The Long-Run Effects of Teacher Cheating on Student Outcomes

A Report for the Atlanta Public Schools

Tim R. Sass, Distinguished University Professor, Georgia State University
Jarod Apperson, Ph.D. student, Georgia State University
Carycruz Bueno, Ph.D. student, Georgia State University

May 5, 2015

Executive Summary

The manipulation of student test scores by teachers and administrators in the Atlanta Public Schools (APS) has been carefully documented and dozens of teachers and leaders have either accepted administrative sanctions, plead guilty to criminal charges or have been convicted of crimes. Little is known, however, about how the falsification of test scores by teachers and administrators ("teacher cheating") has impacted students. Using a panel of individual-level data on students and teachers from APS, we investigated the effects of teacher cheating on subsequent student achievement, attendance and student behavior. Key findings include:

- Of the 11,553 students who were in classrooms flagged by the Governor's Office of Student Achievement (GOSA) for high levels of wrong-to-right erasures on the 2009 CRCT exam, 5,888 (51 percent) were still enrolled in APS in fall 2014.
- Not all students in classrooms where cheating occurred had their test scores manipulated equally. For both reading and English Language Arts (ELA), over one-third of students in flagged classrooms had two or fewer wrong-to-right (WTR) erasures and nearly one-fourth of students in flagged classrooms had two or fewer WTR erasures on their math CRCT exam.
- There is strong evidence that teachers were selective in their manipulation of test scores. In particular, teachers were more likely to change answers for students of lower apparent ability.
- Using the frequency of erasures in 2013 as a benchmark, of the 11,553 students in flagged classrooms in 2008/09 we estimate that 7,064 (61 percent) likely had their test answers manipulated in one or more subjects on the 2009 CRCT exam. Of these 7,064 students, 3,728 were still enrolled in APS in fall 2014.
- There is relatively robust evidence that manipulation of students' test answers had negative consequences for later student performance in reading and English Language Arts (ELA). The estimated impacts are equivalent to a one-time reduction in achievement of roughly one-fourth to one-half of the average annual achievement gain for middle school students. Put differently, the estimated loss in achievement is one to two times the difference in achievement between having a rookie teacher rather than a teacher with five years of experience for a single year. In contrast, the effects of teacher cheating for subsequent achievement in math are mixed.
- There is little or no evidence that teacher cheating had deleterious effects on subsequent student attendance or student behavior.

1. Introduction

Much attention has been paid to the teachers and administrators caught up in the Atlanta Public Schools (APS) cheating scandal. However, little is known about how the falsification of student test scores has impacted students. While illicit behavior by teachers and administrators (henceforth "teacher cheating") obviously boosted student test scores in the short term, it is unknown what effect this cheating had on subsequent student achievement and other outcomes such as attendance, discipline and high school completion. In this report we investigate the effects of test-score manipulation on student outcomes to learn the extent to which students may have been harmed in the long-run. We focus on the relationship between teacher cheating and post-cheating student test scores, attendance and behavior. In addition to measuring the impact of teacher cheating on student outcomes, we seek to quantify the number of students potentially affected by teacher cheating and determine if they are still enrolled within APS.

Specifically, this report addresses the following questions:

1. How many students were potentially impacted by teacher cheating in APS?
2. Among potentially impacted students, what is the effect of teacher cheating on test score outcomes in the long run and does it vary across students and across subjects?
3. Were students who received inflated test scores less likely to attend school after they learned that their prior scores were incorrect?
4. How has the disciplinary behavior of potentially impacted students changed since teacher cheating was uncovered?
5. How many students who were potentially impacted are still enrolled in APS?
6. What proportion of students who were potentially impacted transferred from traditional to charter schools or left the Atlanta Public School system?

2. Possible Mechanisms and Related Research

The existing research related to teacher cheating has focused on the identification of cheating (Jacob and Levitt, 2003a, 2003b; van der Linden and Jeon, 2012; Kingston and Clark, 2014). There is no prior research that specifically considers the possible impacts of altered test scores on subsequent student performance. However, there are four related strands of literature, student self-esteem, grade inflation, teacher incentives and achievement-based interventions, each of which provides some possible mechanisms by which teacher cheating could impact students in the long run.

A. Self-esteem

If a student receives an unexpectedly high test score it is conceivable that this could increase their own perception of their academic ability and consequently boost their self-esteem. Of course, if it is later learned that the prior test score was false, any increase in self-esteem could be eliminated or even reversed. A number of studies have shown that self-esteem may have positive effects on both academic and employment outcomes. For example, Waddell (2006) finds that high school graduates with low self-esteem attain fewer years of postsecondary education, are more likely to be unemployed and have lower earnings than others in their high school graduation cohort. de Araujo and Lagos (2013) consider the simultaneous determination of self-esteem, educational attainment and earnings and find that positive self-esteem boosts wages by increasing educational attainment, but has no direct effect on wages.

In the case of manipulated test scores any impacts of inflated standardized test scores on self-esteem would likely be muted by signals of lower performance from course grades. If teacher cheating did in fact affect student self-esteem, we would expect that student outcomes would improve (beyond the direct effect of falsification on scores) until falsification became known and then they would decline, perhaps to levels below what would be expected prior to the initiation of cheating.

B. 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. Using survey data from college course evaluations, he finds average study time is reduced by half in a class where the average expected grade is an "A," relative to a class in which the average expected grade is a "C." Thus it is possible that, at least in the short term, students could react to inflated test scores by devoting less effort in school, believing that they can perform well with little effort. If this is the case, one would expect even lower scores once test answers were no longer manipulated. Any such effects on student performance may be tempered, however, if the achievement exams were considered to be less consequential than course grades by students.[1]

C. Rewards, Motivation and Teacher Effort

It is alleged that a primary cause of teacher cheating in APS was 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

[1] The impact may also be dampened if achievement test results are not known until after the end of the academic year.

would be manipulated). This in turn could result in a 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 (2013) 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.

D. Interventions Based on Student Achievement

One of the main concerns with teacher cheating is that because of artificially inflated test scores, students are not identified to receive remedial services, such as intervention programs, summer school or retention. If remedial programs increase student achievement, denial of these services could potentially have lasting effects on the student. However, the evidence on the efficacy of retention and summer school placement for low-achieving students is somewhat mixed, with positive effects occurring primarily in elementary school. Most of the rigorous research studies conducted thus far exploit achievement based rules for mandating retention and/or summer school and compare outcomes for students just below the intervention threshold (who receive services) with those whose scores are just above the threshold (and thus do not receive remediation). Jacob and Lefgren (2004) find that both summer school and retention have a positive (but small) impact on student achievement for third graders and no effect for sixth graders in Chicago. Mariano and Martorell (2012) analyze a two tier-cutoff system in New York City where students who score below the cutoff on their spring assessment are assigned to summer school; if these students fall below the summer evaluation cutoff they are retained. The authors find that summer school had a small positive impact on student's 6^{th}-grade ELA achievement if they attended for missing the ELA cutoff, but not for students who scored low on the math test. Using a same-year comparison, they find that retention in fifth grade has a large, significant and positive effect on student performance for the following two years. Matsudaira (2008) studies a mandatory summer school policy and finds that positive and quantitatively substantial average effects of summer school on subsequent achievement in both math and reading achievement. However, the estimated impacts vary substantially across grade-level and subject combinations. Winters and Greene (2012) evaluate the impacts of Florida's third-grade retention policy. Not only were students who fell below the threshold retained, they also were assigned a "high-quality" teacher in the retention year and were required to attend summer school. This combination of remediation efforts had substantial positive effects on achievement in multiple subjects, though the effects diminish over time.

