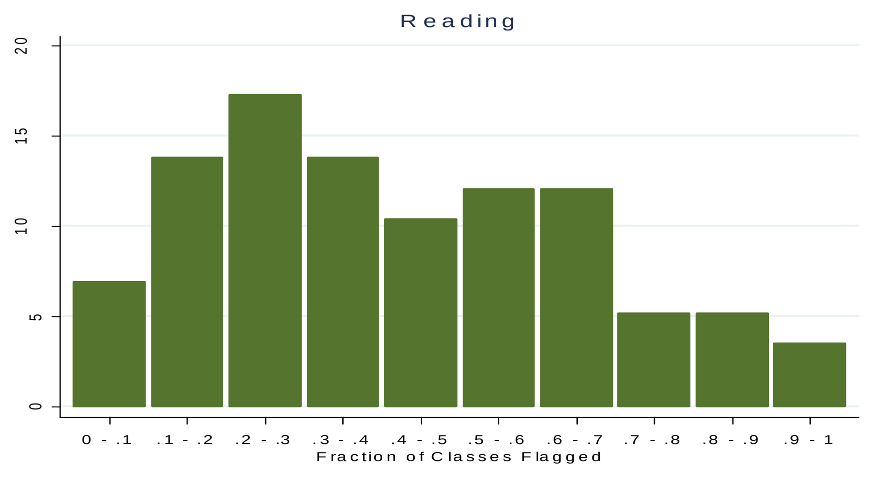
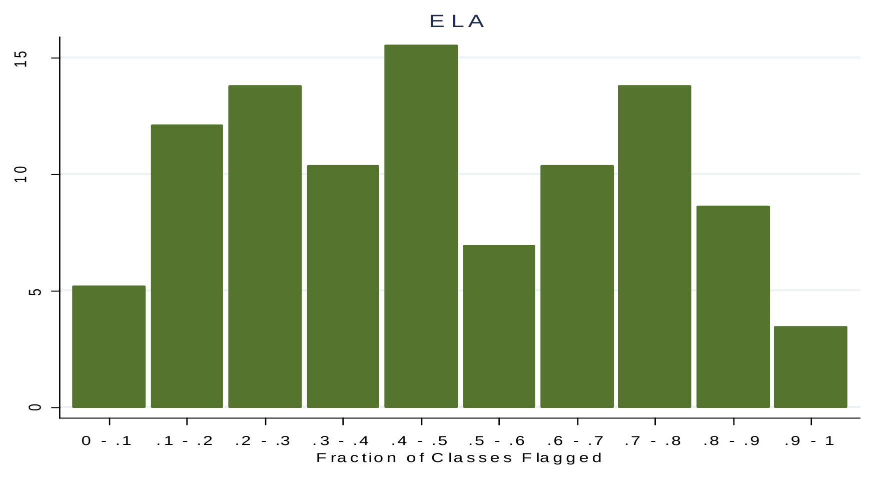


Figure 3: Distribution of Schools by Fraction of Classrooms Flagged for High Numbers of Wrong-to-Right Erasures in 2009 (Schools with Flagged Classrooms)







Tables

Table 1: Summary Statistics by Presence of Classrooms Flagged for High Numbers of Wrong-to-Right Erasures in 2008/09

	All Schools		Schools with Flagged Classrooms		Schools without Flagged Classrooms	
	Mean	S.D.	Mean	S.D.	Mean	S.D.
Black	0.83	0.37	0.95	0.22	0.77	0.42
Hispanic	0.05	0.21	0.03	0.18	0.05	0.23
White	0.10	0.29	0.01	0.07	0.14	0.35
Female	0.50	0.50	0.50	0.50	0.50	0.50
Special Education	0.10	0.30	0.09	0.29	0.10	0.30
Gifted	0.08	0.28	0.06	0.23	0.10	0.29
Limited English Proficiency	0.00	0.07	0.00	0.07	0.00	0.07
Free and Reduced Lunch	0.58	0.49	0.77	0.42	0.49	0.50
NRT National Percentile	42.61	24.30	41.16	22.39	43.34	25.18
Early Intervention Program	0.15	0.35	0.25	0.43	0.10	0.30
Retained	0.02	0.13	0.03	0.17	0.01	0.10
Dropped Out	0.29	0.45	0.31	0.46	0.28	0.45
Graduated	0.65	0.48	0.62	0.49	0.66	0.47
Observations	54,356		18,542		35,814	

