The Long-term Consequences of Teacher Discretion in Grading of High-stakes Tests*

Rebecca Diamond† Petra Persson‡

This Draft: October 2, 2017
First Draft: February 28, 2016

Abstract

We examine the long-term consequences of teacher discretion in grading of high-stakes tests. Bunching in Swedish math test score distributions reveal that teachers inflate students who have "a bad test day," but do not to discriminate based on immigrant status or gender. By developing a new estimator, we show that receiving a higher grade leads to far-reaching educational and earnings benefits. Because grades do not directly raise human capital, these results emphasize that grades can signal to students and teachers within the educational system, and suggest important dynamic complementarities between students' effort and their perception of their own ability.

*A previous version of this paper was circulated under the title "The Long-term Consequences of Grade Inflation." We are grateful to Raj Chetty, John Friedman, Matt Gentzkow, Caroline Hoxby, Guido Imbens, Eddie Lazear, Paul Oyer, Luigi Pistaferri, Kathryn Shaw, Alan Sorensen, Chris Taber, and seminar and conference participants at Boston College, Brown, Cornell, Gothenburg, Houston, Michigan, the Minneapolis Fed, NYU, Oslo, Santa Clara, Stanford, UC Berkeley, UC Davis, UC San Diego, UC Santa Cruz, SIEPR, Stockholm, the Swedish Institute for Social Research, Uppsala, Wisconsin, Yale, the 2016 Applied Microeconomics Jamboree at Duke, the 2015 All California Labor Conference at UCLA, the 2016 Utah Winter Business Economics Conference, and the 2015 Western Economic Association International for helpful comments. We are especially grateful to Bjorn Ockert, Jonas Vlachos, and Olof Aslund. Persson gratefully acknowledges funding from the Jan Wallander and Tom Hedelius Foundation. All remaining errors are our own.

†Stanford University & NBER. Email: diamondr@stanford.edu.

‡Stanford University, NBER, CEPR, IFAU, & IFN. Email: perssonp@stanford.edu.

1 Introduction

The increased reliance on standardized testing in educational systems around the world has gradually reduced teachers' influence over students' grades. China, India, Israel and Japan all use standardized tests, graded without any teacher discretion, to determine university admissions and entrance into prestigious civil service occupations. In contrast, college admissions in the US base student achievement on a mix of standardized measures (such as SAT or ACT exams) and those which contain discretionary evaluations by teachers (GPA and recommendation letters). Indeed, there is current debate about the merits of standardized testing in the US and whether the large emphasis that No Child Left Behind places on standardized testing should be reassessed.

In this paper, we study a unique context of discretionary grading in Sweden that allows us to measure when teachers use discretion in their grading of nationwide math tests. This discretion essentially enables teachers to selectively manipulate students' test scores by "bumping up" certain students over key grade cutoffs. We analyze the consequences of such test score manipulation using administrative population-level data that enables us to follow the universe of students in Sweden from before high school, throughout adolescence, and into adulthood, all the while tracking key educational, economic, and demographic outcomes.

Allowing teacher discretion in measuring student achievement raises two key questions. First, who benefits from discretionary test score manipulation? In particular, do teachers act in a corrective fashion, by raising scores of students who failed although they ought to have passed; or do they use their discretion to discriminate based on factors such as gender or ethnicity? Second, and even more crucially, does discretionary test score manipulation matter, in the sense that it conveys real, long-term economic gains? This is a priori unclear, since test score manipulation gives a student a higher test grade *without raising knowledge*. In order for this to have any effect, grades per se must matter for long-term outcomes. In this paper, we examine both which students benefit from manipulation, and how test score manipulation matters in the long-run.

We start by documenting extensive test score manipulation in the nationwide math tests taken in the last year before high school, by showing significant bunching in the distribution of test scores just above two discrete grade cutoffs. We model teachers' incentives to manipulate students' grades and show that if manipulation occurs, then it is concentrated in two parts of the test score distribution, in the vicinity of each of the two test score thresholds. We then employ a novel methodology to recover the un-manipulated test score distribution that allows us to *estimate* where manipulation begins and ends[1], and find that this varies substantially:

[1] Existing bunching estimators have relied on visual inspection for determining where manipulation begins

in some places, students' test scores are not manipulated at all; in others, students who lack as much as seven test score points are bumped up. Moreover, even within a given test score point, teachers treat students differently.

We analyze the characteristics of the students who are selectively inflated by teachers and find that teachers act in a corrective fashion, by being more likely to inflate students who have "a bad test day." In particular, teachers appear to use their discretion to undo idiosyncratic performance drops below what would be expected from each student's previous performance. Teachers do not selectively inflate based on gender, immigrant status, or whether the student has a stay at home parent who potentially might have more free time to pressure teachers into higher grades.

We then analyze the consequences of receiving test score manipulation. To do this, we cannot simply compare the (outcomes of) students whose test scores were manipulated with students whose tests scores were left un-manipulated – our first set of results show that students who are chosen to receive test score manipulation are different from students who are not. To overcome this issue and identify the effect of test score manipulation on students' longer-term outcomes, we develop a Wald estimator that builds on key ideas in the bunching literature. So far, bunching strategies have been used chiefly to analyze *the distribution that is being manipulated*: distributions of reported incomes (Saez, 2010; Chetty et al., 2011; Kleven and Waseem, 2012) or dates of marriage (Persson, 2014), for example. In each case, bunching methodologies have been used to predict how the manipulated distribution would have looked, had there been no manipulation, by using the un-manipulated parts of the distribution to help "fill in" the shape inside the manipulation regions. The key underlying assumption is that, in the absence of manipulation, the distribution would follow the polynomial that can be estimated from the unmanipulated parts of the distribution.

The core idea in this paper is to use a bunching methodology to examine the impact of manipulation on *other variables than the one that is being directly manipulated*, that is, other outcomes than the test score distribution.[2] Intuitively, just like we can plot the test score density, we can plot the mean of a future outcome, such as earnings, by test score. In the two test score ranges where manipulation occurred (in the vicinity of the Pass and PwD thresholds, respectively), the observed earnings distribution *partly captures the impact of test score manipulation on earnings*. In contrast, in the test score ranges where no manipulation occurred, the observed earnings captures the underlying relationship between (un-manipulated) test scores and earnings. Thus, we can use the relationship between

and/or ends (Saez, 2010; Chetty et al., 2011; Kleven and Waseem, 2012).

[2] This type of idea has related, although distinct, predecessors in the literature. Chetty et al. (2013) use bunching as an "instrument" (but implemented as a reduced form) for information about tax incentives, and Marx (2015) uses bunching to estimate causal effects. See Kleven (2016) for a discussion.

earnings and test scores estimated outside of the two manipulation regions to predict a counterfactual relationship between *earnings* and test scores inside the test score ranges where manipulation occurred. This, coupled with the counterfactual test score distribution – which gives the "head count" of students at each test score point in the absence of manipulation – allows us to calculate the expected earnings, for students inside each manipulation region, in a counterfactual scenario without manipulation.

The difference between these counterfactual earnings and the average earnings observed in the data captures the reduced form impact of (potentially being exposed to) test score manipulation on earnings. Our Wald estimator scales these excess earnings within the bunching region by excess levels of test passage (calculated in the same fashion) and identifies the local average treatment effect of receiving a higher test grade *due to test score manipulation* on labor market earnings.[3]

This methodology could be used more generally to study the consequences of manipulation in many other contexts, such firm earnings manipulation to meet analysts forecasts (Terry, 2016) or securitization of loans around key credit score thresholds (Keys et al., 2010). All these cases would have lent themselves to a Regression Discontinuity (RD) analysis *in the absence of manipulation*; however, these cases also constitute "textbook examples" of manipulation of the running variable (McCrary, 2008). Our methodology does not attempt to "solve" this issue and recover the effect of narrowly crossing the threshold in the absence of manipulation.[4] Instead, we identify *another* parameter – the causal impact of being manipulated across this threshold.[5] Because teachers select whom to manipulate, the LATE that our Wald estimator identifies captures the causal impact of manipulation on the subset of students that are chosen for manipulation. In the same vein, in cases of manipulation of firm earnings or loan securitization, we may be primarily interested in the effect of this manipulation for the firms or lenders that choose to manipulate.[6]

[3] In the language of the potential outcomes framework (Imbens and Angrist, 1994), among students who reach the test score manipulation window, those who de facto are graded up are akin to "compliers"; whereas those who are left un-manipulated can be thought of as "never-takers."

[4] For estimates of the former effect, Ebenstein et al. (forthcoming) provide important evidence on the impact of narrowly passing a high-stakes test in Israel on earnings. Their results highlight that the test allocates high-stakes rewards in an essentially random fashion to those who narrowly fare better (due to lower pollution exposure during the test day).

[5] Dee et al. (2016) seek to identify the same parameter in the context of New York City Regent's exams, using a different methodology. The only other paper that, to our knowledge, analyzes the causal effect of manipulation is Chen et al. (2017); they employ the empirical methodology that we develop to analyze the consequences of manipulation of R&D investment on firm-level productivity.

[6] Gerard et al. (2016) develop a methodology to identify yet another distinct, third, parameter – the impact of narrowly crossing the threshold for the un-manipulated in a context with manipulation; in our context, this would correspond to analyzing the impact of crossing the high-stakes threshold for those who do not receive manipulation and whose un-manipulated scores are right below the threshold.

Our results suggest far-reaching consequences of receiving test score manipulation, at all subsequent stages of the student's life that we observe. As expected, students who are manipulated (across any threshold) on the test, taken in February, receive a higher final grade in math in June; indeed, the nationwide test's importance for the final grade is precisely what makes it a high stakes test. In addition, manipulation of the math test causes students to perform better in *in other subjects* in the immediate future, between February (when the test is taken) and June (when final grades are awarded). Interestingly, these spill-over effects on other subjects arise only at the higher end of the ability distribution, around the PwD cutoff. We conjecture that these effects are driven either by self-signaling, where a higher test grade boosts the student's self confidence and effort, or potentially by signaling to teachers to give higher grades as well.

We then examine outcomes at the next educational stage, high school. Being graded up above the lower (Pass) threshold raises the likelihood of high school graduation three years later by 20 percentage points. This reflects a "gatekeeper effect" as a passing final math grade is a requirement for high school eligibility.[7] Manipulation around the PwD threshold does not alter students' eligibility for high school and, predictably, has a substantially smaller effect on high school graduation; intuitively, most of the students at this higher point in the ability distribution would proceed to high school directly after grade nine, regardless of whether they get a Pass or PwD in math. We also find that manipulated students perform better in high school: Those inflated above the lower (higher) threshold have 11% (7%) higher high school GPA. Conditional on attending high school, however, manipulation does not push these students into high schools with higher peer GPAs.

Manipulated students continue to benefit eight years after test score manipulation. Students who are bumped across the lower (higher) threshold are 12 (8) percentage points more likely to enroll in college, and complete 0.33-0.5 more years of education by age 23. Moreover, manipulated students are less likely to have a child during their teenage years. These effects translate in to substantial income gains at age 23 (the end of our sample period): Around both thresholds, manipulated students earn \$ 3400 - \$ 4400 more in annual income.

In sum, despite the fact that test score manipulation does not, per se, raise human capital, it has far-reaching consequences for the beneficiaries.

The immediate impacts on other subjects than math suggest that, in the context that we analyze, the "self-signaling" effect of receiving manipulation is large: First, grade nine is the second time that Swedish students have *ever* received grades in school, which makes their priors on their own ability relatively weak. Second, students do not observe their numeric

[7] The only option available to a student who does not obtain a passing grade in math is a one-year remedial program that serves to get the student ready for a three-year high school program with a one year delay.

score on the test – only the letter grade – which means that bumping a student up creates a large signal change.[8] The large long-term effects are consistent with a mechanism that suggests important dynamic complementarities: Getting a higher grade on the test serves as a signal *within the educational system*, motivating students and potentially teachers; this, in turn, raises human capital in the immediate future; and the combination of higher effort and higher human capital ultimately "snowballs," generating large labor market gains.

These insights are related to the sheepskin literature, which analyzes the signaling value of education in the labor market. It has generally found a small or zero signaling value (Cameron and Heckman, 1993; Kane and Rouse, 1995; Jaeger and Page, 1996; Kane et al., 1999; Tyler et al., 2000; Clark and Martorell, 2014), though a recent study demonstrates positive returns among low ability students in France (Canaan and Mouganie, Forthcoming). Our results suggest that, particularly for high ability students, the signaling value of grades may be very important *inside the educational system itself*, by raising students' motivation or other teachers' perceptions.[9]

The importance of signaling inside the educational system ties to the literature on the impact of receiving a lower or higher grade (Jacob and Lefgren, 2006; Manacorda, 2012; Papay et al., 2015; Smith and Avery, 2017). This literature focuses on educational attainment within a few years of receiving a certain grade, and do not follow the students into the labor market. Our paper essentially marries this literature with the literature on the sheepskin effect, by drawing on data that allows us get inside the "black box" of how a potential signaling effect within the educational system affects each step of the educational trajectory, and ultimately outcomes in the labor market. Moreover, while the literatures on sheepskin effects and the impact of receiving a certain grade generally focuses on adversely selected student populations, we study a test taken by the universe of all students in Sweden and demonstrate that self-signaling is particularly important at the higher end of the ability distribution.

A few papers assess the potential causes of test score manipulation.[10] Previous work

[8] "Self-signalling" is closely related to the concept of perceived self-efficacy (Wuepper and Lybbert, forthcoming). Our results are consistent with Bandiera et al. (2015) who show motivational effects of feed-back about academic performance when students lack perfect information about their own ability. An alternative interpretation is that the extra effort in other subjects is driven by a change in actual incentives, rather than in motivation. However, the improvement in other subjects is concentrated among higher ability students, for whom the high school incentive effect is unimportant (moreover, manipulation around the higher ability cutoff does not push these students into better high schools), suggesting that self-efficacy is at play rather than a change in incentives. Hvidman and Sievertsen (2017) analyze a context where students respond to a change in actual incentives stemming from a reform that recalculated students' GPA.

[9] Around the Pass threshold, the "gatekeeper" effect of receiving a passing grade on high school eligibility also relates to the returns to additional educational attainment around the dropout age (Angrist and Krueger, 1991; Oreopoulos, 2007; Brunello et al., 2009).

[10] While we focus on *ex post* test score manipulation, a related literature analyzes *ex ante* manipulation,

has focused on school-level incentives to manipulate scores, such as penalties for poor school performance (Jacob and Levitt (2003)), school competition (Tyrefors Hinnerich and Vlachos (2013))[11], and teacher-level monetary incentives (Lavy (2009)). Dee et al. (2011) document manipulation of test scores in New York City, and show that it is driven by teachers' desire to help their students avoid a failure to meet exam standards.

We contribute to the small, growing literature that analyzes the impact of teacher discretion on academic achievement. Lavy and Sand (2015) demonstrate in Israel that teachers' grading display a gender bias favoring boys, and that this bias boosts (depresses) boys' (girls') achievements and raises (lowers) boys' (girls') likelihood of enrolling in advanced courses in math.[12] Apperson et al. (2016) study the impacts of explicit cheating of teachers who erase and correct wrong answers of students on high stakes tests in a US urban school district, finding that students whose answers were changed performed worse on future achievement tests. Concurrent work by Dee et al. (2016) builds on Dee et al. (2011) and analyzes the effect of manipulation of the New York City's Regent's Exam on students' high school outcomes, finding that manipulation reduces the likelihood of high school graduation and raises the likelihood of advanced placement course taking. Our paper differs from this literature by studying key long term outcomes including labor market earnings, child bearing, and educational attainment in the longer run, over 7 years after test taking.

Methodologically, our paper contributes to the small, but growing literature on estimators that use cut-offs and notches to estimate causal effects other than those identified by standard regression discontinuity methods. Dong and Lewbel (2015) and Angrist and Rokkanen (2015) develop methods for identifying RD effects away from the RD cutoff. Gerard et al. (2016) develop treatment effect bounds for RD estimators when the RD itself is manipulated. As far as we are aware, ours is the first paper to develop a method for estimating the treatment effect of *being manipulated*, by extending the ideas underpinning the bunching literature.

The remainder of the paper proceeds as follows. Section 2 describes the institutional setting and data, and Section 3 provides a primer of the theory and estimation (i.e., of

including efforts to "teach to the test" (Jacob, 2005), alter the test-taking population (Figlio and Getzler, 2002; Cullen and Reback, 2006), and target instruction to students who are believed to be at risk for falling close to target thresholds (Neal and Whitmore Schanzenbach, 2010).