3. Data and Background

To answer the research questions outlined above, we constructed a longitudinal dataset from APS administrative data. With the assistance of APS staff we were able to collect anonymous individual-level data on enrollment, attendance, discipline, program participation (e.g. special education, subsidized lunch, English as a second language), school types, student demographics and diploma receipt. APS also provided state test results from the Criterion-Referenced Competency Test (CRCT) in grades 1-8 and high school End-of-Course Tests (EOCTs).[2] For both the CRCT and EOCTs we computed normalized scores for each grade (or course for the EOCTs) and year based on statewide means and standard deviations. All of the data were assembled into a panel covering all students enrolled in APS from the 2004/05 school year through fall of 2014.

Another key data source is the Georgia Bureau of Investigation (GBI) report on teacher cheating in APS (Office of the Governor, 2011). The GBI investigation included information from the Governor's Office of Student Achievement's (GOSA's) analysis of erasures on the spring 2009 CRCT exam. Results of the erasure analysis were used by the GBI 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. Interview data from the report provided valuable contextual information about how teacher cheating was carried out.

The data provided by APS include individual-level erasure data for the CRCT in school years 2008/09 through 2012/13. The erasure data from 2008/09 and 2009/10 only cover schools that were investigated by the GBI, based on high levels of wrong-to-right (WTR) erasures in one or more classrooms and other evidence of testing improprieties (henceforth "investigated schools"). Thus we have erasure data for 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.[3] The erasure data provided to us contain the raw number of correct answers and the number of WTR erasures. The individual-level erasure data were combined with information on the criteria used by the GBI to "flag" classrooms suspected of cheating.[4] Using the erasure data and the GOSA

[2] 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.

[3] Our 2008/2009 erasure data cover all schools initially targeted for investigation, including two schools that did not receive a detailed investigation because initial inquiries uncovered no evidence of improprieties.

[4] 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 GBI reports (Office of the Governor, 2011) refer to "flagged classrooms," though they were in

methodology, we identified both schools which were investigated by the GBI and flagged classrooms within the investigated schools for school years 2008/09 and 2009/10.

4. Research Design and Methods

A. Identifying Cheating

A key element in the analysis of the effects of teacher cheating is identifying which students had their test scores manipulated. There are three types of teacher cheating possible. First, teachers with advanced 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."

Interviews conducted during the GBI investigation uncovered evidence that all three types of cheating occurred in APS. 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 methods that have been used in prior research to identify teacher cheating. One method, developed by Jacob and Levitt (2003a, 2003b), is to look for unusual patterns of responses, either within a single student's answer sheet or across student answer sheets. We do not have access to the answer sheets of individual students, so this approach is not an option in the current analysis.

A second approach employed by Jacob and Levitt (2003b) 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 APS, the run-up in test scores associated with cheating will not be observed for students whose test scores were

---

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.

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.

As illustrated in Figure 1, we do observe a drop in test scores after teacher cheating was uncovered in 2009. 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 or students in formerly investigated schools in 2010. Using test scores drops to identify cheating creates problems for our analysis of the impacts of cheating, however. As explained below, we rely on the first score a student receives after cheating ends to control for their true ability. Since test score drops are the difference between the last manipulated score and the first true score, including both a measure of test score drops (to identify having been cheated) and the first post-cheating test score (to measure student ability) in an analysis of future achievement would be akin to using manipulated scores to predict future (true) scores. Absent complete cheating, manipulated scores will be positively related to ability. Thus the relationship between being cheated (large test score drops) and future achievement would be biased upward. We confirmed this to be the case in preliminary analyses where we found a quantitatively large and statistically significant positive relationship between large test score drops from spring 2009 to spring 2010 and post-2010 test scores (holding constant spring 2010 test scores). We therefore do not employ test score drops to identify teacher cheating.

Rather than unusual answer patterns or large year-to-year 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 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, "having a bad 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 under-identification 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

[5] One notable exception is cases where a student initially marks their answers in 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.

were present in APS schools during the 2010 CRCT 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 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 also 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 CRCTs 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 ex-ante cheating in a number of ways. One method noted by the GBI 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 by grade 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 ex-ante 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 and above.[7] Thus, although we cannot reject the possibility that some ex-ante or contemporaneous cheating occurred in higher grades, we can be more confident that WTR erasure counts are a good measure of all forms of teacher cheating in grades 3 and above. Given the likelihood of more substantial ex-ante cheating in grades 1 and 2, when employing WTR erasure counts to identify cheating we conduct separate estimates of the impact of cheating on student outcomes for grades 3-8 and for grades 1-2.

We base our erasure-count measure of cheating on the distribution of WTR erasures on the spring 2013 exam (when all evidence suggests that cheating no longer existed). Table 1 shows the 90th, 95th and 99th percentiles of the distribution of WTR erasure counts in both the last year of pervasive cheating, 2009, and the last year of available erasure data, 2013. A student was designated as having been cheated in 2009 if the number of WTR erasures on a given exam exceeded the number of WTR erasures corresponding to the 95th percentile in 2013. Thus, for example, if a student's CRCT reading exam in 2009 had more than four WTR erasures, they were classified as having been cheated.

[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 less precise.

B. Differential Cheating and Controlling for Student Ability

Prior research on teacher cheating (e.g. Jacob and Levitt, 2003a, 2003b), as well as the GBI investigation, focused on identifying *classrooms* where cheating occurred in order to identify which teachers engaged in illicit behavior. In contrast, the focus of our analysis is on students rather than on teachers. The extent of teacher cheating can vary across students within classrooms, i.e. the impact of a cheating teacher will 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. 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. Evidence of such within-classroom variation is provided in Table 2, which shows the distribution of WTR erasures by subject within GBI-flagged classrooms in 2008/09. For both reading and ELA, over one-third of students in flagged classrooms had two or fewer WTR erasures and over one-fourth of students had only one or no WTR erasures on their reading exam. In math, the proportion of students with few WTR erasures was smaller, but still substantial. Nearly one-fourth of students in flagged classrooms had two or fewer WTR erasures on their math CRCT exam.

Because our interest is in evaluating how cheating impacted long-run student outcomes, we must evaluate the extent to which student ability played a role in selective cheating. If the students selected by a teacher for cheating were weaker students to begin with, a long-run achievement analysis which failed to account for such differences would be biased downward. Indeed, we find strong evidence that teachers selected which students to cheat based on their ability. Figure 3A shows the distribution of WTR erasure patterns in a cheating environment (2009 flagged classrooms) by quintile of student ability in math, where student ability is measured by 2010 CRCT scores. Figure 3B shows the distribution of WTR erasure in a normal environment (2013), broken down by measured student ability in math, where student ability is characterized by performance on the 2013 CRCT exam. In the cheating environment, the fewest erasures are for the highest ability students (Q1) with the distributions of WTR erasures progressively skewed to the right for students of lower apparent ability. In contrast, in the non-cheating environment, the quintile graphs are nearly coincident, save for slightly fewer erasures amongst the most able group (Q1). This is to be expected since higher ability students are likely to be more confident in their initial answers. Taken together, these results clearly suggest that teachers were more likely to change student answers for students of lower ability.