National Percentile score is only available for students in grades 4-9 in 2008/09. "Schools without Flagged Classrooms" includes all non-investigated schools. Retention statistics are limited to students in grades 1-8 in the year 2009/10. Dropout and graduation statistics only cover the 7th and 8th grade cohorts in 2008/09.

Table 2: Number of Wrong-to-Right Erasures on 2009 and 2013 CRE by 2006-2008 Average NRT Achievement Quintile

	Math				Reading				ELA			
NRT Achievement Quintile	2009 (Flagged Classrooms)		2013 (All Schools)		2009 (Flagged Classrooms)		2013 (All Schools)		2009 (Flagged Classrooms)		2013 (All Schools)	
	Mean	S.D.	Mean	S.D.	Mean	S.D.	Mean	S.D.	Mean	S.D.	Mean	S.D.
First (Lowest)	9.73	9.01	1.39	1.65	6.69	6.69	1.00	1.66	7.10	7.42	1.14	1.69
Second	9.09	7.99***	1.69**	2.12***	6.19	5.44***	1.07	1.39***	6.25**	5.77***	1.05	1.48**
Third	9.58	8.14**	1.56	1.84**	5.98**	5.11***	1.08	1.33***	5.84***	5.24***	1.04	1.33***
Fourth	7.83***	6.48***	1.64**	1.98***	5.14***	4.62***	0.95	1.30***	4.85***	4.37***	1.03	1.42***
Fifth (Highest)	6.61***	6.07***	1.76***	2.08***	4.22***	4.09***	0.89	1.57	4.33***	4.34***	1.15	1.76
All	8.67	7.75	1.62	1.96	5.70	5.36	0.99	1.46	5.72	5.64	1.09	1.56

* Indicates quintile mean/standard deviation is significantly different from quintile one mean/standard deviation at the 10% level.
** Indicates quintile mean/standard deviation is significantly different from quintile one mean/standard deviation at the 5% level.
***Indicates quintile mean/standard deviation is significantly different from quintile one mean/standard deviation at the 1% level.

Table 3: Estimated Effect of Being Cheated in 2008/09 on CRE Normed Test Scores, 2009/10 – 2011/2012

Model	School Year		
	2009/10	2010/11	2011/12
Math			
No Fixed Effects	-0.1255*** (0.0248)	-0.0934*** (0.0262)	-0.0722*** (0.0274)
School-by-Grade Fixed Effects	-0.0674*** (0.0217)	-0.0209 (0.0268)	-0.0064 (0.0288)
Classroom Fixed Effects	-0.0692** (0.0274)	-0.0200 (0.0323)	-0.0307 (0.0344)
N	5,060	3,820	2,786
Reading			
No Fixed Effects	-0.1280*** (0.0254)	-0.1338*** (0.0240)	-0.1438*** (0.0294)
School-by-Grade Fixed Effects	-0.0763*** (0.0253)	-0.1071*** (0.0290)	-0.1168*** (0.0343)
Classroom Fixed Effects	-0.1086*** (0.0312)	-0.1426*** (0.0360)	-0.1623*** (0.0420)
N	4,912	3,708	2,731
ELA			
No Fixed Effects	-0.0809*** (0.0250)	-0.0778*** (0.0272)	-0.1384*** (0.0284)
School-by-Grade Fixed Effects	-0.0526** (0.0243)	-0.0531* (0.0284)	-0.0970*** (0.0296)
Classroom Fixed Effects	-0.0690** (0.0277)	-0.0650* (0.0339)	-0.1171*** (0.0338)
N	4,991	3,769	2,769

Dependent variable is the normed CRE score in the given school year. Explanatory variables include: mean of national percentile NRT scores in 2006-2008 averaged across five subjects, gender, race/ethnicity, free/reduced-price lunch status, gifted status, limited English proficiency status, and a set of 15 disability categories. Fixed effects are for the classroom, grade and school attended in 2008/09. Standard errors clustered at the 2008/09 classroom level in parentheses. *significant at the 10% level, **significant at the 5% level, ***significant at the 1% level in a two-tailed test.