[11] While we do not observe re-graded test scores, the counterfactual test score density that we estimate approximates the re-graded distribution. On the relationship between grading leniency and competition for students, also see Vlachos (2010) and Böhlmark and Lindahl (2012) for evidence from Sweden, and Butcher et al. (2014); Bar et al. (2009) for evidence from the U.S.

[12] They quantify gender biased grading leniency by comparing teachers' average grading of boys and girls in a non-blind classroom exam to the respective means in a blind national exam marked anonymously. This, in spirit, is very similar to our methodology: we compare the distributions of test scores under manipulation to what these distributions would have looked like in the absence of manipulation. Contrary to Lavy and Sand (2015), however, we do not readily observe these counterfactuals from blind grading of the same tests, but we develop a methodology to estimate them.

the subsequent three sections). Section 4 provides a model of teachers' grading behavior. In Section 5, we estimate counterfactual test score distributions. Section 6 then uses data on long-term outcomes, coupled with our estimates from Section 5, to quantify the causal impacts of test score manipulation on subsequent schooling and adult labor market outcomes. Our results are presented in Section 7.

2 Institutional Background and Data

2.1 Schooling in Sweden

Schooling is mandatory until age 16, which corresponds to the first nine years of school (ages 7-16). During the sample period, grades are awarded for the first time in the Spring semester of eighth grade. One year later, at the end of ninth grade, *final grades* are awarded and the GPA is calculated.[13] The grading scale is Fail, Pass, Pass with Distinction (henceforth PwD), and Excellent. All students who wish to continue to the next scholastic level, grades 10-12 (ages 16-18; roughly equivalent to high school in the U.S.), must actively apply to grade 10-12 schools (henceforth high schools).[14] To be eligible for high school, the student must have passing final grades in math, English, and Swedish. Conditional on eligibility, the grade nine GPA is the sole merit-based criterion used for acceptance to high school.[15] The GPA cutoff for admittance to a given program, in a given year, is determined by the lowest GPA among admitted individuals. At the end of high school, prospective university students apply to university programs, with admittance determined in a similar fashion by thresholds in (high school) GPA.

Nationwide tests. All students in grade nine take nationwide tests in mathematics, usually in the beginning of the Spring semester (February), i.e., four months before the teacher sets the final grade in mathematics. The test is graded locally, either by the teacher or jointly by the teachers at a given school, according to a detailed grading manual provided by The Swedish National Agency for Education.[16] While some points are awarded based on non-

[13] The GPA reflects the average grade in 16 subjects taught in grades 7-9 and ranges from zero to 320.

[14] There is a range of high school programs, ranging from vocational to university preparatory. See Golsteyn and Stenberg (2015) for a comparison of earnings over the life cycle for students choosing different programs.

[15] In particular, no other merit-based criterion (e.g., essays, school-specific entry tests, etc.) that are commonly administered in the U.S. are used in admittance decisions; the only factors that may be taken into account other than GPA are the distance to school and sibling preferences.

[16] The manual specifies the exact cutoffs in the raw test score, r_i, required for the test grades Fail, Pass, and PwD. In addition, it specifies a lower bound on r_i that constitutes a necessary but not sufficient condition for awarding the top grade, Excellent. The sufficient conditions for obtaining the highest grade are highly subjective criteria; moreover, we cannot observe them in the data. For this reason, our analysis considers the two lower test score thresholds only. Appendix C provides the exact step function from the grading sheet

manipulable criteria ("which number is larger?"), a subset of points may be awarded for partially completed work, beautiful expression, etc. This gives the teacher some leeway in raising a student's test score. When writing the test, the students do not know the cutoffs; further, these cutoffs vary over time. Any bunching that we observe is thus attributable to teacher grading, not student sorting. Finally, students are only informed of their letter grade on the test, but not their raw numeric score. A student who fails is thus not aware of how close (s)he were to the passing cutoff; this removes any incentive for teachers to grade *down* in order to avoid complaints. The tests cannot be appealed.[17]

Nationwide tests are also administered in English and Swedish. The test grades on these two tests are not based on any numeric test scores, however; these test grades are awarded based on assessments of the quality of the students' writing and reading.

Final grades. The final grade in mathematics (that counts towards the GPA) partly reflects the test grade, but the teacher also takes into account all other performance, e.g. on homework and in-class tests, when setting the final grade.

Schools' incentives to manipulate. Why would teachers manipulate their students' test scores? On the one hand, as teachers to some extent know their students personally, they may experience emotional discomfort when awarding bad grades. On the other, awarding a higher grade than a student deserves may constitute a devaluation of the teacher's own professionalism. While these mechanisms, and a myriad of others, likely are at play in all schools and institutional contexts, a combination of two particular features of Sweden's schooling system may make the country particularly susceptible to manipulation: First, municipal and voucher schools compete for students – or, more to the point, for the per student voucher dollars that the municipality pays the school for each admitted student.[18] Second, the key way for a grade 7-9 school to attract students is to produce cohorts with a high average GPA *in the cohort that exits from* ninth grade.[19] This gives school principals strong incentives to encourage their teachers to go easy on grading.[20]

from 2004, along with more detailed information about the tests and r_i.

[17] Teachers are also not subject to any "cap" in how many students can obtain a Pass or PwD.

[18] Nonvoucher tuition payments are forbidden in Sweden.

[19] Indeed, schools are often ranked, in newspapers and online, based on the GPA of the exiting cohort of ninth graders (which is public information in Sweden). This in practice ties a school's reputation for quality to the average GPA of its exiting cohort, even though this measure does not capture school value added. Put differently, if schools believe that parents, especially those with high ability children, rely on these rankings when choosing a suitable school for their child, schools face an incentive to produce cohorts with a high average GPA in grade nine. As a result, schools can compete either in the intended fashion, by providing a better education, which justifiably may raise the grades of the exiting cohorts; or by engaging in manipulation, which artificially raises the school's reputation for quality.

[20] In municipal schools, teachers are not compensated based on the performance of their students, either on nationwide tests or on other performance measures (while voucher schools may engage in such practices). Nonetheless, anecdotally, public school teachers have reported feeling pressure from the principals to "pro-

2.2 Data

We start from the universe of students who finish ninth grade between 2004 to 2010. For these children and their families, we obtain information from various data sources:

Grade nine academic performance and schooling information. We observe precise information on each student's performance on the nationwide tests in mathematics, English, and Swedish. On the math test, we observe both the score (after possible manipulation by the teacher) and the test grade. On the English and Swedish tests, we observe the test grade. Because the English and Swedish nationwide tests are taken before the math test, we can use them as pre-determined measures of student ability. We also observe the final grade in math, grade nine GPA, and the type of school (municipal or voucher).

Demographic and socio-economic characteristics. For each child, we observe the date of birth, the municipality of residence when attending ninth grade, and whether the student has a foreign background.[21] We also observe each parent's year of birth, educational attainment, immigration status and annual taxable earnings.

Medium- and longer-term outcomes. To trace economic outcomes after ninth grade and throughout adolescence into adulthood, we add information from high school records, university records, and tax records. Our key outcome variables in the medium term capture information about high school completion, performance in high school (high school GPA), and the quality of the high school (measured by high school peer GPA). At the university level, we observe whether an individual initiates university studies, which is defined as attending university for two years or less. Moreover, we observe total educational attainment by 2012, which corresponds to the age of 24 for the oldest cohort in our sample. We also observe the exact (employer-reported) taxable income for all years in which the student is aged 16 and above. In 2012, the last year for which we observe income, the individuals in our sample are up to 24 years old. Earnings at age 23-24 likely captures income from stable employment in the sub-population of individuals who do not attend university. Among university enrollees, however, it is too early to capture income from stable employment. Finally, we observe an indicator for teen birth by 2009, which corresponds to the age of 21 for the oldest cohort in our sample.

Sample and Summary Statistics Our sample consists of all students who attend ninth grade between 2004 to 2010 and both took the national test (obtained a non-missing, positive test score) and obtained a final grade in math. The first column of Table 1 presents

duce" high test grades, in order to satisfy parents and boost the school's image in face of the competition for students.

[21] We define foreign background as having a father who is born outside of Sweden.

summary statistics for the full sample, and the second and third columns for two distinct subsamples: students that obtain a test score that is subject to potential manipulation around the threshold for Pass and PwD, respectively. The definition of these three regions varies by year, county, and voucher status of the school, and are derived from our estimates of the size of the manipulation regions, which we discuss in detail in Section 5. In our full sample, 93 percent of the students receive a final grade in math of Pass or better (i.e., seven percent receive the final grade Fail); and 18 percent obtain PwD or better. We let the final grade in math take the value of 0 if the student fails; 1 if the grade is Pass; 2 for PwD; and 3 for Excellent. In the overall sample, the average grade is 1.13.

In the full sample, the average test score is 28.3, and the averages (mechanically) increase as we move from the lower to the higher threshold. 5.7 percent of all students attend a voucher school. The mean GPA in the entire sample is 191.[22] Our key longer-term outcomes of interest are whether the student graduates from high school, college attendance, educational attainment, and income earned at age 23. Income is estimated in 2011 and 2012 for the students who attended grade nine in 2004 and 2005. In the full sample, 76 percent of the students graduate from high school. Finally, 22 percent of the full sample of students have a foreign background.

3 A primer

Here, we provide a roadmap and some intuition for the analysis in the next three sections.

Model of test score manipulation All bunching methodologies rely on the idea that manipulation is confined to some parts of the density, whereas others are un-manipulated; this is why we can use the un-manipulated parts to learn about what would have happened inside the manipulation regions in the absence of manipulation. Section 4 models teachers' behavior and provides a theoretical basis for this in our context. In a nutshell, the model shows that there exists a lowest test score where manipulation may take place (around each grade threshold); below it, the distribution is un-manipulated with probability one. This is true at any level of aggregation of the data: at the teacher level, the school level, the regional level, and so on.[23] The model thus echoes a familiar result: the optimality of threshold strategies in notch settings (Kleven and Waseem, 2012). We show that it holds in

[22] See footnote 13 for more information about the GPA.

[23] This follows trivially from the fact that each teacher's strategy is characterized by a minimum test score at which she is willing to bump any of her students up. To see this, consider a school with two teachers, A and B, with minimums 18 and 20, respectively. The minimum test score at which any student *at the school* may be graded up simply is the minimum of these, i.e., 18. The minimums at any higher level of aggregation are obtained in an analogous fashion.

a very general formulation of the teachers problem in our setting, where teachers are allowed to treat students differently even if they have the same test score. Moreover, we show that the threshold does not depend on the students' *realized* test scores.

Estimation of counterfactual densities In Section 5, we turn to estimating counterfactual test score distributions. Because a minimum test score at which manipulation occurs exists at any level of aggregation, we could in theory do this at the teacher, school, regional, or even national level. We perform the analysis at the county*voucher*year level (henceforth referred to as region*year), a choice we discuss further below.

Existing bunching estimators have relied on visual inspection for determining where manipulation begins and/or ends (Saez, 2010; Chetty et al., 2011; Kleven and Waseem, 2012); we refine these methods and develop an estimator that allows us to *estimate* where manipulation begins and ends. The procedure delivers one counterfactual test score distribution for each region and year. Each counterfactual tells us how far down in the test score distribution manipulation occurs, around the Pass and PwD thresholds, respectively. We find that in some places, there is hardly any manipulation; in others, some students as far as 7 test score points below the (Pass) threshold are bumped up. Figure 1 illustrates the estimated regional counterfactuals, aggregated up to the national level, in year $t = 2010$. The blue connected line plots the observed distribution of test scores, and the red connected line shows the estimated counterfactual density. At the national level, the lowest test score at which any student is manipulated is exactly 7 points below the pass threshold – we knew this already, as this is the lowest test score where anyone gets manipulated in the region with the most manipulation! This is rare, however; the modal width of the manipulation region is 3 test score points around Pass, and 1 test score point around PwD.

In addition to telling us where manipulation occurs, the key piece of information that we get from the (region-year) counterfactual densities is, for each test score point inside a manipulation region, a count of how many students would have gotten that score in a counterfactual scenario without manipulation. Indeed, this is the definition of a counterfactual density; nonetheless, it bears emphasizing here, as this counterfactual "head count" is one of two key inputs into our estimation of causal effects.

The impact of manipulation The central methodological contribution of our paper is to develop an estimator that quantifies the causal effect of manipulation. This analysis is done at the same level of aggregation as the estimation of the counterfactual density. To see how it works, consider an example (region*year) where the Pass threshold is 22 and manipulation occurs in three test score points below this threshold, i.e., in points 19, 20, and 21; further, teachers may bump students up all the way to 24.[24] Thus, the manipulation region around the Pass threshold is 19-24.

A key feature of our data is that we observe only manipulated test scores: If we observe Anna at 23, we do not know whether she truly scored 23 and was left un-manipulated, or if her true score fell below the threshold and she was bumped up to 23. More generally, to the right of the threshold inside the manipulation region, we cannot distinguish manipulated from un-manipulated students.

How, then, can we quantify the impact of manipulation? The key step is to recognize that, while some students who truly scored 19-24 have been moved, no student has been moved into our out of the manipulation region – indeed, this is the very definition of the manipulation region! Thus, if we observe 100 students with scores 19-24 in the (manipulated) data, we know that there would have been 100 students scoring 19-24 in the absence of manipulation (distributed differently within the manipulation region).

We can thus consider *all* students inside the manipulation region. Our Wald estimator is constructed in three steps:

1. Construct an estimate of what would have been the average long run outcome, e.g. earnings, *of all 100 students* scoring 19-24 in a counterfactual world without manipulation. That is, we construct $\mathbf{E}(Earnings|No\ manipulation)$ for all 100 students.

2. Calculate the actual average earnings of these 100 students (from the raw data). If 15 of the 100 students in fact received manipulation, then this average reflects the earnings of 15 manipulated students and 85 un-manipulated students.

3. The difference between (1) and (2) is driven by the (earnings of the) 15 students who were manipulated. It constitutes the reduced form effect of test score manipulation on earnings. The LATE scales this by the the probability of being bumped up (here 15/100, i.e., 0.15).

The key step is (1). To see how we do this, Figure 2a depicts average age 23 earnings, by test score, outside of the manipulation region in our example. Because all these students are un-manipulated, this data reflects the true, underlying relationship between test scores and (average) earnings. We fit a flexible polynomial to this data and predict the relationship inside the manipulation region, allowing for a true effect on earnings of passing the threshold. Figure 2b illustrates the result: six light-grey points inside the manipulation region.

Intuitively, in a world without manipulation, a certain number of students would have scored 20 on the test; then, at age 23, they would have obtained some earnings. The point

[24] As we discuss further below, due to lumpiness in the points, teachers may have to overshoot.

highlighted in yellow in Figure 2b shows our prediction of the average age 23 earnings of all students scoring 20 in a world without manipulation. For each test score point inside the manipulation region, we obtain an analogous prediction.

With the data at hand, we do not know precisely which of the 100 students inside the manipulation region would have scored 20 in a world without manipulation. But we have an estimate of *how many* students would have scored 20 – our counterfactual test score distribution gives us precisely this "head count". So, we have everything that we need to construct $\mathbf{E}(Earnings|No\ manipulation)$ for all 100 students inside the manipulation region:

$$\mathbf{E}(Earnings|No\ manipulation) = {}^{1}\!/_{100} \sum_{s=19}^{24} \mathbf{E}(Earnings|r_i = s)^{*} Headcount(s). \tag{1}$$

That is, we simply multiply each grey predicted point inside the manipulation region, depicted in Figure 2b, with the "head count" in that test score point given by the counterfactual density, and sum it up.

Identifying assumptions. This relies on an assumption that the un-manipulated distribution of (average) earnings between 19 and 24 can be parametrically recovered from a polynomial fit to the un-manipulated regions of the distribution, an assumption akin to what underlies all bunching methodologies. Moreover, we rely on data points outside of the manipulation region to predict the relationship between test scores and earnings inside this region, in the absence of manipulation. In the modal region*year, manipulation occurs in three test score points below the Pass region and in one test score point below the PwD region.

Before proceeding, a few remarks are in order.

Remark 1: Selection on unobservables. It bears emphasizing that it is the distribution of (average) earnings inside the manipulation regions *in a world without manipulation* that we assume can be recovered by fitting a polynomial to the un-manipulated regions of the test score distribution. Suppose, for example, that teachers choose whom to grade up partly based on some characteristic that is unobserved to us, say, charisma. Suppose further that charisma has an independent effect on earnings that does not run through the receipt of manipulation. While this type of story would violate an exclusion restriction in a typical IV model, it is generally not a problem for our estimator. Under such selection-on-charisma, if we could plot the average charisma at each test score point in the presence of manipulation, we would have lower levels of charisma to the left of the cutoff, and higher levels to the right of the cutoff, potentially in a non-smooth fashion. All we assume, however, is that in an un-manipulated world – i.e., if we were to put the charismatic people who were bumped up back down to their proper places to the left of the threshold – then this distribution of charisma (and, hence, of expected earnings) could be recovered from fitting a polynomial outside of the manipulation region.