The challenge imposed by selective cheating can be addressed by appropriately controlling for student ability. If cheating on the 2009 exam were an isolated incident, we could rely on test scores in prior years as a measure of student ability; however, as discussed above, APS experienced widespread, long-term cheating which renders pre-2009 results unreliable. Two other alternatives exist. One is to rely on our 2009 estimate of initial right answers. In the

case of ex-post cheating, this method has the appeal of measuring ability immediately prior to the "treatment." However, because we are aware that some ex-ante and contemporaneous cheating existed, reliance on the initial right scores could lead to bias in our estimates.[8] A second option is to rely on 2010 test scores as a measure of ability. The appeal of this option is that we have greater confidence in its accuracy as the test observers led to a significantly lower incidence of cheating in 2010. However, the downside to this measure is that it limits the mechanisms we are able to evaluate. Our analysis can still measure the effects of students realizing they were cheated. However, we can no longer evaluate the effect of potential self-esteem boosts occurring in the 2009/10 school year as a result of 2008/09 cheating.

C. Choosing a Comparison Group

We are interested in the causal impact of teacher cheating on subsequent student outcomes. Put differently, how would a student have fared if their scores had not been manipulated? One possibility is to compare outcomes for a student before their scores were manipulated to their post-cheating outcomes. Such a within-student analysis would hold any time-invariant student characteristics constant. For outcomes that are observed in each period, such as attendance or disciplinary incidents, this could be accomplished by estimating models of student outcomes that include indicators for the cheated and post-cheated periods along with "student fixed effects" that create a separate baseline for each student. Unfortunately, as discussed above, in many instances teacher cheating occurred across multiple classrooms within a school over several years, meaning that for many students we never observe pre-cheating outcomes. Further, for one-time outcomes such as high school graduation, it is not possible to establish a within-student pre-cheating baseline, making this approach infeasible.

The other alternative is to compare outcomes for students who were cheated with those who were not cheated. This is potentially problematic, however, since non-cheated students may be different from cheated students in ways that are correlated with student outcomes and thus we could falsely attribute differences in outcomes to cheating which might be caused by other factors. For example, many schools in APS were not investigated by the GBI 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. These non-investigated schools tended to serve much lower proportions of students from low-income families than did investigated schools. 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 school leadership, school culture, or average teacher quality. Within-school comparisons can be made by estimating models of student outcomes that include school fixed effects. However, even within investigated schools, sorting of students across teachers (either due to parental preferences for teachers, teacher preferences for

[8] Another problem with using the initial right is that questions vary in difficulty and thus the raw number of initial right answers does not directly map into the true scale score for a student.

students or ability tracking of students) could result in significant differences in the underlying characteristics of students who attended classrooms where teacher cheating occurred vis-à-vis students attending non-cheating classrooms in the same school. It is also possible to compare cheated and non-cheated students in the same classroom by including classroom fixed effects, though non-cheated students are likely to be of higher ability than cheated students since there is less for a teacher to gain from manipulating already-high scores. Also, within-classroom comparisons would not measure any cheating behavior that affects all students within a classroom equally, so as a reduction in teacher effort. Further, making within-classroom comparisons reduces the number of students being directly compared and will therefore diminish the precision of any estimated effects.

For the cross-student comparisons we take a number of steps to minimize any bias associated with the non-random exposure of students to teacher cheating. First, we control for a variety of observable student characteristics, including gender, race/ethnicity, free/reduced-price lunch status, gifted status, limited English proficiency and disability status. In addition, we control for unmeasured ability by including the first not-cheated test score. For students who had been continuously cheated, the first non-manipulated score would be the revelation of "true" achievement for a student. As such, the score would not be influenced by potential loss in self-esteem. However, the first non-cheating score could be a downwardly biased measure of true ability if past teacher cheating led to either reduced teacher or student effort and achievement has persistent effects over time.

D. Econometric Model

The decisions described above regarding the issues of identifying cheating, differential cheating by teachers, controlling for student ability and constructing a valid comparison group lead us to the following empirical model of student achievement:

$$\boldsymbol{A_{i,2009+t} = \beta_0 + \delta(Cheat)_{2009} + \lambda A_{i,2010} + \theta X'_{i,2009+t} + \varepsilon_{i,2009+t}}$$

where $A_{i,2009+t}$, is achievement level for student i, t years after 2009, *Cheat* equals one if the individual was cheated in 2009 and zero otherwise, $X'_{i,t+1}$ is a vector of student and family specific characteristics such as race, gender, lunch status, gifted status, Limited English Proficiency, Special Education status and $\varepsilon_{i,2009+t}$ is the idiosyncratic error term. The main parameter of interest is δ. We estimate three variants of the model, one with no fixed effects (as specified above), a second with school fixed effects that produce within-school comparisons (i.e. with an additional term γ_k where k indexes schools) and a third with classroom-level fixed effects that produce within-classroom comparisons (ie. with an additional term ϕ_j, where j indexes classrooms).

5. Results

A. Descriptive Statistics

In Table 3 we provide descriptive statistics for students, broken down by school-type of enrollment in 2008/09. Investigated schools enrolled over one-fourth of the student population. Relative to non-investigated schools, investigated schools serve 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).

In Table 4 we present a tabulation of the number of students who were enrolled in flagged classrooms in 2008/09 by their continued enrollment in APS in subsequent years, broken down by their grade level in 2008/09. It is clear that the vast majority of flagged classrooms were in elementary schools (grades 1-5). Further, while there was some attrition after the "true" scores were revealed in 2010, there was not a large exodus of students from the district. Few of the students who were in grades 7 and 8 in 2008/09 remain in APS in fall 2014 since their cohorts would have graduated (assuming normal progress and on-time graduation at the end of 12th grade). For students who were in grades 1-8 in 2008/09 about 50-60 percent of them remain in APS in fall 2014. The proportion remaining tends to be higher for younger cohorts (around 60 percent for the 2008/09 1st-grade cohort) than for older cohorts (slightly less than 50 percent for the 2008/09 cohort), which is likely due to higher attrition as a result of dropouts in high school. A total of 5,888 students who were in classrooms flagged for extraordinarily high WTR erasures based on 2009 test results remain in APS in fall 2014.

Table 5 presents a breakdown of enrollment in APS traditional and charter schools by year for those students who were in a flagged classroom in 2008/09. There is some evidence that the cheating scandal lead to a movement away from traditional public schools and into the charter sector. There are fairly large increases in enrollment in charter schools in the two years after the revelation of teacher cheating, 2009/10 and 2010/11. After the 2010/11 school year the absolute enrollment in charters stabilizes, though the relative proportion continues to increase.

Table 6 provides a tabulation of the fall 2014 grade level for students who were in a flagged classroom in 2008/09 by their 2008/09 grade level. For each 2008/09 cohort, the majority of students are on-track, enrolled in a grade that is six levels above their enrollment grade in 2008/09. Of the 5,888 who were in classrooms flagged by the GBI in 2008/09 and are enrolled in an APS school in fall 2014, a bit less than half (2,477 of 5,888 or 42 percent) have not yet reached high school.

Table 7 presents a tabulation of enrollment in APS by year by the level of WTR erasures on the 2009 CRCT. The first row indicates enrollment by students defined as "cheated" based on having more WRT erasures on the 2009 CRCT in any of three subjects (reading, math or ELA)

than the 95th percentile of WRT erasures in a non-cheating environment (the 2013 CRCT). Based on this definition, 7,064 students had their exam answers manipulated; 3,728 of which were still attending APS schools in fall 2014. The second and third rows student enrollment counts provide alternative benchmarks, 5 or more WRT erasures only any of the three subject exams or 10 or more erasures on any of the three subject exams.