Table 4: Estimated Effects of Being Cheated in 2008/09 on CRE Normed Test Scores, 2009/10 – 2011/2012 (School-by-Grade and Classroom Fixed-Effects Models)

Variable	Model					
	School-by-Grade FE			Classroom FE		
	2009/10	2010/11	2011/12	2009/10	2010/11	2011/12
Math						
Cheated – Less than 10 WTR	-0.0013 (0.0271)	0.0479 (0.0313)	0.0788** (0.0330)	-0.0137 (0.0308)	0.0451 (0.0351)	0.0540 (0.0372)
Cheated – 10 or More WTR	-0.1308*** (0.0276)	-0.0893*** (0.0338)	-0.1020** (0.0405)	-0.1381*** (0.0358)	-0.1046** (0.0427)	-0.1427*** (0.0492)
N	5,060	3,820	2,786	5,060	3,820	2,786
Reading						
Cheated – Less than 10 WTR	-0.0305 (0.0290)	-0.0622* (0.0329)	-0.0555 (0.0370)	-0.0663* (0.0344)	-0.1065*** (0.0386)	-0.1029** (0.0432)
Cheated – 10 or More WTR	-0.1634*** (0.0344)	-0.1923*** (0.0356)	-0.2244*** (0.0454)	-0.2040*** (0.0407)	-0.2269*** (0.0452)	-0.2895*** (0.0581)
N	4,912	3,708	2,731	4,912	3,708	2,731
ELA						
Cheated – Less than 10 WTR	-0.0119 (0.0276)	-0.0072 (0.0300)	-0.0505 (0.0321)	-0.0327 (0.0304)	-0.0238 (0.0341)	-0.0737** (0.0359)
Cheated – 10 or More WTR	-0.1509*** (0.0357)	-0.1621*** (0.0435)	-0.1925*** (0.0466)	-0.1742*** (0.0409)	-0.1854*** (0.0535)	-0.2243*** (0.0541)
N	4,991	3,769	2,769	4,991	3,769	2,769

Dependent variable is the normed CRE score in the given school year. Explanatory variables include: mean of national percentile NRT scores in 2006-2008 averaged across five subjects, gender, race/ethnicity, free/reduced-price lunch status, gifted status, limited English proficiency status, and a set of 15 disability categories. Fixed effects are for the classroom, grade and school attended in 2008/09. Standard errors clustered at the 2008/09 classroom level in parentheses. *significant at the 10% level, **significant at the 5% level, ***significant at the 1% level in a two-tailed test.

Table 5: Estimated Effect of Being Cheated in Either Math, Reading or ELA in 2008/09 on the Probability of Dropout and the Probability of Graduation, 2008/09 7th and 8th Grade Cohorts

Model	Dropout	Graduation
Probit	0.0679** (0.0331)	-0.0426 (0.0362)
Linear Probability	0.0647** (0.0313)	-0.0389 (0.0315)
Linear Probability with School-by-Grade Fixed Effects	0.0176 (0.0323)	0.0017 (0.0322)
Linear Probability with Classroom Fixed Effects	0.0237 (0.0326)	0.0067 (0.0391)
N (Probit) N (OLS)	1,134 1,139	1,047 1,069

Reported coefficients are marginal effects. Explanatory variables include: mean of national percentile NRT scores in 2006-2008 averaged across five subjects, gender, race/ethnicity, free/reduced-price lunch status, gifted status, limited English proficiency status, and a set of 15 disability categories. Standard errors clustered at the 2008/09 classroom level in parentheses. *significant at the 10% level, **significant at the 5% level, ***significant at the 1% level in a two-tailed test.