Remark 2: The choice of polynomial. While we allow for an effect of crossing the threshold in the absence of manipulation, we are not focused on this estimate per se; it may be sensitive to the polynomial with which we are making our prediction, and this is reflected in the standard errors. Instead, we are interested in (1) – or, more precisely, in the reduced form that we obtain by taking the difference between (1) and its sample analogue. The sum (1) is *not* sensitive to the exact choice of polynomial. To see the intuition for this, again consider 2b. Suppose that we would fit a polynomial of another order that would result in a slightly smaller estimated 'gap' around the Pass region. This would be associated with slightly higher predicted average earnings for the test score points 19, 20, and 21, and sightly lower predicted average earnings for the test score points 22, 23, and 24. These differences would move (1) in opposite directions, leaving the sum largely unaffected.

4 A model of test score manipulation

Set-up For simplicity, we model test score manipulation around a single threshold. Student i is taught by teacher j, who observes his ability a_i.[25] He takes the nationwide test and, in the absence of any manipulation, receives a numeric ("raw") test score $r_i = r\left(a_i, \varepsilon_i\right)$, where $\varepsilon_i \tilde{} F\left(\varepsilon_i\right)$ captures the fact that student i may have a "good " or "bad" test day. Because the teacher grades the test, she observes the raw test score r_i.

The teacher can award some amount of additional test points to student i, Δ_i. The test grade is an indicator function (for Pass): $t_i = t\left(a_i, \varepsilon_i, \Delta_i\right) = 1 \; if \; r\left(a_i, \varepsilon_i\right) + \Delta_i \geq \bar{p}$, where $\bar{p}$ is the passing threshold; and zero otherwise. The teacher chooses the amount of manipulation of student i's test score, Δ_i, to maximize the per-student utility function $u_{ij}(\Delta_i) = \beta_{ij} t\left(a_i, \varepsilon_i, \Delta_i\right) - c_{ij}\left(\Delta_i\right)$.

Here, β_{ij} measures teacher j's desire to raise student i's grade from a Fail to a Pass. Its dependence on j permits teachers to be heterogenous in their desire to inflate grades. Such heterogeneity may stem from teacher-specific factors, such as a teacher's aversion against incorrectly assigning a test score below the threshold, or from factors stemming from the school at which teacher j works, e.g., the competitive pressure that the school faces from other schools to attract students, pressure from the school principal to "produce" higher grades, etc. Moreover, the dependence of β_{ij} on i permits a given teacher to place a heterogenous value on raising different students' grades from Fail to Pass. Importantly, this permits the

[25] Strictly speaking, a_i need not reflect student i's true, innate ability; it is sufficient that it reflects the teacher's perception of student i's innate ability.

teacher to use her discretion both in a "corrective" and "discriminatory" fashion.[26] Finally, although we have formulated a per-student utility function above, note that the dependence of β_{ij} on i permits the teacher's desire to raise student i's grade from a Fail to a Pass to depend on the overall ability distribution in teacher j's class of students. Such preferences would entail if, for example, the teacher wants a certain percentage of the students in the class to obtain a passing grade.

In order to inflate a student's test grade by Δ_i, the teacher must pay a cost, $c_{ij}(\Delta_i)$, which satisfies $c_{ij}(0) = 0$ and is strictly increasing and convex, $c'_{ij}(\Delta_i) > 0, c''_{ij}(\Delta_i) > 0$.[27] Convexity captures the fact that it is increasingly hard for a teacher to add an additional test point.[28] For now, we assume that when the teacher chooses Δ_i, she is free to pick any (positive) value that she wishes.[29]

Teacher grading behavior We immediately see that $\Delta_i^* \in \{0, \bar{p} - r_i\}$, that is, the teacher either leaves student i's test score un-manipulated, or raises the student's final numeric grade to exactly $\bar{p}$.[30] Thus, the teacher's decision of whether to bump up a given student i who would fail in the absence of manipulation ($r_i < \bar{p}$) hinges on whether β_{ij}, the teacher's utility from raising the final grade from Fail to Pass, (weakly) exceeds the cost of the manipulation that is required to push the student just up to the passing threshold, i.e., on whether $\beta_{ij} \geq c_{ij}(\bar{p} - r_i)$. As β_{ij} is a constant and $c_{ij}(\bar{p} - r_i)$ is increasing in $\bar{p} - r_i$ (decreasing in r_i), we obtain:

[26] For example, a teacher may have corrective preferences if she places a higher value on inflating a student who had a bad test day. But this formulation also permits the teacher to have discriminatory preferences, e.g., placing a higher value on inflating students of a certain gender or from a certain socioeconomic group (whose parents, for example, may impose stronger pressure on the teacher). In Section 6, we empirically assess whether teachers appear to have corrective or discriminatory preferences; here, we keep a general formulation that permits each of these interpretations (as well as an interpretation where the teacher has a combination of corrective and discriminatory preferences).

[27] The fact that manipulation is costly immediately yields that the teacher never would manipulate to move a student *down*; this captures the institutional features discussed in Section 2.2 above: Because students are only informed of their letter grade on the test, but not of their raw numeric score, a student who fails is not aware of how close (s)he were to the passing cutoff; this removes any incentive for teachers to grade down in order to avoid complaints. Moreover, teachers are also not subject to any "cap" in how many students can obtain a Pass or PwD.

[28] For example, as discussed in Section 2 above, there are some points awarded on the math test that require subjective grading, while others are clearly right or wrong answers. Inflating a test score by a few points would only require somewhat generous grading on the subjective parts of the test, while a large amount of manipulation would require awarding points for more clearly incorrect answers. These costs may also be convex due to the possibility that a school might get audited and have to justify their grading, which is harder to do with larger amounts of manipulation.

[29] In reality, this is not always possible due to lumpiness in the points.

[30] Intuitively, the teacher never raises a student's score less than up to $\bar{p}$ because any amount of manipulation is costly; hence, a necessary condition for manipulation is that it alters the student's test grade from Fail to Pass.

Proposition 1. *Teacher j inflates student i if and only if he would fail in the absence of manipulation and $r_i \geq r_{ij,\min}$, where $r_{ij,\min}$ is implicitly defined by $\beta_{ij} = c_{ij}\left(\bar{p} - r_{ij,\min}\right)$.*[31]

Proposition 1 highlights two key things: **(i) Differential treatment**. Because $r_{ij,\min}$ varies at the student level, if the teacher has two students (say, Anna and Ben) and $\bar{p} = 20$, then the teacher's rule may be to bump Anna up if she receives 16 or more on the test, but to only bump Ben up if he receives 18 or more.

(ii) Decision rules independent of students' *realized* test scores. While the ultimate decision of whether to bump a student up or not depends on the student's test score – Anna is bumped up if she scores 16, but not if she scores 15, for example – the teacher's *decision rule* when it comes to Anna can be thought of as pre-determined.

Proposition 2. *For each teacher, there exists a lowest test score at which test score manipulation (of* any *student) occurs, $r_{j,\min}$. Consequently, students whose un-manipulated test score r_i falls below $r_{j,\min}$ have a zero probability of being inflated. Students whose un-manipulated test score r_i falls above $r_{j,\min}$ have a weakly positive probability of being inflated (to $\bar{p}$).*[32] *The threshold $r_{j,\min}$ is pre-determined and does not depend on students' realized test scores.*

Proposition 2 follows immediately from Proposition 1: For each teacher j, the minimum test score at which *any* of her students get manipulated is simply given by the smallest $r_{ij,\min}$ of all her students, $r_{j,min} = \min_i(r_{ij,min})$. Moreover, importantly, because each of the teacher's student-specific thresholds $r_{ij,\min}$ are pre-determined, the teacher-specific threshold $r_{j,min}$ is pre-determined as well. The same logic yields:

Corollary 1. *For each school s, there exists a lowest test score at which test score manipulation occurs, $r_{s,\min}$. Similarly, for each geographical region g, there exists a lowest test score at which test score manipulation occurs, $r_{g,\min}$.*

5 Estimation of counterfactual densities

Because a minimum test score at which manipulation occurs exists at any level of aggregation, we could in theory estimate counterfactual densities at the teacher, school, regional, or even

[31] Because student i would fail in the absence of manipulation so long as $r_i < \bar{p}$, the teacher inflates student i if and only if his raw test score falls in the interval $r_i \in [r_{ij,\min}, \bar{p})$.

[32] Specifically, among the students whose test scores fall above $r_{j,\min}$: (i) the probability of receiving inflation is one for students whose raw test score satisfies $r_i \in [r_{ij,\min}, \bar{p})$ – these can be thought of as "compliers;" and (ii) the probability of receiving inflation is zero for students whose raw test score satisfies $r_i \in [r_{j,\min}, r_{ij,\min})$ – these can be thought of as "never-takers." This is discussed further in Section 6 below.

national level. In practice, however, there are too few students per school to give us sufficient statistical power to do this at the teacher or school level. Intuitively, we need "enough data" close to the thresholds to quantify bunching. We estimate counterfactual densities for each county, separately for voucher and non voucher schools, in each year from 2004 to 2010.[33] By aggregating many schools together, we identify the minimum test score where manipulation occurs across all these schools. Below it, the distribution is un-manipulated with probability one in all (voucher or public) schools in the region and year.

Existing bunching estimators have relied on visual inspection for determining where the manipulation region begins and/or ends (Saez, 2010; Chetty et al., 2011; Kleven and Waseem, 2012).[34] We cannot manually choose any parameter that defines the width of the manipulation window, for two reasons: First, we analyze a large number of distributions, which makes visual inspection tedious; moreover, we want to be able to determine the manipulation region even when diffuseness of the bunching mass renders visual inspection imprecise. Second, we want to allow for the possibility that there is no manipulation in some locations. We therefore refine existing methods and develop a "fully automatic" estimator, which does not require picking any parameters using visual inspection.

Because we want the estimator to identify where manipulation begins and ends, as well as allow for the possibility of zero manipulation, we first must make a shape restriction on the un-manipulated density. Without this, one could not reject that any observed bunching simply represents an unusual looking test score distribution *without* manipulation.[35] We assume that the un-manipulated density is log concave. Log concavity is an appealing assumption because it is a sufficient condition for a single peaked and continuous test score distribution, and it is easy to mathematically implement. It also allows for considerable generality: many commonly used probability distributions are log concave (normal, gumbel, gamma, beta, logistic). Further, we assume that the densities of the missing and excess mass (to the left and right of each threshold) are log-concave and monotonic. This imposes that, as one moves further away from the threshold (in either direction), there is less manipulated mass. The assumption thus governs how "fat" the tails of the manipulation can be, and effectively limits the possibility that manipulation reaches far into the overall distribution

[33] We aggregate voucher schools in counties where fewer than 200 students are in voucher school in 2004. We maintain the definition of this "aggregate voucher*county" throughout the time period 2004-2010. We cannot pool data across years because the grade cutoffs move around each year.

[34] Most closely related to our setup is Kleven and Waseem (2012) (henceforth KW), the first paper to develop a method to estimate where the manipulated region of a histogram around a notch begins. KW's method relies on visual inspection of where manipulation of the analyzed distribution ends; this is suitable (only) when excess bunching is very sharp, so that the end of bunching can be determined visually.

[35] An assumption of smoothness would be the weakest restriction needed to test for whether there is any manipulation present at all, but we will need more than this to identify where manipulation begins and ends.

without making the aggregate distribution "appear" manipulated. Intuitively, this is a way to mathematically mimic what one would do if manually picking the point where manipulation ends – one would pick the point where the distribution no longer appears manipulated.

Having specified the shapes of the un-manipulated distribution and of manipulation permits the second novel feature of our estimator, namely, that it iterates over all possible widths of the manipulation region (including zero) – as well as over a number of other parameters to be specified below. It then uses a mean squared error criterion function to compare different possible estimates of the width of the manipulation region and the shape of the un-manipulated distribution.[36] Finally, cross-validation will be used to pick the orders of the polynomial used to fit the log-concave distributions. Given the orders of the polynomials, our method identifies where manipulation begins and ends by selecting the point in which the data best fits switching from the polynomial fitting the manipulation to the region fit to the un-manipulated polynomial.

The un-manipulated distribution Let $h_{jt}(r)$ equal the frequency of students in region j in year t that would receive a test score of r in the absence of test score manipulation.[37] We define $h_{jt}(r) = B^{N_{jt}}(\theta_{jt}, r)$, where $B^{N_{jt}}$ is a function of an $N_{jt}th$ order polynomial with coefficients θ_{jt} where the coefficients of $B^{N_{jt}}$ are constrained to force $h_{jt}(r)$ to be log-concave. We implement this by using Bernstein polynomials, which are a family of orthogonal polynomials where linear inequality constraints on the polynomial coefficients can directly impose global concavity (Wang and Ghosh (2012)). We discuss these details more in Appendix D.

Missing and excess mass Let $m_{jt}^{low,k,p_{kjt}^{low}}(\theta_{jt}^{low,k}, r)$ and $m_{jt}^{high,k,p_{kjt}^{high}}((\theta_{jt}^{high,k}, r)$ equal the amount of missing and excess mass in the vicinity of grade threshold k at test score r in region j and year t. We parameterize each as exponentiated polynomials, with coefficients $(\theta_{jt}^{high,k}, \theta_{jt}^{low,k})$ and orders $(p_{kjt}^{high}, p_{kjt}^{low})$. We constrain the missing and excess mass around each threshold to be equal.[38] In addition, we impose that $m_{jt}^{high,k,p_{kjt}^{high}}(\theta_{jt}^{high,k}, r)$ and $m_{jt}^{low,k,p_{kjt}^{low}}(\theta_{jt}^{low,k}, r)$ are monotonic and non-negative at all test scores r. This guarantees that there can only be missing (excess) mass below (above) each threshold k and that the amount

[36] The KW method instead selects the *narrowest* manipulation region which *could* be consistent with the data, and does not systematically compare all possible widths of the manipulation regions and their overall model fit. Our method thus provides a broader search of possible estimates and uses a criterion function to compare them.

[37] As mentioned above, we will estimate grading leniency at the county*voucher*year level. We let j indicate a county*voucher, and will for simplicity refer to it as a "region."

[38] In the language of the bunching literature, this condition rules out an "extensive margin" response (Persson, 2014). In our setting, this is very intuitive: the presence of manipulation moves students around in the test score distribution, but it does not make any student disappear from the test score distribution altogether. Consequently, all students that are moved up from below the threshold are located above the threshold in the manipulated distribution.

of manipulated mass is always shrinking as one moves away from any threshold. Again, we use Bernstein polynomials as our polynomial basis to enable us to easily impose the desired shape restrictions.

Width of the manipulation region Define β_{kjt} as the difference between the test grade threshold k and the minimum test score to ever receive manipulation in region j in year t. β_{kjt} is a key parameter of interest, as it measures how many points below the test score threshold a student has any chance of receiving test score manipulation. β_{kjt} also yields an upper bound on how far above the grade cutoff a test score can be manipulated – this is needed because the test score points are lumpy, so that it is not always possible for the teacher to award exactly the number of points needed to put the student at the threshold:[39] if β_{kjt} is the number of points that the most lenient teacher is willing to add to a student's score, then this bounds any overshooting to $\beta_{kjt} - 1$ points above the cutoff. This gives us our final restrictions: the amount of missing mass below $k - \beta_{kjt}$ is equal to 0 and the amount of excess mass above $k + \beta_{kjt} - 1$ is zero.

Estimation We estimate the model using constrained nonlinear-least squares. To select the orders of the Bernstein polynomials, we use k-fold (k=5) cross-validation.[40] The parameters to be estimated are $(\beta_{1jt}, \beta_{2jt}, \theta_{jt}^{low,1}, \theta_{jt}^{high,1}, \theta_{jt}^{low,2}, \theta_{jt}^{high,2}, \theta_{jt})$. We estimate the model separately for voucher and municipal schools, within each county, in each year. Appendix D provides additional technical details.