B. Estimated Effects of Teacher Cheating on Student Outcomes

Our primary results are presented in Tables 8A – 8C. The tables report estimates of δ in equation (1), which represents the relationship between being a "cheated" student (i.e. having a number of WTR erasures in 2009 that exceeds the 95th percentile for WTR erasures in 2013) and subsequent normed test scores, holding constant observable student characteristics and the student's 2010 normalized test score. Each row represents estimates from a specification with varying levels of fixed effects. Each column represents the estimate of δ on achievement from a different year. Results are presented for students in grades 1 and 2 in 2008/09 (where contemporaneous teacher cheating was likely more prevalent), grades 3-6 in 2008/09 and all grades 1-6 together.

In Table 8A, which reports results for math, we find consistent negative effects of being cheated on future test scores for students who were in grades 1 and 2 in 2008/09. The effects range from -0.06 to -0.10 standard deviations and are fairly constant over time. This is on par with the effect of having a rookie teacher rather than one with five years of experience (Clotfelter, et al., 2006). Alternatively, the effect is equivalent to 18 to 29 percent of the average annual learning gain in math for middle school students of 0.34 (Lipsey, et al., 2012). The model that includes classroom fixed effects, that is the one that compares students within a classroom, produces the largest effects. However, the results do not change dramatically with school fixed effects or even with no fixed effects at all. Results for students who were in grades 3-6 in 2008/09 are quite different. Save for one positive and significant coefficient, the remaining 11 estimates are all insignificantly different from zero. When all students are grouped together, the point estimates are all negative, but only estimates for 2013/2014 are significantly different from zero across all three specifications. For 2013/14 the negative effect of having been cheated in 2008/09 is estimated to be in the range of -0.04 to -0.07 standard deviations or about 12 to 21 percent of annual achievement gains for middle school students in math.

In contrast to math, results for reading achievement, which are displayed in Table 8B, are almost uniformly negative and statistically significant. The estimated magnitudes of the effect of being cheated for students who were in grades 1 and 2 are substantially larger than the size of those for students who were in grades 3-6 in 2008/09. When all grades are combined, the estimated effect of being cheated on reading test scores in future years falls in the range of -0.06 to -0.14, which is equivalent to 22 to 52 percent of annual learning gains in reading for middle school students (which equal 0.27). The largest estimates once again come

from the models which make within-classroom comparisons, i.e. the models with classroom fixed effects.

Estimates for ELA, presented in Table 8C, are similar to those for reading. Estimated impacts of being cheated are almost always negative and statistically significant. Results for grades 1 and 2 are larger in magnitude than for students who were in grades 3-6 during the last year of widespread teacher cheating, 2008/09. Combining all grades, the estimated impact of being cheated on student achievement is in the range of -0.06 to -0.12 standard deviations.

In addition to estimating effects on later student achievement, we also estimated versions of equation (1) that replaced future test scores with measures of attendance and behavior. For these alternative outcomes we define being cheated as having unusually high WTR erasures (greater than the 95th percentile of the 2013 distribution) in either math, reading or ELA. Similarly, rather than include only a single-subject 2010 test score as a control for student ability, we include 2010 scores in math, reading and ELA. We also include 2009 values of the outcome variable (attendance or behavior) as a control. Finally, in the equation estimating the number of disciplinary incidents we include grade level indicators as control since disciplinary incidents to be much higher in middle and high school than in elementary school. Otherwise, the specifications are identical to those for the analysis of test scores.

Table 9 presents estimates of the impact of being cheated on the percent of days in attendance in each of the years 2010/11 through 2013/14. In 2010/11 (the first year following the revelation of true scores) there appears to be no effect; all of the estimated coefficients are positive and statistically insignificant. However, in later years the point estimates are almost always negative and often are significantly different from zero. It is important to remember that in the latter years students who were in tested grades (1-8) in 2008/09 are mostly all in middle and high school, where absenteeism may be more directly a matter of student choice. However, even if we focus on the results that are statistically significantly different from zero, the estimates are quite small in magnitude, on the order of 0.3 to 0.8 percentage points or about 1 school day out of the typical 180-day school year.

Finally, in Table 10, we present estimates of the impact of being cheated in 2008/09 on the number of disciplinary incidents in years 2010/11 through 2013/14. There appears to be little in the way of effects of teacher cheating on subsequent student behavior. While point estimates are mostly positive, in only one case are they significantly different from zero at standard confidence levels.

6. Summary and Conclusions

Much effort has been devoted to identifying the teachers and administrators responsible for manipulating test scores in APS and bringing those responsible to justice. At the same time little is known about the victims of the cheating scandal. This report represents the

first attempt to rigorously analyze the impact of teacher cheating on the long-run outcomes of students.

In conducting the analysis we faced several challenges, including identifying which students were cheated, deriving a measure of the students' true ability and coming up with a reasonable counterfactual group for comparison. Given the available data and the history of test manipulation, we chose to identify "cheated students" as those who had high numbers of wrong-to-right erasures on the 2009 CRCT exam and used their 2010 CRCT scores as a measure of their true ability. Based on these choices, we estimated the effect of being cheated on later test scores, attendance and behavior of students. Comparisons were made between cheated students and non-cheated students generally, between cheated and non-cheated students who attended the same school in 2008/09 and between cheated and non-cheated students within the same classroom in 2008/09.

Our results indicate that being cheated had negative consequences for later student performance in both reading and ELA. The estimated impacts are in the range of one-fourth to one-half of the average annual achievement gain for a middle school student. This is equivalent to one to two times the difference between having a rookie teacher and one with 5 or more years of experience in a single year. In contrast, the effects of teacher cheating for subsequent achievement in math are mixed. There is little evidence that teacher cheating had any deleterious effects on subsequent student attendance or student behavior. Any impacts that may have occurred were very small.

References

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

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

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

Dee, Thomas and James Wyckoff (2013). "Incentives, Selection and Teacher Performance" Evidence from IMPACT." Washington DC: National Bureau of Economic Research, NBER Working Paper No. 19529.

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

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.

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

Lipsey, Mark W., Kelly Puzio, Cathy Yun, Michael A. Hebert, Kasia Steinka-Fry, Mikel W. Cole, Megan Roberts, Karen S. Anthony, and Matthew D. Busick (2012). "Translating the Statistical Representation of the Effects of Education Interventions into More Readily Interpretable Forms." Washington, DC: National Center for Special Education Research, Institute of Education Sciences, U.S. Department of Education.

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.

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.

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.