Table 6. Estimated Effect of Being Cheated in 2008/09 on Percent Attendance, 2009/10 – 2013/2014

Model	School Year				
	2009/10	2010/11	2011/12	2012/13	2013/14
School-by-Grade Fixed Effects	0.0330 (0.1665)	-0.0548 (0.2461)	0.0521 (0.3526)	-0.3221 (0.3941)	-0.7622 (0.5191)
Classroom Fixed Effects	0.1350 (0.2005)	0.1430 (0.2874)	0.2219 (0.4550)	-0.1957 (0.4724)	-0.5099 (0.6365)
N	5,852	5,364	5,006	4,619	3,786

Dependent variable is the percentage of enrollment days a student attended school in the given school year. Explanatory variables include: mean of national percentile NRT scores in 2006-2008 averaged across five subjects, gender, race/ethnicity, free/reduced-price lunch status, gifted status, limited English proficiency status, and a set of 15 disability categories. Fixed effects are for the classroom, grade and school attended in 2008/09. There are no explicit controls for teacher or peer characteristics. Standard errors clustered at the 2008/09 classroom level in parentheses. *significant at the 10% level, **significant at the 5% level, ***significant at the 1% level in a two-tailed test.

Table 7: Estimated Effect of Being Cheated in 2008/09 on the Number of Discipline Incidents, 2009/10 – 2013/2014

Model	School Year				
	2009/10	2010/11	2011/12	2012/13	2013/14
School-by-Grade Fixed Effects	-0.0241 (0.0598)	0.0320 (0.0918)	-0.0521 (0.1291)	0.1123 (0.0930)	0.0007 (0.0780)
Classroom Fixed Effects	-0.0579 (0.0715)	-0.0066 (0.1082)	-0.0397 (0.1639)	0.1546 (0.1192)	-0.0466 (0.0984)
N	5,851	5,360	5,004	4,610	3,784

Dependent variable is the percentage of enrollment days a student attending school in in the given school year. Explanatory variables include: mean of national percentile NRT scores in 2006-2008 averaged across five subjects, gender, race/ethnicity, free/reduced-price lunch status, gifted status, limited English proficiency status, and a set of 15 disability categories. Fixed effects are for the classroom, grade and school attended in 2008/09. Standard errors clustered at the 2008/09 classroom level in parentheses. *significant at the 10% level, **significant at the 5% level, ***significant at the 1% level in a two-tailed test.

Table 8: Estimated Effect of Being Cheated in Either Math, Reading or ELA in 2008/09 on the Probability of Being Retained in 2009/10

Model	Retained in 2009/10
Probit	-0.0017* (0.0011)
Linear Probability	-0.0051* (0.0027)
Linear Probability with School-by-Grade FE	-0.0060* (0.0034)
Linear Probability with Classroom FE	-0.0040 (0.0041)
N (Probit) N (OLS)	5,462 6,160

Reported coefficients are marginal effects. Explanatory variables include: mean of national percentile NRT scores in 2006-2008 averaged across five subjects, gender, race/ethnicity, free/reduced-price lunch status, gifted status, limited English proficiency status, and a set of 15 disability categories. Standard errors in parentheses. *significant at the 10% level, **significant at the 5% level, ***significant at the 1% level in a two-tailed test.

Table 9: Estimated Effect of Being Cheated in Either Math, Reading or ELA in 2008/09 on the Probability of Receiving Early-Intervention Program Services in 2009/10, 2008/09 4th Grade Cohort

Model	Enrolled in Early Intervention Program in 2009/10
Probit	-0.0473* (0.0271)
Linear Probability	-0.0461* (0.0237)
Linear Probability with School-by-Grade FE	-0.0279 (0.0225)
Linear Probability with Classroom FE	-0.0065 (0.0248)
N (Probit) N (OLS)	2,057 2,061

Reported coefficients are marginal effects. Explanatory variables include: mean of national percentile NRT scores in 2008 averaged across five subjects, gender, race/ethnicity, free/reduced-price lunch status, gifted status, limited English proficiency status, and a set of 15 disability categories. Standard errors in parentheses. *significant at the 10% level, **significant at the 5% level, ***significant at the 1% level in a two-tailed test.