To give some intuition behind how our estimator identifies β_1 and β_2, Appendix Figure A1a plots what our model would estimate for the manipulated and un-manipulated test score distributions if β_1 were set to 1 and β_2 were set to 0. These data are for municipal schools in Stockholm in 2005. Note how this fits the data very poorly around the PwD cutoff of 41 and does not match well the test score distribution for low scores below 20. In contrast, Appendix Figure A1b shows the estimated distributions if β_1 is set to 4 and β_2 is set to 1. This allows the estimator to match the observed distribution of test scores in the data much better; and in fact, for municipal schools in Stockholm in 2005, we obtain the estimates $\hat{\beta}_1 = 4$ and $\hat{\beta}_2 = 1$.

Appendix Figure A2 displays histograms of the estimates of $\hat{\beta}_{1jt}$ (upper panel) and $\hat{\beta}_{2jt}$ (lower panel) – that is, of the estimated widths of the manipulation region around the thresh-

[39] For example, either the teacher must assign 2 or 3 extra points (or 0).

[40] This is implemented by splitting the histogram for each county-voucher year into 5 subsamples of data. We minimize the mean-squared error over 4 of the subsamples, and predict out of sample using the estimated parameters on the 5th "hold out" sample and calculate the out-of-sample mean squared error. We do this for each of the 5 subsamples and sum together each of the 5 out-of-sample mean squared errors. We pick the orders of the polynomials that minimize this out-of-sample mean-squared error. We do this separately for each county-voucher-year.

olds for Pass and PwD, respectively. In some regions, students' test scores are manipulated as much as 7 test score points around the Pass region. This is rare, however; the modal width of the manipulation region is 3 test score points around Pass, and 1 test score point around PwD. Note that we do not use this cross-sectional variation when identifying the impact of test score manipulation on future outcomes – we simply show this distribution to give an idea of how much manipulation takes place in Swedish schools.

6 The impacts of manipulation

So far, bunching strategies have been used to analyze *the distribution that is being manipulated*: for example, distributions of reported incomes (Saez, 2010; Chetty et al., 2011; Kleven and Waseem, 2012), dates of marriage (Persson, 2014), or, as in the previous section of this paper, distributions of test scores. We develop a bunching methodology to examine the impact of manipulation on *other variables* than the one that is being directly manipulated. Intuitively, just like we can plot the test score distribution, we can plot the mean of a future outcome (say, earnings) by test score. In the two test score ranges where manipulation occurred, the observed earnings distribution *partly captures the impact of test score manipulation on earnings*. In contrast, in the test score ranges where no manipulation occurred, the observed relationship between mean earnings and test scores captures the underlying relationship between (un-manipulated) test scores and earnings. Thus, we can use data outside of the two manipulation regions to predict a counterfactual relationship between earnings and test scores inside these regions. In each manipulation region, the difference between the average observed and average counterfactual earnings captures the reduced form impact of (potentially being exposed to) test score manipulation on earnings. Finally, the LATE scales this reduced form effect by the probability of being graded up.

Appendix Figure A3 illustrates how students in the the manipulated regions of the test score distribution can be thought of in terms of the potential outcomes framework. To make the figure as clear as possible, we use the national distribution in 2010, from Figure 1, in this example – thus, the manipulation region is the widest that we ever estimate, starting at 14 around the Pass threshold at 21. (The modal width of the manipulation region is substantially smaller: 3 test score points around Pass, and 1 test score point around PwD.) Among the students whose raw test scores fall into the interval $14 - 20$, teachers choose to grade up a subset; these can be thought of as the compliers, who are "missing" below 21 in the observed test score distribution. The students whose observed test scores lie in the interval $14 - 20$ can be thought of as never-takers, as they are left un-manipulated even though their test score was close enough to the threshold for the teacher to consider them for manipulation. Finally, the students whose raw *and* observed test scores lie at or above 21 (but remain within the manipulation region around the Pass threshold) can be thought of as always-takers. In the data, we can identify the never-takers; however, we cannot distinguish the compliers from the always-takers, as both groups' observed test scores fall at or above 21 and we do not observe the raw test scores.

Identifying assumption First, the key identifying assumption is that, in the absence of manipulation, the distribution of outcomes inside the bunching region would have followed the polynomial estimated outside of the bunching region, an assumption akin to what underlies all bunching methodologies. This ascertains that data from outside of the manipulation regions is informative of the counterfactual distribution within these regions. Recall from Section 4 that the teacher-specific (and, hence school-specific and region-specific) thresholds do not depend on students *realized* test scores. Second, we assume that we can rely on data points outside of the manipulation region to predict the relationship between test scores and earnings inside this region, in the absence of manipulation. In the modal region*year, manipulation occurs in three test score points below the Pass region and in one test score point below the PwD region. Third, to interpret our LATE, we assume that there are no defiers.[41]

It is worth re-emphasizing that we *do not* need to assume that teachers allocate test score manipulation to a random subset of the students. In contrast, from our theoretical framework in Section 4, we expect that the students who are bumped up are a select subset; moreover, selection may be based on characteristics that are observable *as well as unobservable* to us. The LATE that our Wald estimator identifies captures the causal impact of manipulation *on the subset of students that are chosen* for manipulation. In Section 7.1, we analyze what observables teachers' use in selecting which students to bump up. Intuitively, this is akin to an analysis of complier characteristics.

6.1 Identifying the causal impact of test score manipulation

The "reduced form" The first step is to estimate the relationship between students' earnings and their un-manipulated test scores. We fit a third order polynomial to the data from the un-manipulated parts of the test score distribution, and then predict this relationship inwards into the manipulation regions. Specifically, denote by g_{ijt} is the earnings of

[41] While this is innocuous in our setting, where teachers have no incentive to grade down (see Section 2.2 and footnote 27), it limits the set of *other* applications for which our estimator could be used; however, while this is outside of the scope of the current paper, we conjecture that minor modifications to the estimator would yield bounds on treatment effects also in the presence of defiers.

student i (who is enrolled in region j in year t). We estimate:

$$g_{ijt} = \hat{g}_{kjt}\left(r_{ijt}, \theta_{kjt}^{grade}\right) + \alpha_{kjt} * (r_{ijt} \geq k) + \epsilon_{ijt}^{g}, \tag{2}$$

$$\text{where } (r_{ijt} < k - \beta_{kjt} \text{ or } r_{ijt} > k + \beta_{kjt} - 1) \text{ and}$$

$$r_{ijt} > (k-1) + \beta_{k-1jt} + 1 \text{ and } r_{ijt} < (k+1) - \beta_{k+1jt}.$$

Here, $\hat{g}_{kjt}\left(r_{ijt}, \theta_{kjt}^{grade}\right)$ is a third order polynomial with coefficients θ_{kjt}^{grade}, which captures the relationship between students' un-manipulated test scores, r_{ijt}, and their expected earnings. $(r_{ijt} < k - \beta_{kjt}$ or $r_{ijt} > k + \beta_{kjt} - 1)$ ensures that the data used to estimate equation (2) is outside of the manipulation region around test grade threshold k. $r_{ijt} > (k-1) + \beta_{k-1jt} + 1$ and $r_{ijt} < (k+1) - \beta_{k+1jt}$ ensures that the data is also not within the manipulation region around the higher $(k+1)$ or lower $(k-1)$ test grade threshold. We allow for a discrete jump in students' expected earnings at the test grade cut-off k, represented by $\alpha_{kjt} * (r_{ijt} \geq k)$, which represents the payoff of just passing the test in world with no test score manipulation. Equation (2) yields the expected earnings, at each test score point inside the manipulation region, *in a counterfactual world where no student receives test score manipulation.*

Next, we compute the counterfactual expected earnings across the *entire* set of students in each manipulation region. We do this by combining the estimates from equation (2) with our estimates of the counterfactual test score distribution, recovered in Section 5 above, which gives us the share of students who would have received each test score, had there been no test score manipulation, $\hat{h}_{jt}(r)$:

$$\begin{aligned} &E\left(\text{earnings in jt}|\text{teacher can't manipulate, r in manipulation region k}\right) \equiv \bar{g}_{jt}\left(k\right) \\ &= \int_{k-\beta_{kjt}}^{k+\beta_{kjt}-1} E\left(g_{kjt}|r, \text{No manip}\right) * \frac{Pr\left(r|\text{No manip, jt}\right)}{\int_{k-\beta_{kjt}}^{k+\beta_{kjt}-1} Pr\left(r|\text{No Manip, jt}\right)dr} dr. \\ &= \int_{k-\beta_{kjt}}^{k+\beta_{kjt}-1} \left[\hat{g}_{kjt}\left(r, \hat{\theta}_{kjt}^{grade}\right) + \hat{\alpha}_{kjt} * (r \geq k)\right] * \frac{\hat{h}_{jt}(r)}{\int_{k-\beta_{kjt}}^{k+\beta_{kjt}-1} \hat{h}_{jt}(r)dr} dr. \end{aligned} \tag{3}$$

For students inside the manipulation region, we now compare the estimated counterfactual average earnings, had there been no test score manipulation, calculated in (3), with the actual average earnings for students in the manipulation region (observed in the data), g_{ijt}. This difference is entirely driven by the fact that some students inside the manipulation region received test score manipulation. Thus, this difference is our "intent-to-treat" estimate of the average increase in a student's earnings due to the student having a raw test score that falls within the manipulation region of the test score distribution:

$$
\begin{aligned}
ITT = \quad & E\left(\text{earnings}|\text{teacher can manipulate}\right) - E\left(\text{earnings}|\text{teacher can't manipulate}\right) \\
= \quad & \underbrace{\frac{\sum_{jt}\left(\sum_{i\epsilon \text{manip region k}} g_{ijt}\right)}{\sum_{jt}\left(N_{kjt}^{\text{manip}}\right)}}_{\text{Average observed earnings across all students in manipulation region (across all j regions and t years)}} - \underbrace{\frac{\sum_{jt} N_{kjt}^{\text{manip}} \bar{g}_{jt}(k)}{\sum_{jt}\left(N_{kjt}^{\text{manip}}\right)}}_{\text{Average predicted earnings for students in manipulation region, had there been no manipulation (across all j regions and t years)}},
\end{aligned}
$$

where N_{jt}^{manip} is the number of students in the manipulation region around threshold k in region(*voucher) j in year t.

The "first stage" and LATE The procedure above can be repeated with math test grade instead of earnings. This yields our first stage effect of falling into the manipulation region on the test grade (intuitively, this captures the share of students whose grades are manipulated, i.e., the probability of being graded up). The ratio of the reduced-form effect on earnings to the first-stage effect, in turn, identifies the local average treatment effect (LATE) of receiving a manipulated math exam grade on future income.[42]

A comment on the level of aggregation This identification strategy is equally valid if we were to pool all regions and schools into an aggregate distribution. We would then define the manipulation region (around Pass and PwD, respectively) as the widest across all schools in the dataset, and only use information outside of this very wide region to extrapolate inward. This method would still lead to un-biased estimation, but would produce an estimate with a higher variance, since we would be throwing away information that could be used for the inward extrapolation from schools that engaged in less aggressive test score manipulation. By splitting the data into regions (by voucher, by year), we can identify places with narrower test score manipulation regions, which makes the extrapolation inward less noisy. As we slice the data into smaller and smaller aggregation units (e.g., if we were to split the data into school-level histograms instead of histograms at the county*voucher*year level), our estimates of where manipulation occurs become more noisy. This is because, when the histograms of the test score distributions contain fewer data points, the variance in our estimates of the manipulation windows increases. The choice of level of aggregation thus

[42] We block bootstrap the entire procedure to calculate standard errors, sampling at the county by voucher by year level. This is the same level at which we estimated the widths of the manipulation regions.

represents a trade-off of these two sources of variance. We chose to estimate the width of the manipulation regions at the county*voucher*year level to balance these two sources of variance.[43]

6.2 Identifying the beneficiaries of test score manipulation

Here, we develop a method to recover observable summary statistics of the types of students that teachers select to manipulate. This can be used more generally, in any type of bunching estimation, to characterize the types responding to the incentive to bunch at the threshold.

For any observable characteristic, Y, we can use students outside of the manipulation region to estimate $E(Y|r)$ at any test score r inside the manipulation region:

$$Y_{ijt} = \hat{g}^{Y}_{kjt}\left(r_{ijt}, \theta^{grade}_{kjt}\right) + \epsilon^{g}_{ijt}, \tag{4}$$

$$\text{where } (r_{ijt} < k - \beta_{kjt} \text{ or } r_{ijt} > k + \beta_{kjt} - 1) \text{ and}$$

$$r_{ijt} > (k-1) + \beta_{k-1jt} + 1 \text{ and } r_{ijt} < (k+1) - \beta_{k+1jt}.$$

For example, if Y is a dummy variable for being an immigrant, this yields an estimate of the expected share of immigrant children at each test score inside the manipulation region, had there been no test score manipulation. We can then calculate the actual share (observed in the data) of immigrant children in the manipulation region, above the cutoff threshold, $\bar{Y}^{up\_all}$, and below the cutoff threshold, $\bar{Y}^{down\_all}$:[44]

$$\bar{Y}^{up\_all} = \frac{1}{N^{tot}_{up}} \sum_{it} Y_{ijt} \text{ where: } k \leq t_{ijt} \leq k + \beta_{kjt} - 1, \tag{5}$$

$$\bar{Y}^{down\_all} = \frac{1}{N^{tot}_{down}} \sum_{it} Y_{ijt}, \text{ where: } k - \beta_{kjt} \leq t_{ijt} \leq k - 1. \tag{6}$$

Here, $\bar{Y}^{up\_all}_{t}$ is an average of students who were bumped up (the "compliers"), as well as students who naturally received a test score just above the threshold in the absence of manipulation ("always-takers"):

$$\bar{Y}^{up\_all} = \frac{N_{up}}{N_{up} + N_{compliers}} * \bar{Y}^{up} + \frac{N_{compliers}}{N_{up} + N_{compliers}} * \bar{Y}^{compliers}.\text{[45]} \tag{7}$$

[43] There likely is some alternative aggregation level which minimizes the variance of the estimates. Solving for the optimal level of aggregation to minimize the variance is left for future research.

[44] In the observed (manipulated) test score distribution, N^{tot}_{up} is the number of students who fall into the manipulation region above the passing threshold. N^{tot}_{down} is the number of students who fall into the manipulation region below the passing threshold.

[45] N_{up} is the number of students who earned an un-manipulated test score above the grade cutoff, within

Similarly, $\bar{Y}^{down\_all}$ is an average of those who selectively *not* were inflated to passing scores ("never-takers"):

$$\bar{Y}^{down\_all} = \frac{N_{down}}{N_{down} - N_{compliers}} * \bar{Y}^{down} - \frac{N_{compliers}}{N_{down} - N_{compliers}} * \bar{Y}^{compliers}.^{46} \tag{8}$$

We can recover the expected share of immigrant students within these regions of the distribution, using our extrapolation from equation (4) and the estimated un-manipulated distribution, $\hat{h}_{jt}(r)$:

$$\bar{Y}^{up} = \sum_j \left(N_j \int_k^{k+\beta jt-1} \hat{g}^Y_{kjt}\left(r, \theta^{grade}_{kjt}\right) * \hat{h}_{jt}(r)dr \right) \tag{9}$$

$$\bar{Y}^{down} = \sum_j \left(N_j \int_{k-\beta jt}^{k-1} \hat{g}^Y_{kjt}\left(r, \theta^{grade}_{kjt}\right) * \hat{h}_{jt}(r)dr \right). \tag{10}$$

Finally, the number of students within each region can be calculated as:

$$N^{tot}_{up} = N_{up} + N_{compliers}, \quad N^{tot}_{down} = N_{down} - N_{compliers},$$

$$N_{up} = \sum_j \left(N_j \int_k^{k+\beta jt-1} \hat{h}_{jt}(r)dr \right), \quad N_{down} = \sum_j \left(N_j \int_{k-\beta jt}^{k-1} \hat{h}_{jt}(r)dr \right).$$

Plugging these into equations (9) and (10) and solving for the mean immigrant share of the compliers gives:

$$\bar{Y}^{compliers} = 0.5 * \left(\frac{N^{tot}_{up}}{N^{tot}_{up} - N_{up}} * \bar{Y}^{up\_all} - \frac{N_{up}}{N^{tot}_{up} - N_{up}} \bar{Y}^{up} \right)$$

$$+0.5 * \left(\frac{N_{down}}{N^{down} - N^{tot}_{down}} \bar{Y}^{down} - \frac{N^{tot}_{down}}{N^{down} - N^{tot}_{down}} * \bar{Y}^{down\_all} \right).^{47}$$

Intuitively, if teachers disproportionately choose to manipulate the test scores of immigrant children, there will be an unexpectedly high share of immigrants right above the grade cutoff, and an unexpectedly low share of immigrants right below the grade cutoff, relative to what we would have expected from an extrapolation inwards into the manipulation re-

the manipulation region. These are the always-takers. $\bar{Y}^{up}$ is the average level of Y for the always-takers.