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

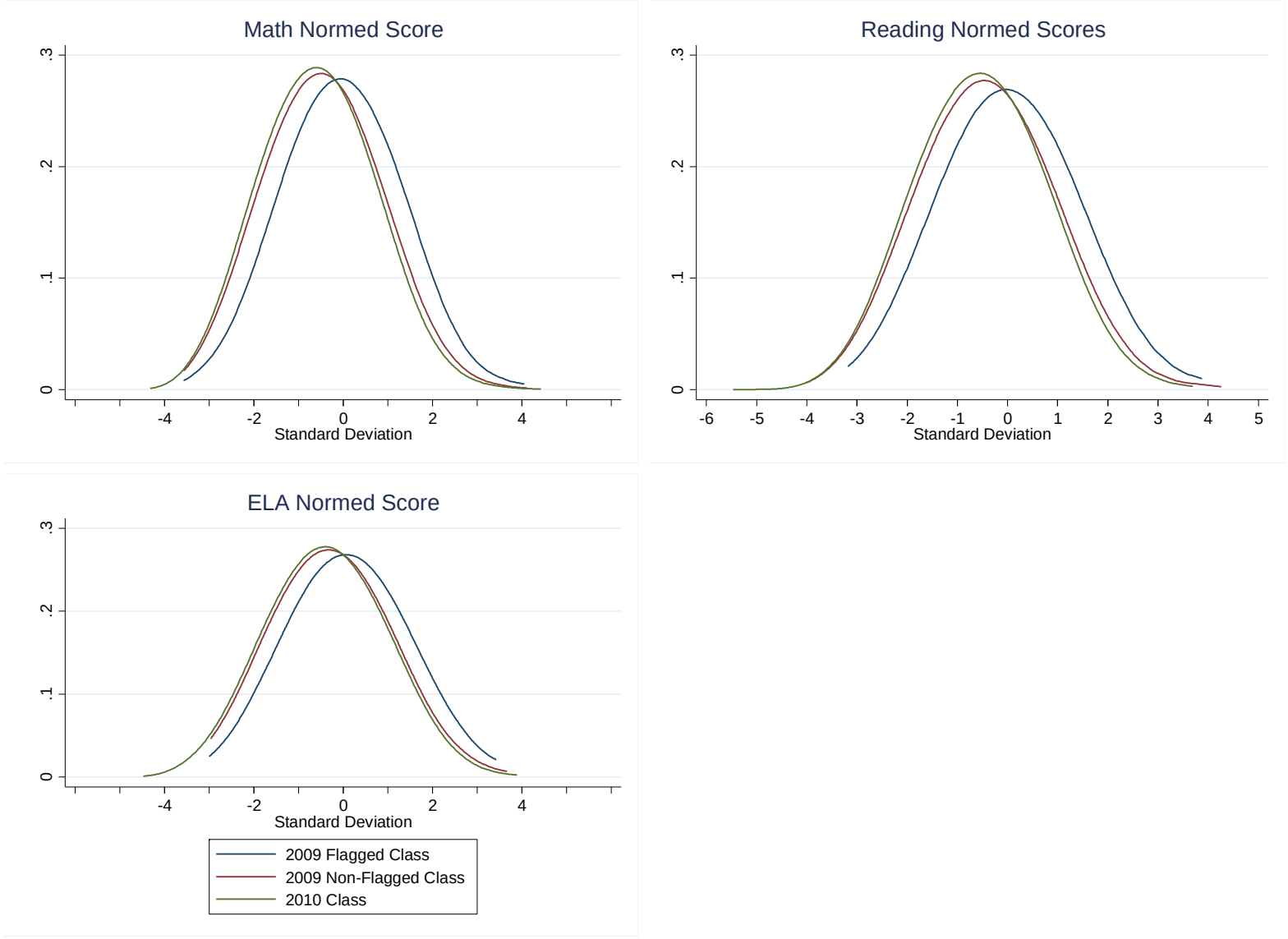
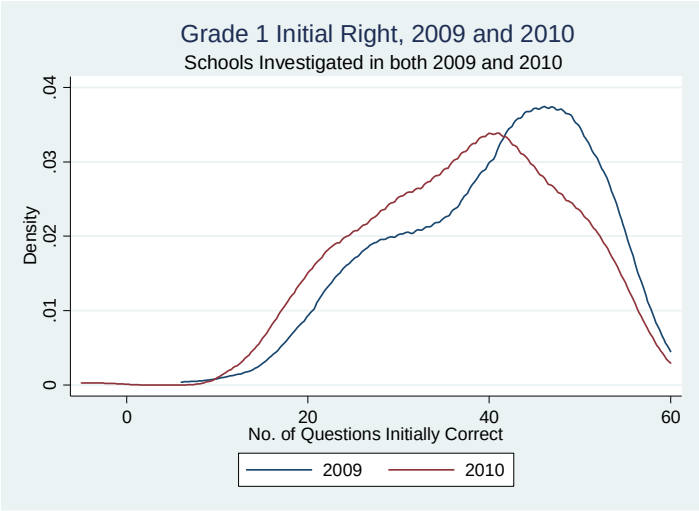
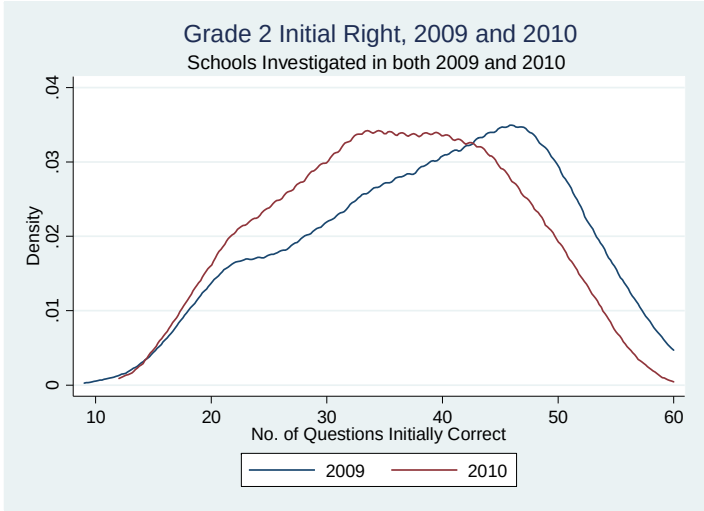
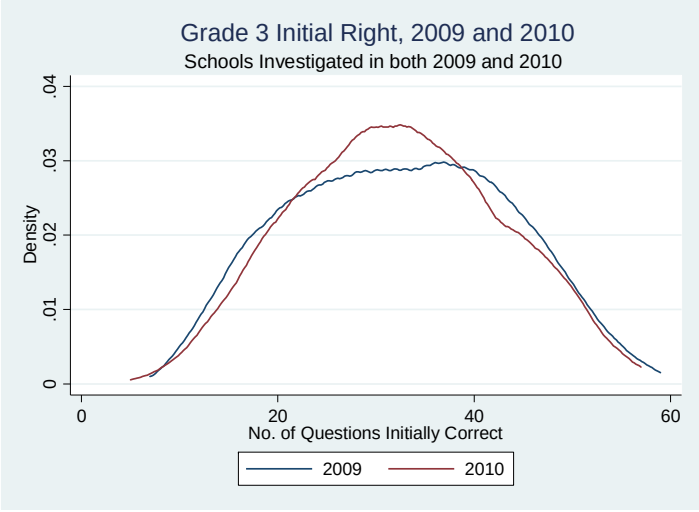
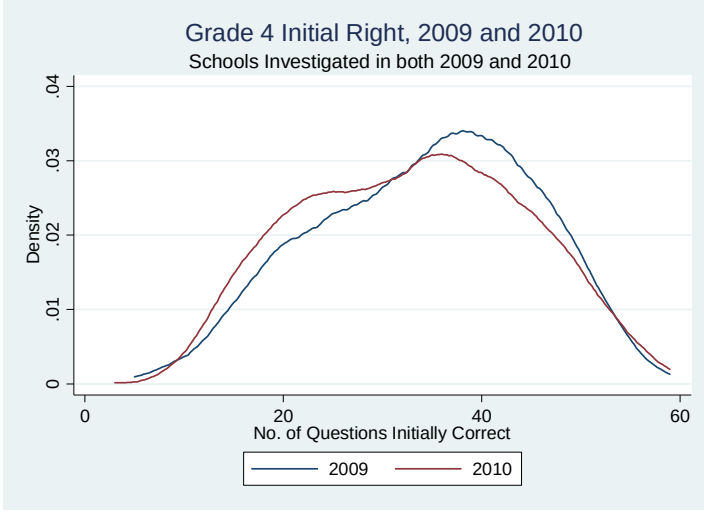