Table 10: Estimated Effects of Losing Eligibility for Early Intervention Program Services as a Result of Being Cheated in 2008/09 on CRE Normed Test Scores, 2009/10 – 2011/2012 (School-by-Grade and Classroom Fixed-Effects Models)

	Model					
	School-by-Grade FE			Classroom FE		
	2009/10	2010/11	2011/12	2009/10	2010/11	2011/12
Math						
Lost Eligibility for EIP due to Being Cheated	-0.0354 (0.0447)	-0.0465 (0.0465)	-0.1189** (0.0509)	-0.0590 (0.0535)	-0.0785 (0.0528)	-0.1158** (0.0570)
Cheated in 2008/09	-0.0617*** (0.0231)	-0.0151 (0.0289)	0.0204 (0.0319)	-0.0599** (0.0284)	-0.0082 (0.0338)	-0.0089 (0.0374)
N	5,060	3,777	2,744	5,060	3,777	2,744
Reading						
Lost Eligibility for EIP due to Being Cheated	-0.0906** (0.0455)	-0.0472 (0.0573)	-0.0308 (0.0560)	-0.1184** (0.0530)	-0.0549 (0.0661)	-0.0390 (0.0711)
Cheated in 2008/09	-0.0703*** (0.0259)	-0.0998*** (0.0291)	-0.1121*** (0.0348)	-0.1013*** (0.0317)	-0.1551*** (0.0423)	-0.1342*** (0.0361)
N	4,912	3,665	2,687	4,912	3,665	2,687
ELA						
Lost Eligibility for EIP due to Being Cheated	-0.1422*** (0.0485)	-0.1048* (0.0583)	-0.1179* (0.0624)	-0.1497*** (0.0574)	-0.1403** (0.0691)	-0.1001 (0.0709)
Cheated in 2008/09	-0.0404 (0.0248)	-0.0444 (0.0292)	-0.0843*** (0.0307)	-0.0580** (0.0283)	-0.0544 (0.0347)	-0.1059*** (0.0352)
N	4,991	3,725	2,725	4,991	3,725	2,725

Dependent variable is the normed CRE score in the given school year. Explanatory variables include: mean of national percentile NRT scores in 2006-2008 averaged across five subjects, gender, race/ethnicity, free/reduced-price lunch status, gifted status, limited English proficiency status, and a set of 15 disability categories. Fixed effects are for the classroom, grade and school attended in 2008/09. Standard errors clustered at the 2008/09 classroom level in parentheses. *significant at the 10% level, **significant at the 5% level, ***significant at the 1% level in a two-tailed test.

Table A1: Estimated Effect of Being Cheated in 2008/09 on CRE Normed Test Scores, 2009/10 – 2011/2012 (Models with Teacher Experience and Peer Characteristics Controls)

Model	School Year		
	2009/10	2010/11	2011/12
Math			
No Fixed Effects	-0.1195*** (0.0261)	-0.1125*** (0.0277)	-0.0609** (0.0251)
School-by-Grade Fixed Effects	-0.0700*** (0.0239)	-0.0467 (0.0297)	-0.0241 (0.0287)
Classroom Fixed Effects	-0.0700** (0.0304)	-0.0483 (0.0357)	-0.0497 (0.0349)
N	3,858	3,026	2,272
Reading			
No Fixed Effects	-0.1157*** (0.0252)	-0.1193*** (0.0255)	-0.1250*** (0.0283)
School-by-Grade Fixed Effects	-0.0858*** (0.0262)	-0.0982*** (0.0305)	-0.1017*** (0.0327)
Classroom Fixed Effects	-0.1209*** (0.0322)	-0.1379*** (0.0380)	-0.1371*** (0.0411)
N	4,150	3,285	2,481
ELA			
No Fixed Effects	-0.0746*** (0.0251)	-0.0719*** (0.0276)	-0.1252*** (0.0288)
School-by-Grade Fixed Effects	-0.0598** (0.0254)	-0.0535* (0.0285)	-0.0879*** (0.0310)
Classroom Fixed Effects	-0.0715** (0.0297)	-0.0623* (0.0349)	-0.1071*** (0.0366)
N	4,240	3,353	2,518