[46] N_{down} is the number of students who earned an un-manipulated test score below the grade cutoff, within the manipulation region. These are the never takers and the compliers. $\bar{Y}^{down}$ is the average level of Y for the never-takers and the compliers.

[47] $\bar{Y}^{compliers}$ can either be estimated by investigating what types of students are "missing" below the cutoff; or by investigating what types of students are found "in excess" above the cutoff. We estimate both, and average them together to increase power.

gion using the immigrant share outside of the manipulation region. We can compare the characteristics of the compliers, $\bar{Y}^{compliers}$, with the characteristics of all students whose unmanipulated test scores fell within the manipulation region of the test score distribution below the test grade threshold (that is, all students who were "eligible" for manipulation), $\bar{Y}^{down}$, to assess whether teachers were targeting their manipulation at certain types of students:

$$\Delta Y = \bar{Y}^{compliers} - \bar{Y}^{down}.$$

7 Results

7.1 Who receives test score manipulation?

There are number of criteria that teachers may use to select which students' test scores to bump up above the grade thresholds. They may choose students whom they deem would have the largest benefit; they may choose students who come from disadvantaged backgrounds; or they may choose the students who simply had a bad day on the test, but who have performed at a higher level in class. It also possible that teachers inflate the most pushy or grade grabbing students, in order to minimize future disagreement with those students (or their parents). To shed light on teachers' selection criteria, we use the methods described in Section 6.2 to analyze the observable characteristics of students who are chosen to be graded up, and compare them to all students who fall within the relevant manipulation region and thus *could have been chosen* for manipulation (we refer to the latter group as students who were "eligible" for manipulation).

For a set of predetermined outcomes, Table 2 compares the average among all eligible students (Column one) with the average for the subset students whose test scores were inflated (Column two). The first two outcomes are the test grade on the national tests in Swedish and English, both of which are taken before the national test in math and hence cannot be influenced by the outcome on the math test. (In the next subsection, we perform a "sanity check" that verifies that the national math test indeed has no impact on the results on the national tests in English and Swedish). Around the Pass margin, we see that inflated students are 7.4 percentage points more likely (than the average eligible student) to have passed their national test in English. Around the the PwD margin, manipulated students are 33 percentage points more likely than eligible students to have received a high grade on the test in English. In the subsequent row, we see a similar pattern when we look at the students' Swedish test grades: manipulated students are positively selected on their pre-determined Swedish test grade.

If teachers were grading up students who had a bad test day, we would expect them to choose to bump up students who have higher grades on other, pre-determined tests than the average student who is eligible for manipulation. This is consistent with the observed selection of students for manipulation based on the pre-determined test grades. Teacher discretion thus appears to be correcting for idiosyncratically poor performance on the math test, given what can be expected based on previous achievement. This may be a desirable outcome, compared to a high-stakes testing environment that sorts students who fall close to the Pass and PwD margins solely based on their idiosyncratic performance on the test day. This suggests that, to the extent that the math test grade carries long-term consequences – a question that we analyze in the next subsection – this type of teacher discretion may be desirable.

Turning to whether teachers' manipulation choices are related to students' demographics, the next row in Table 2 compares the male share of students eligible for manipulation with the male share of manipulated students. We see a precisely estimated zero effect around both thresholds, showing that teachers treat boys and girls equally when choosing whom to bump up. Similarly, the next row of Table 2 shows that manipulated students are not selected based on whether they come from an immigrant household. These results suggest that teachers do not appear to bias their math test grading based on race or gender.

Next, we turn to whether manipulated students come from disadvantaged backgrounds. Around the Pass margin, manipulated students are positively selected on household income: manipulated students' household income is 3.9 percent higher than the household income of the average eligible student. Around the PwD margin, in contrast, the point estimate implies that household incomes of manipulated students are 3.2 percent *lower* than the household income of the average eligible student; however, the effect is not statistically significant. We find similar patterns of selection on fathers' years of education, presented in the subsequent row of Table 2. Manipulated students around the Pass margin have fathers with 0.072 more years of education; however, there is no statistically significant selection effect around the PwD margin. The selection on income and education around the Pass margin is somewhat worrying, as it could exacerbate inequality of opportunity between rich and poor students. However, the point estimate is economically quite small. Moreover, it could be driven by the fact that teachers grade up students who had a bad day on the test; as shown above, these students are higher achievers on previous tests. Thus, to the extent that higher achievement is correlated with income, the estimated effect on parental income may reflect the fact that teachers are targeting students who are truly higher achievers – which happens to be correlated with parental income – rather than targeting income per se.

To analyze whether teachers manipulate students whose parents may have more free time to pressure teachers into giving their children high grades, we look at whether manipulation is selected on whether the student has a stay at home parent. The last row of Table 2 shows a zero effect of selectively bumping up students with a stay at home parent around both thresholds, with negative point estimates. This suggests that if anything, the teachers are more likely to inflate students without a stay at home parent.

Taking all of these dimensions of selection together, it appears that, both at the low and high ends of the ability distribution, teachers primarily help students who had a bad day on the test, as indicated by their achievement on predetermined tests. This suggests that the teachers use their discretion to "undo" having a bad day on the test.

7.2 The long-term consequences of test score manipulation

Table 3 presents results from our first stage, where we compare *the expected final math grade absent manipulation* to *the average observed math grade*, inside the manipulation region of the test score distribution. The coefficient quantifies how much "getting a raw test score that falls into the manipulation region" raises the probability of receiving a higher final math grade (due to test score manipulation).

Around the Pass threshold, falling into the manipulation region of the test score distribution raises the probability of obtaining a higher final grade by 5.5 percentage points.[48] Around the PwD threshold, falling into the manipulation region raises the probability of getting a higher final grade by 10 percentage points. All estimates are statistically significant at the 1 percent level. The F-statistic is far above the conventional level of 10.

These estimates represent the average effects of manipulation on students within the manipulation region; hence, they represent intent-to-treat effects on the final grade. But as predicted by our model in Section 4, only a subset of the students in the manipulation regions are *de facto* manipulated; thus, the students that receive manipulation ("the compliers") are experiencing a larger gain in the final grade than the intent-to-treat estimate. Below, when we turn to our sanity checks and main outcomes, we present LATE estimates, which capture the treatment effect of manipulation on the subset of students who are graded up, that is, on the compliers.

Before turning to the sanity checks and results on our main outcomes, however, we discuss an alternative first stage specification. Test score manipulation leads to a direct change in

[48] We recall that the final grade in math takes the value of 0 if the student's grade is Fail; 1 if the grade is Pass; 2 for PwD; and 3 for Excellent. In the overall sample, the average grade is 1.13. Around the Pass threshold, the average grade is .99, reflecting the fact that most of the variation around this threshold stems from whether of not the student receives a Pass. We have also run our first stage using an indicator variable for whether the student receives a grade of Pass or higher (and PwD or higher, respectively), and the results look similar.

the math test grade (awarded in February), which ultimately can lead to a change in the student's final math grade (awarded in June). We use the final math grade as the endogenous variable of interest when analyzing longer-term outcomes – so, the estimates presented in Table 3 represent our first stage estimates – but one could also use the math test grade (awarded in February). Appendix Table B1 estimates the impact of receiving a manipulated math test grade on the final math grade. Around the Pass margin, receiving a manipulated test grade leads to a 35 percentage point increase in the probability of receiving a passing final grade, and around the PwD margin, a manipulated test grade raises the likelihood of receiving a higher final math grade by 87 percentage points. These effects are not 100 percent because, as discussed in Section 2 above, the teacher takes into account more than the math test grade when assigning the final grade, including students' classroom performance. If the reader prefers to view the endogenous variable of interest as the math test grade, instead of the final grade in math, then simply multiply the treatment effects by these estimated effects.

Before turning to the long-term outcomes, we perform several sanity checks to validate our methodology. We first estimate the causal effect of receiving a higher final math grade (through teachers' discretion) on characteristics of the students that are pre-determined at the time of the math test. Clearly, we know that test score manipulated cannot change their grades on previous tests. We expect to see that our estimator finds this to hold. Appendix Table B2 shows that for both the Pass and PwD thresholds, there is no causal effect of manipulation on the test grade on the English nationwide test, which is taken before the nationwide math test. We find similar zero effects on students' (predetermined) Swedish test grades (Appendix Table B3). Panel B of the two tables report the simple OLS relationship between these outcomes and dummy variables indicating students' final math grades, controlling for county*voucher*year fixed effects. Unlike in our placebo tests, we see very strong correlations between students' final math grades and their test grades in other subjects. The fact that our identification strategy breaks these very strong OLS correlations in the data provide confidence in our estimation methods.

The first outcome that we consider, grade nine GPA, captures student performance in the immediate future following the nationwide math test. GPA in grade nine is calculated based on the average of the final grade in math and other subjects, and is awarded in June of the final year before high school, i.e., within four months of the nationwide math test. Grade nine GPA ranges from zero to 320. Table 4 presents the LATE for the outcome GPA. Around the Pass threshold, exposure to manipulation raises the GPA by 10.6 points for those who are graded up, or by roughly 6 percent of the mean GPA around the threshold (177). Around the PwD threshold, manipulation raises GPA by 21.4 points for those who are graded up, or by approximately 9 percent. The direct effect of receiving a higher math grade *mechanically* increases a student's GPA by 10 points when moving from Fail to Pass, and by 5 points when moving from Pass to PwD. Around the Pass threshold, we cannot reject that the effect is equal to a 10 point increase in the student's GPA. Around the PwD threshold, however, the results suggest that there is a motivational effect, since manipulation raises the student's performance substantially over and above the mechanical effect induced by test score manipulation. Receiving a PwD on the math test thus either encourages the students to work harder in their *other* classes, or their other teachers to choose to inflate them as well, on future tests and assignments.

An alternative interpretation is that the extra effort is driven by a change in actual incentives, rather than in motivation. For example, if students get utility from moving on to high school or teachers get utility from students who move on to high school, then a passing grade on the nationwide math test – which influences the final grade in math and potentially the probability of going to high school – changes the return to effort (or lenient grading) going forward. As we shall see below, however, test score manipulation has a large impact on high school attendance (graduation) only when it moves a student across the Pass threshold; in contrast, the impact on high school graduation is much smaller around the PwD threshold. Thus, the GPA response is concentrated among higher ability students, for whom the high school incentive effect is unimportant.

The effect on GPA highlights that there is a strong signaling value from the math test grade at the *higher* end of the ability distribution: Receiving a higher grade signals to the student and potentially to his or her teachers that the student's ability is higher, and this appears to be complementary with increased effort on the part of the student, or more generous grading in other subjects on the part of other teachers. Panel B compares these estimates to the OLS relationship between math grades and overall GPA. These point estimates are much larger, showing that students who pass math have 82.2 higher GPAs than those who fail. Going from Pass to PwD is associated with 50.6 more GPA points. These OLS results further highlight the fact that math grades are highly endogenous in the cross-section.

We then examine a set of outcomes measured at the end of high school, three years after test score manipulation. Table 5 presents results on high school graduation by age 19 (i.e., "on time"). We find that test score manipulation that pushes a student above the Pass threshold raises his or her probability of finishing high school on time (by age 19) by 20 percentage points. The large impact at the lower end of the ability distribution is consistent with our finding that test score manipulation around the lower threshold raises the student's likelihood of receiving a passing final grade in math (awarded in June in the last year before high school), which is a necessary condition for admittance to *any* high school (other than one-year remedial programs that serve to get the student ready for a three-year high school program with a one year delay). However, this magnitude is smaller than the OLS relationship in the cross-section: Panel B shows that passing math class is associated with a 53.8 percentage point increase in on time high school graduation. Around the PwD threshold, test score manipulation increases the probability of on time high school graduation by 5.5 percentage points, a 6 percent increase over the base mean of 87 percent. This much smaller effect is likely driven by the fact that most of the students at this higher point in the ability distribution would proceed to high school directly after grade nine, regardless of whether they get a Pass or PwD in math. Further, this point estimate is smaller than the observed OLS relationship, an 11.8 percentage point increase.

To analyze whether inflated students perform better in high school, Table 6 reports effects on high school GPA (measured in the last year of high school), among students who complete high school. This is interesting to analyze because a student who is inflated in ninth grade may be at risk of obtaining *lower* grades in high school, as the student may be tracked with better high school peers (Malamud and Pop-Eleches, 2011). Interestingly, however, we do not find any statistically significant negative effects of test score manipulation in grade nine on high school GPA. On the contrary, among students at the lower end of the ability distribution, test score manipulation appears to raise high school GPA. Specifically, we find that manipulation over the Pass threshold causes a 1.4 point higher high school GPA, relative to a base of 11.9. Manipulation over the PwD threshold has a similar effect, with the point estimate suggesting an increase in GPA of 1 point, relative to a mean of 14. This further highlights the fact that the signaling value of a higher math test grade in the last year before high school can substantially change future human capital investment decisions. A possible alternative explanation is that manipulated students have enrolled in different high schools that give *all* students better grades. To test this, we analyze whether receiving an inflated math grade causes a higher *peer* high school GPA. Appendix Table B4 shows that receiving an inflated math grade in grade nine does not increase the GPA of one's peers in high school. This further substantiates that the positive impacts on one's own GPA likely arises through an effort and human capital investment margin.

Our last set of outcomes captures student well-being eight years after test score manipulation. Table 7 reports impacts of test score manipulation on the probability of enrolling and initiating college by age 23. Our point estimates are economically significant, with manipulation around the Pass threshold leading to a 12 percentage point increase (which represents a 86% increase, relative to the mean of 14%) in the probability of initiating college. Interestingly, we find a similar estimate of 16 percentage points in the OLS cross-section between passing math and college initiation. We cannot reject that the two effects are equal. Around the PwD threshold, we find a slightly smaller effect, 7.9 percentage points; however, this estimate is noisy and we cannot reject zero. Nonetheless, the point estimate of 7.9 is economically meaningful, relative to a mean of 38%; moreover, it is much smaller than the OLS relationship of 25.8.

When looking at total years of completed education, we find effects around both thresholds. Table 8 shows that students who get manipulated to a Pass (PwD) have 0.33 (0.48) more years of education by age 23. This is equivalent to a 3-4% increase relative to the mean years of education for these groups. This is a much smaller effect than the observed OLS relationship of a passing grade leading to 1.3 more years of schooling and a PwD leading to an additional 0.77 years of schooling (relative to a passing grade).

Thus, around both margins, inflated students are more likely to remain in school for longer. Through this channel, test score manipulation in grade nine may also help students "stay on track" and avoid outcomes that force them to drop out of school, such a teen pregnancy. Table 9 shows that, indeed, inflated students around the Pass threshold are 0.027 percentage points less likely to have a teen birth. This is a large (although marginally insignificant) effect, relative to the mean of 0.014, suggesting that the students chosen for inflation were particularly at risk for having a teen birth. We see a similar relationship in the OLS estimates. We see similar, large and statistically significant effects around the PwD threshold, where test score manipulation lowers the teen birth rate by 3.5 percentage points.

Our final long-term outcome captures income at age 23 (the end of our sample period). Table 10 shows that being graded up above the Pass threshold in grade nine raises age 23 income, with a point estimate of 34000 SEK, relative to a mean of 158000. (Recall that our income variable is measured in 100 SEK.) 100 SEK is roughly $10; thus, this corresponds to a point estimate of roughly $ 3400, relative to a mean of $ 15800. This is a large, 20% increase in earnings at age 23, and it is quite similar to the OLS relationship of 37000 SEK. However, the mean is quite low, since many of these students are still in school. Further, this mean income is for *all* students in the manipulation region, not just for the compliers. Given the nature of selection into being a complier, their mean income may indeed be much higher at this stage of life, absent manipulation. Students inflated over the PwD margin receive a 44800 SEK higher age 23 income, relative to a base mean of 146100. This is quite different than the OLS relationship which shows an income *decrease* of 19300 SEK. This discrepancy highlights that many of these students are still in school, which depresses the group's average labor market earnings. However, since the increase in years of education due to manipulation is quite small, this "still in school"-effect is likely less reflected in the LATE estimates than in the OLS estimates.

8 Conclusion

Despite the fact that test score manipulation does not, per se, raise human capital, this paper demonstrates that its beneficiaries receive large, long-term gains in educational attainment and earnings. The mechanism at play suggests important dynamic complementarities: Getting a higher grade on a high-stakes test can serve as an immediate signaling mechanism *within the educational system*, motivating students and potentially teachers; this, in turn, can raise human capital; and the combination of higher effort and higher human capital can ultimately generate substantial labor market gains.