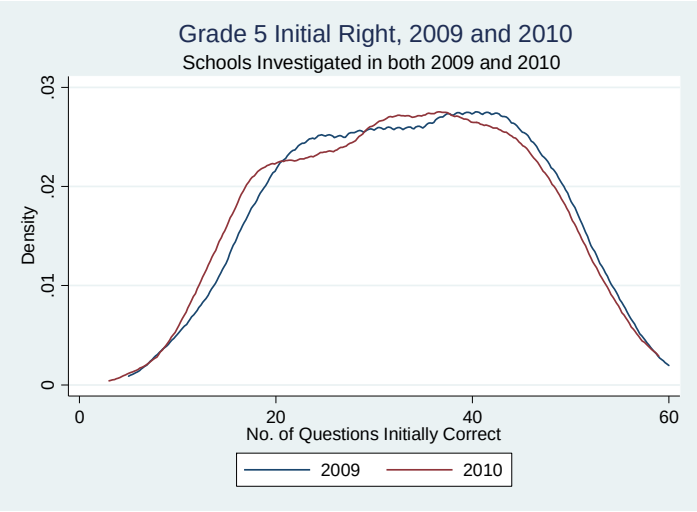
Figure 2: Distributions of the Number of "Initial Right" Answers by Grade, 2009 and 2010

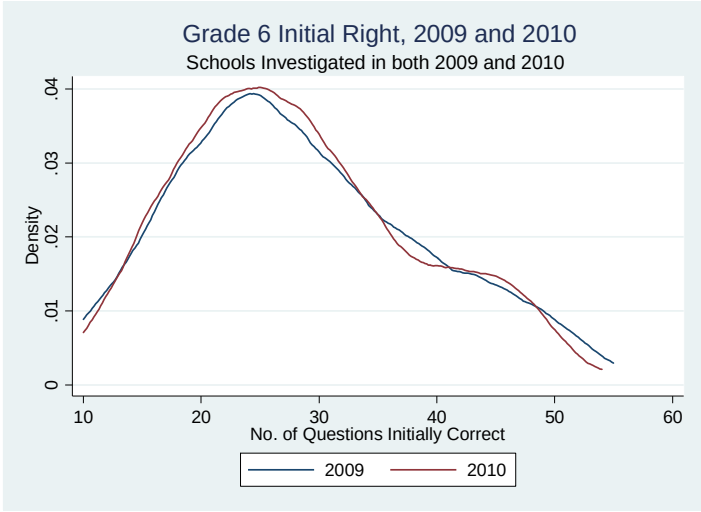


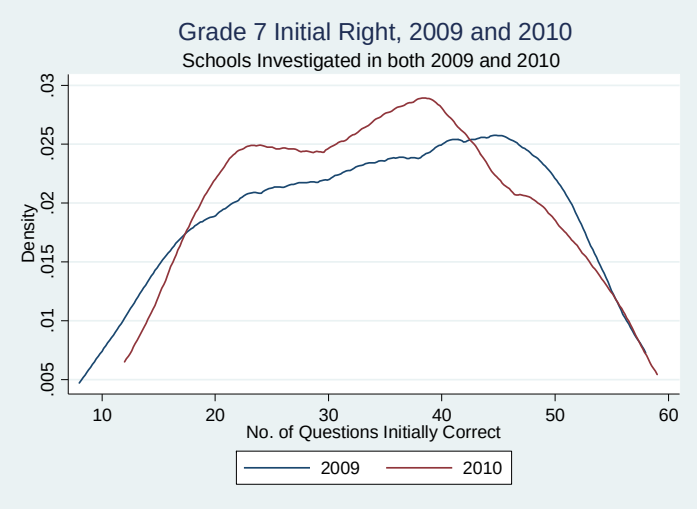







Grade 5 Initial Right, 2009 and 2010
Schools Investigated in both 2009 and 2010
Density
0
.01
.02
.03
0
20
40
60
No. of Questions Initially Correct
2009
2010


Grade 6 Initial Right, 2009 and 2010
Schools Investigated in both 2009 and 2010
Density
0
.01
.02
.03
.04
10
20
30
40
50
60
No. of Questions Initially Correct
2009
2010


Grade 7 Initial Right, 2009 and 2010
Schools Investigated in both 2009 and 2010
Density
.005
.01
.015
.02
.025
.03
10
20
30
40
50
60
No. of Questions Initially Correct
2009
2010

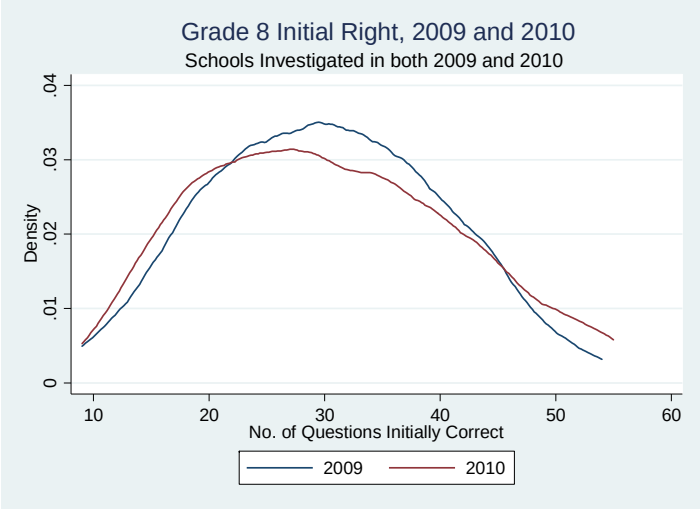

Grade 8 Initial Right, 2009 and 2010
Schools Investigated in both 2009 and 2010
Density
0
.01
.02
.03
.04
10
20
30
40
50
60
No. of Questions Initially Correct
2009
2010

Figure 3A: Distribution of 2009 Wrong-to-Right Erasures by 2010 Achievement Quintile in Math (Students in a flagged classroom in 2008/09 and not in a flagged classroom in 2009/10)

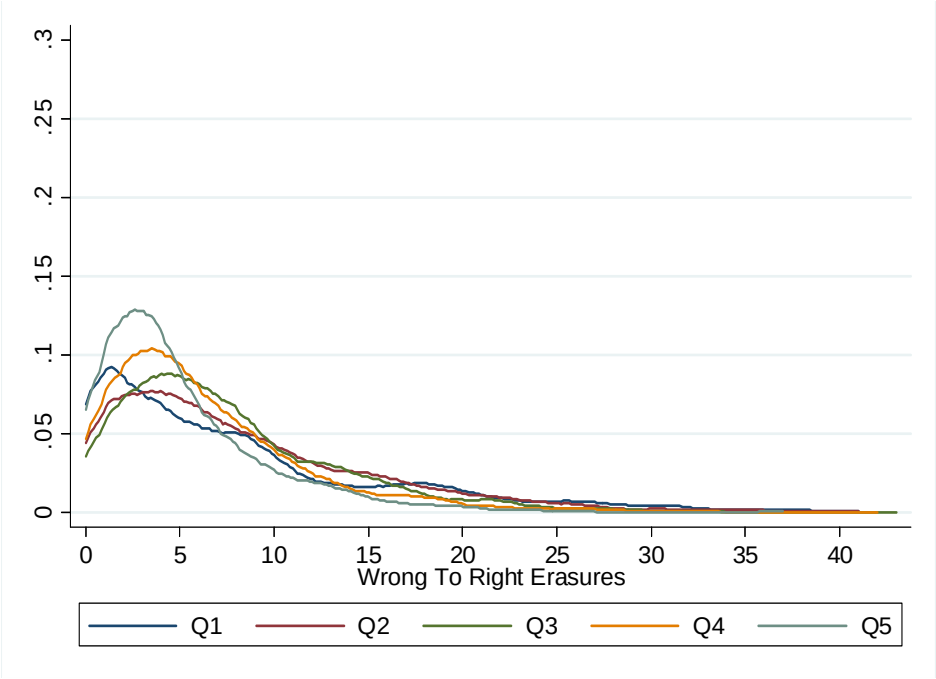


Figure 3B: Distribution of 2013 Wrong-to-Right Erasures by 2013 Achievement Quintile in Math (Students who attended a school that was flagged in 2008/09)