Dependent variable is the normed CRE score in the given school year. Explanatory variables include: mean of national percentile NRT scores in 2006-2008 averaged across five subjects, gender, race/ethnicity, free/reduced-price lunch status, gifted status, limited English proficiency status, a set of 15 disability categories, mean classroom student characteristics and a set of teacher experience category indicators. Fixed effects are for the classroom, grade and school attended in 2008/09. Standard errors clustered at the 2008/09 classroom level in parentheses. *significant at the 10% level, **significant at the 5% level, ***significant at the 1% level in a two-tailed test.

Table A2: Estimated Effects of Losing Early Intervention Program Services as a Result of Being Cheated in 2008/09 on CRE Normed Test Scores, 2009/10 – 2011/2012 (School-by-Grade and Classroom Fixed-Effects Models)

Specification	Model					
	School-by-Grade FE			Classroom FE		
	2009/10	2010/11	2011/12	2009/10	2010/11	2011/12
Math						
Lost Eligibility for EIP due to Being Cheated	0.0009 (0.0445)	-0.0028 (0.0481)	-0.0679 (0.0532)	-0.0248 (0.0531)	-0.0350 (0.0536)	-0.0649 (0.0586)
Cheated – Less than 10 WTR	-0.0014 (0.0276)	0.0487 (0.0317)	0.0894*** (0.0344)	-0.0113 (0.0313)	0.0483 (0.0356)	0.0637 (0.0390)
Cheated – 10 or More WTR	-0.1310*** (0.0294)	-0.0939** (0.0373)	-0.0804* (0.0459)	-0.1322*** (0.0371)	-0.0981** (0.0454)	-0.1273** (0.0542)
N	5,060	3,777	2,744	5,060	3,777	2,744
Reading						
Lost Eligibility for EIP due to Being Cheated	-0.0655 (0.0463)	-0.0226 (0.0563)	-0.0004 (0.0551)	-0.0929* (0.0534)	-0.0328 (0.0655)	-0.0056 (0.0702)
Cheated – Less than 10 WTR	-0.0287 (0.0291)	-0.0564* (0.0330)	-0.0507 (0.0371)	-0.0636* (0.0345)	-0.1000** (0.0386)	-0.0963** (0.0433)
Cheated – 10 or More WTR	-0.1543*** (0.0360)	-0.1858*** (0.0359)	-0.2243*** (0.0467)	-0.1913*** (0.0421)	-0.2178*** (0.0457)	-0.2860*** (0.0591)
N	4,912	3,665	2,687	4,912	3,665	2,687
ELA						
Lost Eligibility for EIP due to Being Cheated	-0.1208** (0.0520)	-0.1071* (0.0621)	-0.1285* (0.0665)	-0.1412** (0.0596)	-0.1502** (0.0723)	-0.1088 (0.0751)
Cheated – Less than 10 WTR	-0.0011 (0.0236)	0.0380 (0.0311)	0.0505 (0.0349)	0.0156 (0.0285)	0.0512 (0.0362)	0.0419 (0.0419)
Cheated – 10 or More WTR	-0.1057** (0.0458)	-0.0894 (0.0547)	-0.1304** (0.0582)	-0.1084** (0.0499)	-0.1042 (0.0645)	-0.1617** (0.0666)
N	4,991	3,725	2,725	4,991	3,725	2,725

Dependent variable is the normed CRE score in the given school year. Explanatory variables include: mean of national percentile NRT scores in 2006-2008 averaged across five subjects, gender, race/ethnicity, free/reduced-price lunch status, gifted status, limited English proficiency status, and a set of 15 disability categories. Fixed effects are for the classroom, grade and school attended in 2008/09. Standard errors clustered at the 2008/09 classroom level in parentheses. *significant at the 10% level, **significant at the 5% level, ***significant at the 1% level in a two-tailed test.