The large benefits that accrue to the beneficiaries of test score manipulation of those who have “a bad test day” suggest that teachers may find it privately desirable to err on the side of giving their students higher grades, and to thereby improve their students’ outcomes. But although each teacher’s adjustments to his or her students’ test scores would not affect the nationwide grade distribution, the combined effect of many teachers’ manipulation may shift the grade distribution upwards. This suggests that this paper may have identified a micro-mechanism contributing to grade inflation, an increasingly pervasive problem in Scandinavia as well as in the U.S.. In turn, this suggests that, while test score manipulation may be privately optimal from the perspective of each teacher, it may be socially undesirable if grade inflation induces distortionary general equilibrium effects.

Moreover, the fact that we see large regional variation in test score manipulation, and some differences between municipal and voucher schools, suggests that teacher discretion undermines the equality of opportunity in Swedish schools: students who live in a region with substantial test score manipulation are more likely to get inflated, and thereby more likely to enjoy the benefits shown in this paper. Exploring the roots of these differences in grading leniency, as well as the general equilibrium effects of test score manipulation, are left for future work.

References

Angrist, Joshua D. and Alan B. Krueger, “Does Compulsory School Attendance Affect Schooling and Earnings?,” *The Quarterly Journal of Economics*, 1991, *106* (4), 979–1014.

Angrist, Joshua D and Miikka Rokkanen, “Wanna get away? Regression discontinuity estimation of exam school effects away from the cutoff,” *Journal of the American Statistical Association*, 2015, *110* (512), 1331–1344.

Apperson, Jarod, Carycruz Bueno, and Tim R Sass, “Do the Cheated Ever Prosper? The Long-Run Effects of Test-Score Manipulation by Teachers on Student Outcomes,” *mimeo*, 2016.

Bandiera, Oriana, Valentino Larcinese, and Imran Rasul, “Blissful ignorance? A natural experiment on the effect of feedback on students’ performance,” *Labour Economics*, 2015, *34*, 13 – 25. European Association of Labour Economists 26th Annual Conference.

Bar, Talia, Vrinda Kadiyali, and Asaf Zussman, “Grade Information and Grade Inflation: The Cornell Experiment,” *Journal of Economic Perspectives*, 2009, *23* (3), 93–108.

Böhlmark, A and M. Lindahl, “Har den växande friskolesektorn varit bra för elevernas utbildningsresultat på kort och lång sikt?,” *IFAU Rapport*, 2012, *17* (2).

Brunello, Giorgio, Margherita Fort, and Guglielmo Weber, “Changes in Compulsory Schooling, Education and the Distribution of Wages in Europe*,” *The Economic Journal*, 2009, *119* (536), 516–539.

Butcher, Kristin F., Patrick J. McEwan, and Akila Weerapana, “The Effects of an Anti-grade-Inflation Policy at Wellesley College,” *Journal of Economic Perspectives*, 2014, *28* (3), 189–204.

Cameron, Stephen V and James Heckman, “The Nonequivalence of High School Equivalents,” *Journal of Labor Economics*, 1993, *11* (1), 1–47.

Canaan, Serena and Pierre Mouganie, “Returns to Education Quality for Low-Skilled Students: Evidence from a Discontinuity,” *Journal of Labor Economics*, Forthcoming.

Chen, Z., Z. Liu, D. Xu, and J-C. Suarez-Serrato, “Notching RD Investment with Corporate Income Tax Cuts in China,” *mimeo, Duke University*, 2017.

Chetty, Raj, John N. Friedman, and Emmanuel Saez, “Using Differences in Knowledge Across Neighborhoods to Uncover the Impacts of the EITC on Earnings,” *The American Economic Review*, 2013, *103* (7), 2683–2721.

_ , John N Friedman, Tore Olsen, and Luigi Pistaferri, “Adjustment Costs, Firm Responses, and Micro vs. Macro Labor Supply Elasticities: Evidence from Danish Tax Records,” *Quarterly Journal of Economics*, 2011, *126* (2), 749–804.

Clark, Damon and Paco Martorell, “The Signaling Value of a High School Diploma,” *Journal of Political Economy*, 2014, *122* (2), 282 – 318.

Cullen, Julie and Randall Reback, “Tinkering Toward Accolades: School Gaming Under a Performance Accountability System,” *NBER Working Paper No. 12286*, 2006.

Dee, Thomas S., Brian A. Jacob, Justin McCrary, and Jonah Rockoff, “Rules and Discretion in the Evaluation of Students and Schools: The Case of the New York Regents Examinations,” *mimeo*, 2011.

Dee, Thomas S, Will Dobbie, Brian A Jacob, and Jonah Rockoff, “The Causes and Consequences of Test Score Manipulation: Evidence from the New York Regents Examinations,” *National Bureau of Economic Research*, 2016.

Dong, Yingying and Arthur Lewbel, “Identifying the effect of changing the policy threshold in regression discontinuity models,” *Review of Economics and Statistics*, 2015, *97* (5), 1081–1092.

Ebenstein, Avraham, Victor Lavy, and Sefi Roth, “The Long-Run Economic Consequences of High-Stakes Examinations: Evidence from Transitory Variation in Pollution,” *American Economic Journal: Applied Economics*, forthcoming.

Figlio, David and Lawrence Getzler, “Accountability, Ability and Disability: Gaming the System,” *NBER Working Paper No. 9307*, 2002.

Gerard, Francois, Christoph Rothe, and Miikka Rokkanen, “Bounds on Treatment Effects in Regression Discontinuity Designs with a Manipulated Running Variable, with an Application to Unemployment Insurance in Brazil,” *NBER Working Paper # 22892*, 2016.

Golsteyn, Bart H. and Anders Stenberg, “Earnings over the Life Course: General versus Vocational Education,” *Mimeo*, 2015.

Hinnerich, Björn Tyrefors and Jonas Vlachos, "Systematiska skillnader mellan interna och externa bedömningar av nationella prov – en uppföljningsrapport," *Skolverket*, 2013.

Hvidman, Ulrik and Hans Henrik Sievertsen, "Grades and Student Behavior," *Mimeo*, 2017.

Imbens, Guido W. and Joshua D. Angrist, "Identification and Estimation of Local Average Treatment Effects," *Econometrica*, 1994, *62* (2), 467–475.

J., Hurwitz-M. Smith and C. Avery, "Giving college credit where it is due: Advanced placement exam scores and college outcomes.," *Journal of Labor Economics*, 2017, *35* (1).

Jacob, Brian A., "Accountability, incentives and behavior: the impact of high-stakes testing in the Chicago Public Schools," *Journal of Public Economics*, June 2005, *89* (5-6), 761–796.

Jacob, Brian A and Lars Lefgren, "Remedial Education and Student Achievement: A Regression-Discontinuity Analysis," *The Review of Economics and Statistics*, 2006, *86* (1), 226–244.

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

Jaeger, David A. and Marianne E. Page, "Degrees Matter: New Evidence on Sheepskin Effects in the Returns to Education," *The Review of Economics and Statistics*, 1996, *78* (4), 733–740.

Kane, Thomas J. and Cecilia E Rouse, "Labor-Market Returns to Two- and Four-Year College," *The American Economic Review*, 1995, *85* (3), 600–614.

_ , _ , and Douglas Staiger, "Estimating Returns to Schooling when Schooling is Misreported," *NBER Working Paper*, 1999, (7235).

Keys, Benjamin J., Tanmoy Mukherjee, Amit Seru, and Vikrant Vig, "Did Securitization Lead to Lax Screening? Evidence from Subprime Loans," *The Quarterly Journal of Economics*, 2010, *125* (1), 307–362.

Kleven, Henrik, "Bunching," *Annual Review of Economics*, 2016, *8*, 435–464.

Kleven, Henrik J and Mazhar Waseem, "Using Notches to Uncover Optimization Frictions and Structural Elasticities: Theory and Evidence from Pakistan," *Quarterly Journal of Economics*, 2012, *128* (2), 669–723.

Lavy, Victor, "Performance Pay and Teachers' Effort, Productivity, and Grading Ethics," *American Economic Review*, 2009, *99* (5), 1979–2011.

_ and Edith Sand, "On the origins of gender human capital gaps: Short and Long Term Consequences of Teachers' Stereotypical Biases," *NBER Working Paper No. 20909*, 2015.

Malamud, Ofer and Cristian Pop-Eleches, "School tracking and access to higher education among disadvantaged groups," *Journal of Public Economics*, 2011, *95* (11 - 12), 1538 – 1549.

Manacorda, Marco, "The Cost of Grade Retention," *The Review of Economics and Statistics*, 2012, *94* (2), 596–606.

Marx, Benjamin M., "Dynamic Bunching Estimation and the Cost of Reporting Regulations for Charities," *mimeo, University of Urbana Champaign*, 2015.

McCrary, Justin, "Manipulation of the running variable in the regression discontinuity design: A density test," *Journal of Econometrics*, 2008, *142* (2), 698–714.

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

Oreopoulos, Philip, "Do dropouts drop out too soon? Wealth, health and happiness from compulsory schooling," *Journal of Public Economics*, 2007, *91* (11â12), 2213 – 2229.

Papay, John P, Richard J Murnane, and John B Willett, "The Impact of Test-Score Labels on Human-Capital Investment Decisions," *Journal of Human Resources*, 2015.

Persson, Petra, "Social Insurance and the Marriage Market," *mimeo*, 2014.

Saez, Emmanuel, "Do Taxpayers Bunch at Kink Points?," *American Economic Journal: Economic Policy*, August 2010, *2* (3), 180–212.

Terry, Stephen J., "The Macro Impact of Short-Termism," *Mimeo*, 2016.

Tyler, John H., Richard J. Murnane, and John B. Willett, "Estimating the Labor Market Signaling Value of the GED," *The Quarterly Journal of Economics*, 2000, *115* (2), 431–468.

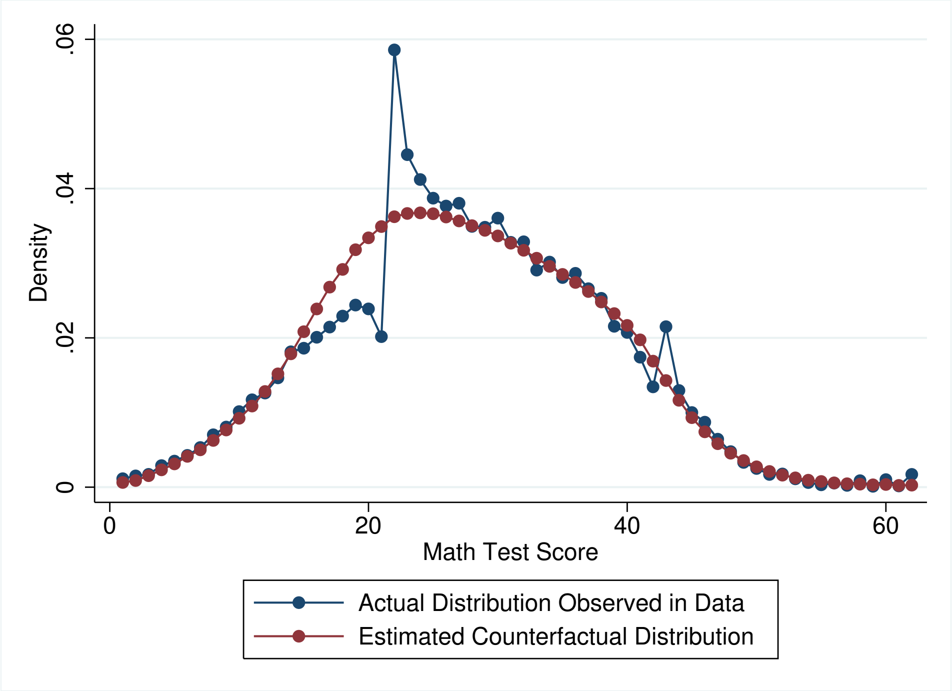
Vlachos, Jonas, "Betygets värde. En analys av hur konkurrens påverkar betygssättningen vid svenska skolor," *Uppdragsforskningsrapport 2010:6, Konkurrensverket*, 2010.

Wang, Jiangdian and SK Ghosh, "Shape restricted nonparametric regression with Bernstein polynomials," *Computational Statistics & Data Analysis*, 2012, *56* (9), 2729–2741.

Wuepper, David and Travis Lybbert, "Perceived Self-Efficacy, Poverty and Economic Development," *Annual Review of Resource Economics*, forthcoming.

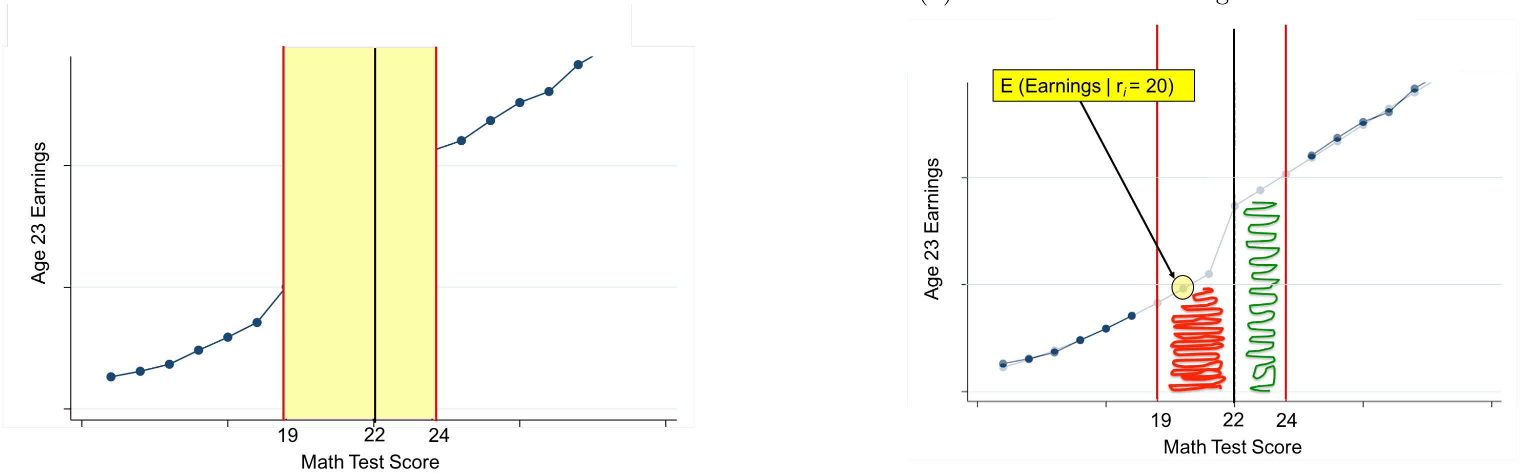
9 Figures and Tables

Figure 1: National Test Score Distribution and Aggregated Counterfactual, 2010



Note: The figure illustrates the national test score distribution in 2010 and the counterfactual (aggregated from the county*voucher estimated counterfactuals for 2010). The estimation of the counterfactual density (at the county*voucher*year level) is described in detail in Section 5. The blue connected line plots the observed (manipulated) distribution of test scores, and the red connected line shows the estimated (un-manipulated) counterfactual density. The counterfactual tells us where manipulation occurs: At the national level, the lowest test score at which any student is manipulated is exactly 7 points below the pass threshold – by definition, this is the cutoff for manipulation in the (county*voucher) with *the most* manipulation in 2010. This is rare, however; the modal width of the manipulation region is 3 test score points around the Pass threshold and 1 test score point around the PwD threshold. In addition to telling us where manipulation occurs, the key piece of information that we get from the (county*voucher*year) counterfactual densities is, for each test score point inside a manipulation region, a count of how many students would have gotten that score in a counterfactual scenario without manipulation. This "head count" (for each county*voucher*year) is one of two critical inputs into our estimation of causal effects.