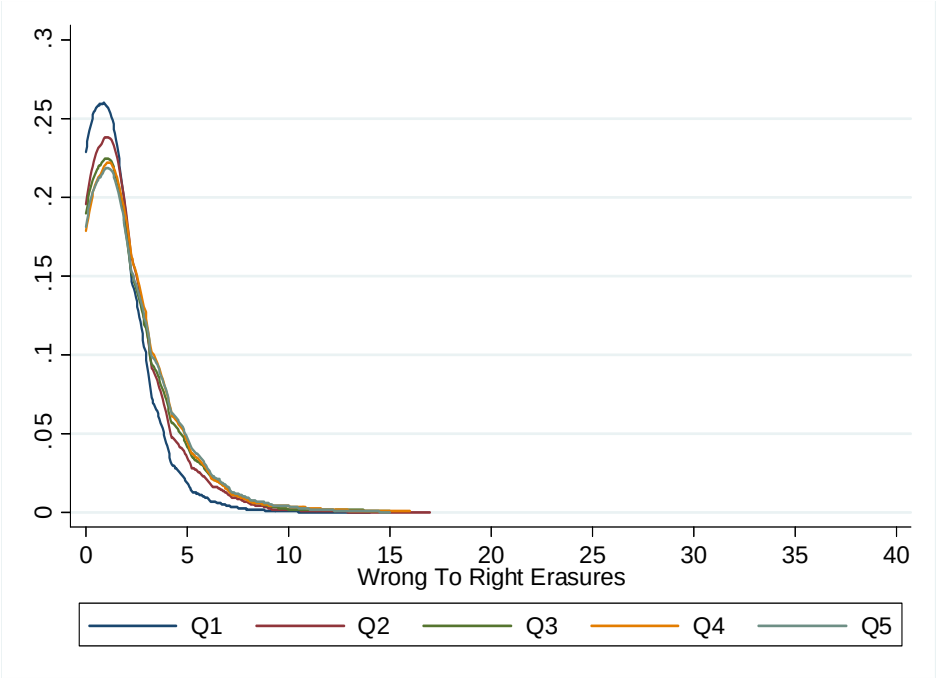


Table 1. 90th, 95th and 99th Percentiles of the Wrong-to-Right Erasure Count on 2009 and 2013 CRCT Exams in Reading, ELA and Math

	2009 CRCT			2013 CRCT		
	Reading	ELA	Math	Reading	ELA	Math
90th Percentile	8	8	12	3	3	4
95th Percentile	12	12	16	4	4	5
99th Percentile	19	19	26	6	8	9

Table 2. Wrong-to-Right Erasure Count Frequencies for CRCT Reading, ELA and Math in Flagged Classrooms in Investigated Schools in 2008/09

No. of WTR Erasures	Reading		ELA		Math	
	Number	Percent	Number	Percent	Number	Percent
0	1,038	12.23	865	10.50	659	7.28
1	1,113	13.11	945	11.47	791	8.73
2	1,061	12.50	978	11.87	800	8.83
3	917	10.80	876	10.63	857	9.46
4	741	8.73	796	9.66	757	8.36
5	621	7.32	663	8.05	725	8.01
6	523	6.16	583	7.08	577	6.37
7	412	4.85	449	5.45	523	5.78
8	349	4.11	374	4.54	492	5.43
9	281	3.31	313	3.80	419	4.63
10	243	2.86	259	3.14	332	3.67
11	214	2.52	192	2.33	286	3.16
12	188	2.22	165	2.00	252	2.78
13	150	1.77	146	1.77	208	2.30
14	118	1.39	127	1.54	190	2.10
15	99	1.17	99	1.20	167	1.84
More than 15	419	4.94	410	4.98	1021	11.27
Total	8,487	100.00	8,240	100.00	9,056	100.00

Table 3. Summary Statistics for APS Schools by Investigated School Status in 2008/09

	All Schools		Investigated Schools		Non-Investigated Schools	
	Mean	S.D.	Mean	S.D.	Mean	S.D.
Black	0.83	0.37	0.95	0.22	0.77	0.42
White	0.10	0.29	0.01	0.07	0.14	0.35
Hispanic	0.05	0.21	0.03	0.18	0.05	0.23
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
Gifted Served	0.08	0.27	0.06	0.23	0.09	0.28
Early Intervention Program	0.15	0.35	0.25	0.43	0.10	0.30
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
Observations	54,356		18,542		35,814	

Table 4. Enrollment in APS by Year by 2008/09 Grade Level for Students in One or More Flagged Classrooms in 2008/09

Grade Level in 2008/09	Enrollment in APS by Year						
	2008/09	2009/10	2010/11	2011/12	2012/13	2013/14	Fall 2014
1	2068	1901	1709	1581	1465	1310	1210
2	2018	1837	1647	1516	1332	1238	1154
3	2001	1829	1642	1401	1358	1303	1167
4	1822	1685	1476	1375	1307	1232	1084
5	1611	1383	1278	1197	1106	1046	895
6	706	642	582	535	486	412	338
7	588	538	465	423	370	297	29
8	739	664	603	525	454	46	11
Total	11553	10479	9402	8553	7878	6884	5888

Table 5. Enrollment in APS by Year and School Type for Students in One or More Flagged Classrooms in 2008/09

School Type	Enrollment in APS by Year						
	2008/09	2009/10	2010/11	2011/12	2012/13	2013/14	Fall 2014
Traditional	11285	10038	8796	7898	7187	6203	5274
Charter	268	441	606	655	691	681	614
Total	11553	10479	9402	8553	7878	6884	5888

Table 6. Enrollment in APS by Grade Level in Fall 2014 by 2008/09 Grade Level for Students in One or More Flagged Classrooms in 2008/09

Grade Level in 2008/09	Enrollment in APS in Fall 2014 by Grade Level								
	5	6	7	8	9	10	11	12	Total
1	11	197	1000	1	1	0	0	0	1210
2	0	8	153	992	1	0	0	0	1154
3	0	0	5	101	1058	3	0	0	1167
4	0	0	1	7	491	580	5	0	1084
5	0	0	0	1	120	264	508	2	895
6	0	0	0	0	13	30	69	226	338
7	0	0	0	0	1	3	9	16	29
8	0	0	0	0	0	0	1	10	11
Total	11	205	1159	1102	1685	880	592	254	5888

Table 7. Enrollment in APS by Year by Erasure Counts in 2009 for Students in Flagged Classrooms in 2008/09