Figure 2: Estimation of Expected Counterfactual Outcomes: Intuition



(a) Un-manipulated earnings data around Pass threshold

(b) Counterfactual earnings around Pass threshold

Note: An example (county*voucher*year) where the Pass threshold is 22 and the manipulation region (around Pass) is 19-24. Our Wald estimator is constructed in three steps, and these two panels illustrate the first of these three steps: estimating what would have been the average outcome (e.g., earnings), *of all students* scoring 19-24 in a counterfactual scenario without manipulation. That is, we want to construct $\mathbf{E}(Earnings|No\ manipulation)$ for all students inside the manipulation region (not only those who were manipulated). Figure 2a depicts average age 23 earnings, by test score, outside of the manipulation region. Because all these students are un-manipulated, this data reflects the true, underlying relationship between test scores and (average) earnings. We fit a flexible polynomial to this data and predict the relationship inside the manipulation region, allowing for a true effect on earnings of passing the threshold. Intuitively, in a world without manipulation, a certain number of students would have scored 20 on the test; then, at age 23, they would have obtained some earnings. The point highlighted in yellow in Figure 2b shows our prediction of the average age 23 earnings of all students scoring 20 in a world without manipulation. For each test score point inside the manipulation region, we obtain an analogous prediction. Even though we do not know precisely which of the students inside the manipulation region would have scored 20 in a world without manipulation, we have an estimate of *how many* students would have scored 20 – our counterfactual test score distribution (for the same county*voucher*year) gives us precisely this "head count." So, we have everything that we need to construct $\mathbf{E}(Earnings|No\ manipulation)$ for all n students inside the manipulation region: $\mathbf{E}(Earnings|No\ manipulation) = {}^1/_n \sum_{s=19}^{24} \mathbf{E}(Earnings|r_i = s)^* Headcount(s)$. That is, we simply multiply each grey predicted point inside the manipulation region, depicted in Figure 2b, with the "head count" in that test score point given by the counterfactual density, and sum it up.

Table 1: Summary Statistics

	Overall	Pass Region	PwD Region
Math Test Score	28.4	22.6	40.6
Father Foreign Born	0.22	0.23	0.18
Household Income	4772.2	4421.3	5561.2
Male	0.51	0.52	0.50
Father's Years of Education	11.8	11.6	12.2
Has Non-Working Parent	0.25	0.27	0.21
Math Final Grade	1.13	0.99	1.62
Math Test Grade	0.88	0.71	1.39
English Test Grade	1.50	1.32	1.88
Swedish Test Grade	1.33	1.18	1.67
Overall Grade Point Average	191.6	177.1	227.2
High School Graduate (for 2004-2009 Pupils)	0.76	0.73	0.87
Initiated College (for 2004-2005 Pupils)	0.061	0.040	0.099
Years of Education (for 2004-2005 Pupils)	12.0	11.8	12.5
High School GPA (for 2004-2006 Pupils)	12.8	11.9	14.0
Teen Birth (for 2004-2005 Pupils)	0.0057	0.0073	0.0021
Age-23 Labor Income (for 2004-2005 Pupils)	1517.8	1579.9	1461.2
Observations	490519	114049	64397

Note: Our baseline sample consists of all students who attended ninth grade between 2004 to 2010 and both took the national test and obtained a final grade in math. For variables that are measured at a certain duration after graduation from ninth grade, we only include the cohorts that we observe at that duration (see the text for more details). Income is measured in 100 SEK (roughly $10). See text for further details defining the two subpopulations around the two test score grading thresholds.

Table 2: Who Benefits From Teacher Discretion?

	Eligible for Inflation	Inflated	Difference
English Test Grade			
Pass/Fail Margin	1.07	1.15	0.074***
	(0.064)	(0.055)	(0.017)
Pass/PWD Margin	1.40	1.73	0.33**
	(0.14)	(0.068)	(0.14)
Swedish Test Grade			
Pass/Fail Margin	0.96	1.02	0.060***
	(0.057)	(0.051)	(0.014)
Pass/PWD Margin	1.24	1.60	0.35***
	(0.13)	(0.068)	(0.13)
Share Male			
Pass/Fail Margin	0.51	0.51	-0.0045
	(0.0030)	(0.0081)	(0.0097)
PWD/Pass Margin	0.51	0.49	-0.019
	(0.0028)	(0.030)	(0.032)
Share Foreign Background			
Pass/Fail Margin	0.25	0.24	-0.0065
	(0.0072)	(0.0094)	(0.0077)
Pass/PWD Margin	0.18	0.13	-0.046
	(0.0068)	(0.038)	(0.037)
Household Income			
Pass/Fail Margin	4272.3	4439.4	167.1***
	(57.3)	(70.2)	(45.6)
PWD/Pass Margin	5389.6	5218.8	-170.7
	(96.3)	(482.0)	(477.1)
Father's Years of Education			
Pass/Fail Margin	11.5	11.6	0.072*
	(0.023)	(0.047)	(0.043)
PWD/Pass Margin	12.1	12.3	0.20
	(0.034)	(0.23)	(0.24)
Having Stay At Home Parent			
Pass/Fail Margin	0.28	0.27	-0.011
	(0.0050)	(0.0075)	(0.0070)
PWD/Pass Margin	0.22	0.16	-0.056
	(0.0065)	(0.035)	(0.040)

* p < 0.10, ** p < 0.05, *** p < 0.01.

Note: To shed light on teachers' selection criteria, this table presents observable, pre-determined characteristics of the students that teachers select to grade up, and compare these to the characteristics of all students who could have been chosen for inflation by the teacher. Specifically, Column 1 presents the predicted mean characteristic of all students whose un-manipulated math test score falls in the manipulation region of the test score distribution and thus could have been chosen by teachers to receive an inflated grade (all students eligible for inflation), $\bar{Y}^{down}$. Column two presents the predicted mean characteristic among the compliers, i.e., the students who were actually chosen to receive inflation, $\bar{Y}^{compliers}$. Column three tests the difference. To obtain the predictions, we use use students outside the manipulation region to estimate the expected characteristic, at any test score r inside the manipulation region, and then use the method described in detail in Section 6.2 to calculate $\bar{Y}^{down}$ and $\bar{Y}^{compliers}$. Standard errors that are block bootstrapped at the county*voucher*year level in parentheses.

Table 3: First stage: Impact of Inflation on Final Math Grade

	Pass	PWD
Change in Final Math Grade	0.055***	0.10***
	(0.0031)	(0.0085)
Fstat	317.3	141.6

* p < 0.10, ** p < 0.05, *** p < 0.01.

Note: The table presents estimates of the impact of exposure to inflation on the final math grade (on everyone in the manipulation region; though this estimate is in practice driven by the impact on those who are graded up, i.e., the compliers). The sample includes all cohorts in our sample, i.e., all students who attend ninth grade between 2004 and 2010. The predicted final grade absent manipulation is estimated from regressions of students' final grades on a dummy for whether the test score is above the cutoff and 3rd order polynomials in the test score, for each year and county*voucher. These regressions only use data from students outside of the manipulation regions of the test score distribution. See the text for more details. Standard errors block bootstrapped at the county*voucher*year level in parentheses.

Table 4: Impact of Grade Inflation on GPA (LATE)

	Panel A. Causal Impact Estimate	
	Pass	PWD
Δ Final Math Grade	10.6***	20.4***
	(4.10)	(5.47)
F Stat	317.3	141.6
Dep Variable Mean	177.1	227.2

	Panel B. OLS Estimate
Pass	82.19***
	(0.558)
PWD	132.8***
	(0.699)
Observations	488707

* p < 0.10, ** p < 0.05, *** p < 0.01.

Note: Panel A presents estimates of the impact of receiving a higher final math grade due to teachers' discretionary grading on GPA in grade nine. This should, mechanically, be equal to 10 around the Pass margin and 5 around the PwD margin if all that test score manipulation does is to raise the final grade in math (given how GPA is calculated in Sweden) and manipulation does not encourage or discourage student effort or teacher grading in other subjects. Panel B displays the OLS estimate. The sample includes all cohorts in our sample, i.e., all students who attend ninth grade between 2004 and 2010. Standard errors block bootstrapped at the county*voucher*year level in parentheses.

Table 5: Impact of Grade Inflation on High School Graduation (LATE)

	Panel A. Causal Impact Estimate	
	Pass	PWD
Δ Final Math Grade	0.20*** (0.044)	0.055* (0.034)
F Stat	308.6	185.4
Dep Variable Mean	0.73	0.87

	Panel B. OLS Estimate
Pass	0.538*** (0.00641)
PWD	0.656*** (0.00709)
Observations	409295

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Note: Panel A presents estimates of the impact of receiving a higher final math grade due to teachers' discretionary grading on the likelihood of high school graduation on time, i.e., within 3 years of ninth grade. Panel B displays the OLS estimate. The sample includes all students who attend ninth grade between 2004 and 2009, who are 18-19 years old in 2007-2012, respectively (and hence have had the opportunity to graduate from high school within 3 years of completing ninth grade in our sample). Standard errors block bootstrapped at the county*voucher*year level in parentheses.

Table 6: Impact of Grade Inflation on High School GPA (LATE)

	Panel A. Causal Impact Estimate	
	Pass	PWD
Δ Final Math Grade	1.36*** (0.49)	1.01* (0.61)
F Stat	308.6	185.4
Dep Variable Mean	11.9	14.0

	Panel B. OLS Estimate
Pass	2.067*** (0.0455)
PWD	4.169*** (0.0575)
Observations	141426

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Note: Panel A presents estimates of the impact of receiving a higher final math grade due to teachers' discretionary grading on high school GPA at graduation. The sample includes all students who attend ninth grade in 2004 through 2006. Panel B displays the OLS estimate. Standard errors block bootstrapped at the county*voucher*year level in parentheses.

Table 7: Impact of Grade Inflation on Initiating College (LATE)

	Panel A. Causal Impact Estimate	
	Pass	PWD
Δ Final Math Grade	0.12**	0.079
	(0.052)	(0.12)
F Stat	67.3	57.9
Dep Varaible Mean	0.14	0.38

	Panel B. OLS Estimate
Pass	0.160***
	(0.00278)
PWD	0.418***
	(0.00478)
Observations	134448

* p < 0.10, ** p < 0.05, *** p < 0.01.

Note: The table presents estimates of the impact of receiving a higher final math grade due to teachers' discretionary grading on the likelihood of enrolling in college within 7 years of completing ninth grade. The sample includes all students who attend ninth grade between 2004 and 2005, who are 22-23 years old in 2011 and 2012, respectively (and hence we observe whether they initiate college within 7 years of completing ninth grade). Panel B displays the OLS estimate. Standard errors block bootstrapped at the county*voucher*year level in parentheses.

Table 8: Impact of Grade Inflation on Years of Education (LATE)

	Panel A. Causal Impact Estimate	
	Pass	PWD
Δ Final Math Grade	0.33*	0.48*
	(0.20)	(0.30)
F Stat	67.3	57.9
Dep Variable Mean	11.8	12.5

	Panel B. OLS Estimate
Pass	1.261***
	(0.0199)
PWD	2.038***
	(0.0222)
Observations	131756

* p < 0.10, ** p < 0.05, *** p < 0.01.

Note: The table presents estimates of the impact of receiving a higher final math grade due to teachers' discretionary grading on educational attainment within 7 years of ninth grade. The sample includes all students who attend ninth grade in 2004 and 2005. Panel B displays the OLS estimate. Standard errors block bootstrapped at the county*voucher*year level in parentheses.

Table 9: Impact of Grade Inflation on Pr of Teenage Birth (LATE)

	Panel A. Causal Impact Estimate	
	Pass	PWD
Δ Final Math Grade	-0.027 (0.019)	-0.035** (0.017)
F Stat	67.3	57.9
Dep Varaible Mean	0.014	0.0048

	Panel B. OLS Estimate
Pass	-0.0226*** (0.00201)
PWD	-0.0303*** (0.00218)
Observations	134428

* p < 0.10, ** p < 0.05, *** p < 0.01.

Note: The table presents estimates of the impact of receiving a higher final math grade due to teachers' discretionary grading on the probability of having a child before age 20. The sample includes all students who attend ninth grade in 2004-2005. Panel B displays the OLS estimate. Standard errors block bootstrapped at the county*voucher*year level in parentheses.

Table 10: Impact of Grade Inflation on Income (LATE)

	Panel A. Causal Impact Estimate	
	Pass	PWD
Δ Final Math Grade	340.4* (183.0)	448.5** (215.0)
F Stat	67.3	57.9
Dep Variable Mean	1579.9	1461.2

	Panel B. OLS Estimate
Pass	369.8*** (18.46)
PWD	176.8*** (22.85)
Observations	131756

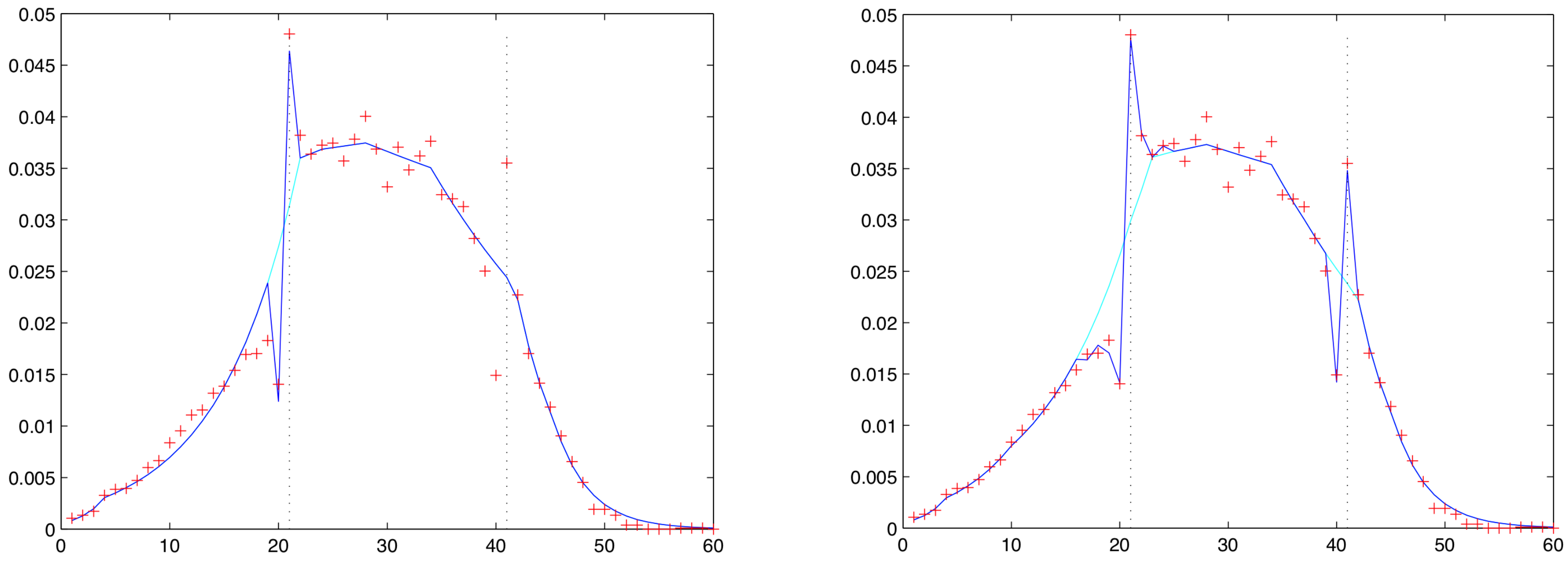
* p < 0.10, ** p < 0.05, *** p < 0.01.

Note: The table presents estimates of the impact of receiving a higher final math grade due to teachers' discretionary grading on age-23 earnings. The sample includes all students who attend ninth grade in 2004 and 2005, who are 22-23 years old in 2011 and and 2012, respectively. Panel B displays the OLS estimate. Standard errors block bootstrapped at the county*voucher*year level in parentheses.