2009 WRT Erasure Measure	Enrollment in APS by Year						
	2008/09	2009/10	2010/11	2011/12	2012/13	2013/14	Fall 2014
> 95th Percentile of WTR Erasures in 2013 in any Subject	7064	6437	5810	5268	4872	4340	3728
5 or More WTR Erasures in Any Subject	7125	6485	5852	5313	4901	4348	3728
10 or More WTR Erasures in Any Subject	3473	3150	2877	2611	2421	2156	1825
All Students in Flagged Classrooms	11553	10479	9402	8553	7878	6884	5888

Table 8A: Estimated Effect of Being Cheated in 2008/09 on Normed Math Test Scores in 2010/11 – 2013/2014 by 2008/09 Grade Level

	Grades 1&2				Grades 3-6				Grades 1-6			
Model	2010/11	2011/12	2012/13	2013/14	2010/11	2011/12	2012/13	2013/14	2010/11	2011/12	2012/13	2013/14
No Fixed Effects	-0.0686** (0.0206)	-0.0680** (0.0218)	-0.0657** (0.0241)	-0.0830** (0.0233)	0.0053 (0.0179)	0.0160 (0.0168)	0.0290 (0.0229)	-0.0111 (0.0341)	-0.0192 (0.0137)	-0.0209 (0.0134)	-0.0184 (0.0167)	-0.0555** (0.0195)
School Fixed Effects	-0.0610** (0.0195)	-0.0648** (0.0212)	-0.0612* (0.0243)	-0.0688** (0.0224)	0.0002 (0.0150)	0.0204 (0.0153)	0.0433* (0.0219)	0.0061 (0.0328)	-0.0154 (0.0123)	-0.0146 (0.0125)	-0.0110 (0.0161)	-0.0405* (0.0184)
Classroom Fixed Effects	-0.0882** (0.0234)	-0.0923** (0.0256)	-0.0622* (0.0307)	-0.0997** (0.0296)	-0.0088 (0.0211)	0.0096 (0.0209)	-0.0056 (0.0297)	-0.0188 (0.0446)	-0.0405* (0.0159)	-0.0364* (0.0163)	-0.0351 (0.0215)	-0.0717** (0.0247)

Standard errors clustered at the classroom level in parentheses. *Significant at the 5% level, **significant at the 1% level. Note: 2012, 2013, and 2014 results are restricted to grades 1-5, 1-4, and 1-3, respectively

Table 8B: Estimated Effect of Being Cheated in 2008/09 on Normed Reading Test Scores in 2010/11 – 2013/2014 by 2008/09 Grade Level

	Grades 1&2				Grades 3-6				Grades 1-6			
Model	2010/11	2011/12	2012/13	2013/14	2010/11	2011/12	2012/13	2013/14	2010/11	2011/12	2012/13	2013/14
No Fixed Effects	-0.1419** (0.0269)	-0.1955** (0.0323)	-0.1858** (0.0371)	-0.1863** (0.0401)	-0.0701** (0.0169)	-0.0762** (0.0196)	-0.0491* (0.0239)	-0.0374 (0.0373)	-0.0798** (0.0142)	-0.0936** (0.0165)	-0.0612** (0.0205)	-0.1076** (0.0268)
School Fixed Effects	-0.1449** (0.0273)	-0.1919** (0.0315)	-0.1775** (0.0370)	-0.1705** (0.0399)	-0.0681** (0.0166)	-0.0685** (0.0193)	-0.0440 (0.0235)	-0.0355 (0.0365)	-0.0782** (0.0142)	-0.0978** (0.0164)	-0.0633** (0.0204)	-0.0932** (0.0268)
Classroom Fixed Effects	-0.1888** (0.0337)	-0.2195** (0.0377)	-0.2536** (0.0462)	-0.2164** (0.0515)	-0.0977** (0.0229)	-0.1045** (0.0266)	-0.0664* (0.0315)	-0.0326 (0.0527)	-0.1263** (0.0190)	-0.1409** (0.0217)	-0.1386** (0.0271)	-0.1334** (0.0370)

Standard errors clustered at the classroom level in parentheses. *Significant at the 5% level, **significant at the 1% level. Note: 2012, 2013, and 2014 results are restricted to grades 1-5, 1-4, and 1-3, respectively

Table 8C: Estimated Effect of Being Cheated in 2008/09 on Normed ELA Test Scores in 2010/11 – 2013/2014 by 2008/09 Grade Level

	Grades 1&2				Grades 3-6				Grades 1-6			
Model	2010/11	2011/12	2012/13	2013/14	2010/11	2011/12	2012/13	2013/14	2010/11	2011/12	2012/13	2013/14
No Fixed Effects	-0.1119** (0.0210)	-0.1053** (0.0252)	-0.1078** (0.0260)	-0.1326** (0.0258)	-0.0380* (0.0167)	-0.0707** (0.0186)	-0.0365 (0.0236)	-0.0578 (0.0338)	-0.0668** (0.0131)	-0.0887** (0.0150)	-0.0664** (0.0176)	-0.1145** (0.0204)
School Fixed Effects	-0.1051** (0.0205)	-0.1013** (0.0241)	-0.0984** (0.0250)	-0.1224** (0.0252)	-0.0406** (0.0159)	-0.0669** (0.0184)	-0.0260 (0.0233)	-0.0644 (0.0335)	-0.0645** (0.0127)	-0.0828** (0.0145)	-0.0603** (0.0173)	-0.1090** (0.0201)
Classroom Fixed Effects	-0.1421** (0.0268)	-0.1191** (0.0300)	-0.1407** (0.0351)	-0.1503** (0.0316)	-0.0426* (0.0210)	-0.0862** (0.0236)	-0.0319 (0.0309)	-0.0567 (0.0425)	-0.0828** (0.0165)	-0.1020** (0.0185)	-0.0824** (0.0237)	-0.1190** (0.0253)

Standard errors clustered at the classroom level in parentheses. *Significant at the 5% level, **significant at the 1% level. Note: 2012, 2013, and 2014 results are restricted to grades 1-5, 1-4, and 1-3, respectively

Table 9. Estimated Effect of Being Cheated in 2008/09 on Percent Attendance 2010/11 – 2013/2014

Model	2010/11	2011/12	2012/13	2013/14
No Fixed Effects	0.0781 (0.1124)	-0.2632 (0.1835)	-0.3145 (0.2299)	-0.7764** (0.2627)
School Fixed Effects	0.0021 (0.0916)	-0.4200** (0.1482)	-0.3878* (0.1959)	-0.6880** (0.2468)
Classroom Fixed Effects	0.0200 (0.1045)	-0.0781 (0.1816)	0.0603 (0.2166)	-0.2992 (0.2850)

Standard errors clustered at the classroom level in parentheses. *Significant at the 5% level, **significant at the 1% level.

Table 10. Estimated Effect of Being Cheated in 2008/09 on Discipline Incidents 2010/11 – 2013/2014

Model	2010/11	2011/12	2012/13	2013/14
No Fixed Effects	0.0070 (0.0408)	-0.0098 (0.0634)	0.1149* (0.0549)	0.0677 (0.0539)
School Fixed Effects	-0.0138 (0.0382)	-0.0847 (0.0597)	0.0038 (0.0561)	-0.0441 (0.0583)
Classroom Fixed Effects	-0.0054 (0.0463)	-0.0629 (0.0770)	0.0505 (0.0733)	-0.0232 (0.0749)

Standard errors clustered at the classroom level in parentheses. *Significant at the 5% level, **significant at the 1% level.