ONLINE APPENDIX – NOT FOR PUBLICATION

A Supplemental Figures

Figure A1: Examples of Estimates of Unmanipulated Distributions for Different Guesses of β_1 and β_2

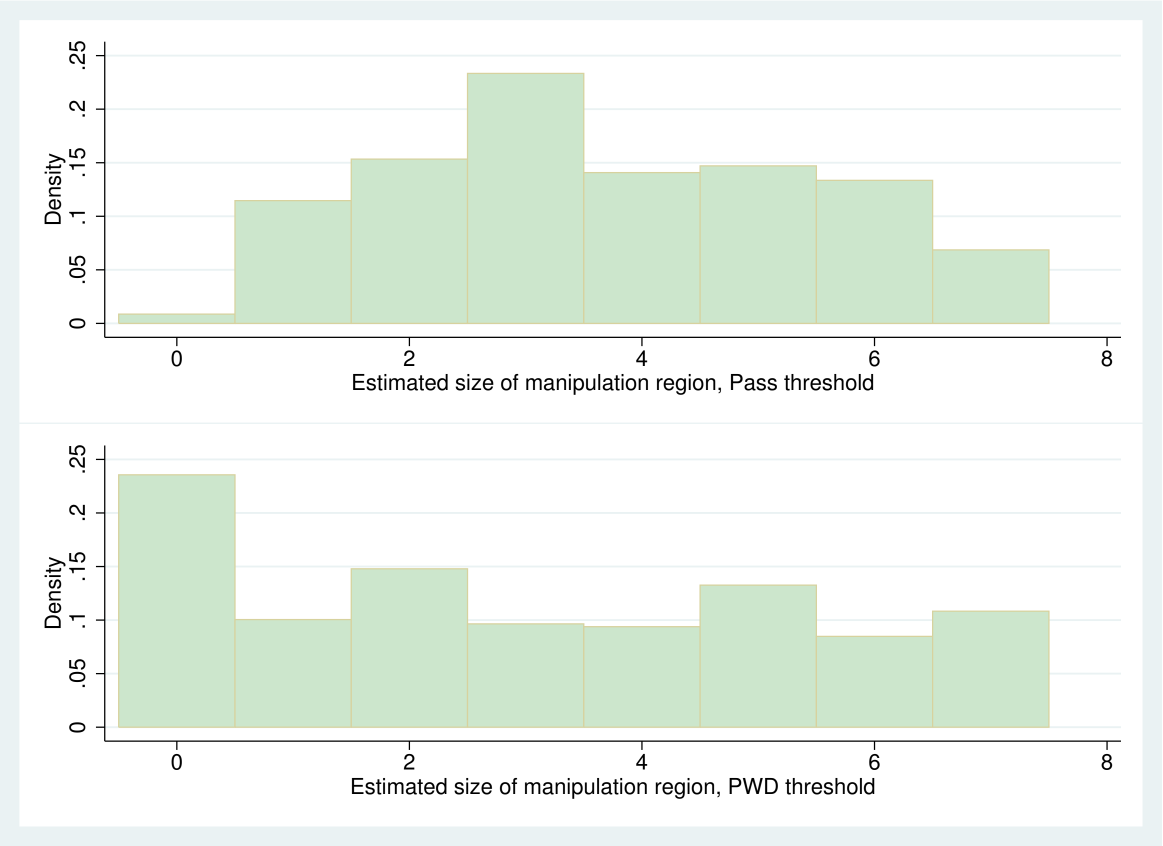


(a) $\hat{\beta}_1 = 1, \hat{\beta}_2 = 0$

(b) $\hat{\beta}_1 = 4, \hat{\beta}_2 = 1$

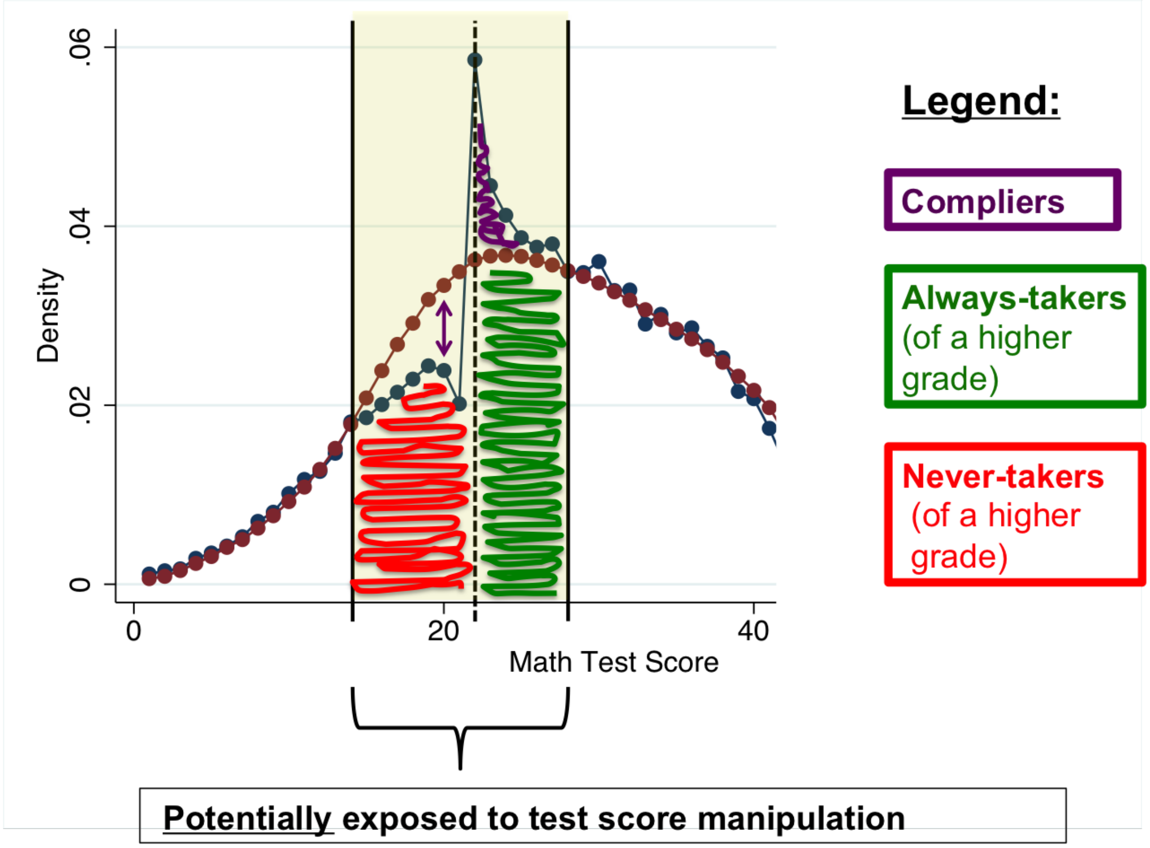
Note: In each subfigure, the red plus signs display the raw data; the blue solid line displays the estimated distribution including manipulation; and the turquoise solid line displays the estimated counterfactual (un-manipulated) distribution (which only deviates from the blue solid line where manipulation occurs). While the raw data is the same in both subfigures, the estimated distribution including manipulation, as well as the estimated counterfactual distribution, differ in the two subfigures. In Figure A1a, we display our estimate of the manipulated and un-manipulated distribution under the hypotheses that $\beta_1 = 1$ and $\beta_2 = 0$. In Figure A1b , we display our estimate of the manipulated and un-manipulated distribution under the hypotheses that $\beta_1 = 4$ and $\beta_2 = 1$. Note how $\beta_1 = 4$ and $\beta_2 = 1$ fit the data much better. These data are for municipal schools in Stockholm county in 2005.

Figure A2: Distribution of Grading Leniency around the Two Thresholds



Note: The figure illustrates the distribution of the estimated sizes of the manipulation region, around the thresholds for Pass and PwD, respectively, by county*voucher*year, from 2004 to 2010.

Figure A3: Wald estimator



Note: The figure illustrates the test score regions of relevance to our Wald estimator in an example where the Pass threshold is 21 and the manipulation region starts at 14. Students who receive a raw (un-manipulated) test score of 13 face a zero probability of being graded up, whereas students who receive a raw test score of 14 (or higher) face a weakly positive probability of being graded up. Among the students whose raw test scores fall into the interval $14-20$, teachers choose to grade up a subset; these can be thought of as the compliers, who are "missing" below 21 in the observed test score distribution. The students whose observed test scores lie in the interval $14-20$ can be thought of as never-takers, as they are left un-manipulated even though their raw test scores put them into the manipulation region. Finally, the students whose raw *and* observed test scores lie at or above 21 can be thought of as always takers. In the data, we can observe the never-takers; however, we cannot distinguish the compliers from the always-takers, as both groups' observed test scores fall at or above 21 and we do not observe the raw test scores.

B Supplemental Tables

Table B1: Impact of Inflated Test Grade on Final Math Grade

	Panel A. Causal Impact Estimate	
	Pass	PWD
Δ Math Test Grade	0.35*** (0.020)	0.87*** (0.064)
F Stat	1683.7	133.1
Dep Varaible Mean	0.99	1.62

	Panel B. OLS Estimate
Pass	0.412*** (0.00465)
PWD	1.235*** (0.00867)
Observations	478675

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Note: The table presents estimates of the impact of receiving an inflated math test grade on the final math grade (on everyone in the manipulation region; though this estimate is in practice driven by the impact on those who are graded up, i.e., the compliers). The sample includes all cohorts in our sample, i.e., all students who attend ninth grade between 2004 and 2010. The predicted final grade absent manipulation is estimated from regressions of students' final grades on a dummy for whether the test score is above the cutoff and 3rd order polynomials in the test score, for each year and county*voucher. These regressions only use data from students outside of the manipulation regions of the test score distribution. See the text for more details. Standard errors block bootstrapped at the county*voucher*year level in parentheses.

Table B2: “Sanity Check”: Impact of Grade Inflation on English Test Grade (LATE)

	Panel A. Causal Impact Estimate	
	Pass	PWD
Δ Final Math Grade	-0.019 (0.062)	-0.021 (0.067)
F Stat	317.3	141.6
Dep Varaible Mean	1.32	1.88

	Panel B. OLS Estimate
Pass	0.626*** (0.00621)
PWD	1.028*** (0.00751)
Observations	399221

* p < 0.10, ** p < 0.05, *** p < 0.01.

Note: The table presents estimates of the impact of receiving a higher final math grade due to teachers' discretionary grading on the student's English test grade. The English test is taken before the math test, so it cannot be affected by the outcome on the math test. Thus, this is a sanity check of our identification strategy. Panel B displays the OLS estimate. The sample includes all cohorts in our sample, i.e., all students who attend ninth grade between 2004 and 2010. Standard errors block bootstrapped at the county*voucher*year level in parentheses.

Table B3: “Sanity Check”: Impact of Grade Inflation on Swedish Test Grade (LATE)

	Panel A. Causal Impact Estimate	
	Pass	PWD
Δ Final Math Grade	0.036 (0.055)	0.072 (0.071)
F Stat	317.3	141.6
Dep Varaible Mean	1.18	1.67

	Panel B. OLS Estimate
Pass	0.556*** (0.00591)
PWD	0.986*** (0.00636)
Observations	399711

* p < 0.10, ** p < 0.05, *** p < 0.01.

Note: The table presents estimates of the impact of receiving a higher final math grade due to teachers' discretionary grading on the student's Swedish test grade. The Swedish test is taken before the math test, so it cannot be affected by the outcome on the math test. Thus, this is a sanity check of our identification strategy. Panel B displays the OLS estimate. The sample includes all cohorts in our sample, i.e., all students who attend ninth grade between 2004 and 2010. Standard errors block bootstrapped at the county*voucher*year level in parentheses.

Table B4: Impact of Grade Inflation on High School *Peers'* GPA (LATE)

	Panel A. Causal Impact Estimate	
	Pass	PWD
Δ Final Math Grade	-0.14 (0.14)	0.012 (0.18)
F Stat	123.1	51.6
Dep Varaible Mean	12.7	13.0

	Panel B. OLS Estimate
Pass	0.260*** (0.0283)
PWD	0.554*** (0.0493)
Observations	141426

* p < 0.10, ** p < 0.05, *** p < 0.01.

Note: The table presents estimates of the impact of receiving a higher final math grade due to teachers' discretionary grading on peer GPA in high school. The sample includes all students who attend ninth grade in 2004 through 2006. Panel B displays the OLS estimate. Standard errors block bootstrapped at the county*voucher*year level in parentheses.

C More information on the nationwide math test

The test is comprised of four subtests, or parts: A, B1, B2, and C. Each subtest has a certain number of questions, and each question is worth a certain number of "Pass points," P and a certain number of "Pass with Distinction points," PwD. (Easier questions are awarded only P-points; harder questions are awarded both P- and PwD-points, or only PwD-points.)

A grading sheet is distributed to teachers with detailed instructions regarding the grading of each question. The P-points are awarded based on objective and hard-to-manipulate criteria (such as "which of the following five numbers are higher?"). The the PwD-points often involve a subjective assessment, however: points are awarded for partially completed work, for "clarity," for "beautiful expression," and so on. The grading sheet thus effectively provides a short list of correct answers to the P-points, and longer descriptions of how to award PwD-points.

A student's test score thus consists of a pair, (P_i, PwD_i). In 2004, the maximum number of P-points was 38, and the maximum number of PwD-points was 32. The highest sum of points that a student could achieve was thus $S_i = P_i + PwD_i = 70$.

In addition to providing guidance on the grading of each question, the grading sheet defines the test grade as a step function of the number of P- and PwD-points; or, equivalently, as a function of S_i and PwD_i:

$$t_i = \left\{ \begin{array}{ccccc} Pass & \text{if} & S_i \geq 23 & & \\ PwD & \text{if} & S_i \geq 43 & \text{and} & PwD_i \geq 12 \\ Excellent & \text{if} & PwD_i \geq 21 & \text{and} & E = 1 \end{array} \right\},$$

$S_i \geq 23$ is a necessary and sufficient condition for obtaining the test grade Pass. Moreover, for the vast majority of students who are on the margin between Pass and PwD, $S_i \geq 43$ is the binding constraint (as opposed to $PwD_i \geq 12$); thus, again, the necessary and sufficient condition for obtaining PwD can be expressed in terms of the sum of Pass and PwD points. The students where the PwD points is the binding constraint for receiving a PwD test grade are dropped from the analysis. In the paper, we therefore define the raw test score r_i as the sum of Pass and PwD points (S_i above).

A subset of the test questions, marked by the symbol #, allow the teacher to judge criteria that capture that the student's answers are worthy of the grade "Excellent" ($E = 1$). We do not observe the teachers' judgements of these criteria – they are awarded based on highly subjective criteria – but we can infer it based on the awarded test grade.[49] In contrast to

[49] These criteria include (i) using general strategies when planning and executing the exercise; (ii) comparing and evaluating the pros and cons of different solution methods; (iii) displaying certainty in the calculations; (iv) displaying structured mathematical language; and (v) displaying an ability to interpret

Pass and PwD, however, the test score that we observe in the data does not provide anything resembling a sufficient condition for receiving the test grade Excellent – a substantial share of the students whose test score satisfies $PwD_i \geq 21$ are *not* awarded the grade Excellent in the data. Because the test score only provides a necessary but not sufficient condition for the grade Excellent, our method is not appropriate, so we do not analyze this threshold.

D Estimation Details

D.1 Recovering the un-manipulated distribution and the width of the manipulation region

Let R_{rjt} equal the observed test score frequency in the data at test score r within region j in year t. Our model predicts:

$$\underbrace{R_{rjt}}_{\text{Observed density}} \underbrace{= B^{N_{jt}}(\theta_{jt}, r) -}_{\text{un-manipulated distribution}} \underbrace{\textstyle\sum_k m_{jt}^{low,k,p_{kjt}^{low}}(\theta_{jt}^{low,k}, r) +}_{\text{missing mass}} \underbrace{\textstyle\sum_k m_{jt}^{high,k,p_{kjt}^{high}}(\theta_{jt}^{high,k}, r) +}_{\text{excess mass}} \underbrace{\epsilon_{rjt}.}_{\text{sampling error}}$$

We parameterize each function above $\{B^{N_{jt}}(\theta_{jt}, r), m_{jt}^{low,k,p_{kjt}^{low}}(\theta_{jt}^{low,k}, r), m_{jt}^{high,k,p_{kjt}^{high}}(\theta_{jt}^{high,k}, r)\}$ as an exponentiated Bernstein polynomial:

$$B^{N_{jt}}(\theta_{jt}, r) = exp(\sum_{n=0}^{N_{jt}} \theta_{njt} f_n(r, N_{jt})),$$

$$m_{jt}^{low,k,p_{kjt}^{low}}(\theta_{jt}^{low,k}, r) = \sum_{n=0}^{p_{kjt}^{low}} \theta_{njt}^{low,k} f_n(r, p_{kjt}^{low}),$$

$$m_{jt}^{high,k,p_{kjt}^{high}}(\theta_{jt}^{high,k}, r) = \sum_{n=0}^{p_{kjt}^{high}} \theta_{njt}^{high,k} f_n(r, p_{kjt}^{high}),$$

where $f_n(x, N))$ is the nth basis function of the Nth order Bernstein polynomial:

$$f_n(x, N)) = \binom{N}{n} x^n (1-x)^{(N-n)}.$$

As shown in Wang and Ghosh (2012), we can constrain all functions to be log-concave by imposing the following linear inequality constraint on the polynomial coefficients. Let θ

and analyze. In order to be awarded the grade Excellent on the test, a student must have demonstrated "most of" these five qualities, on at least three of the six questions marked by the symbol #.

be the coefficients on the Bernstein polynomial of interest. Then:

- $exp(\sum_{n=0}^{N} \theta_n f_n(r, N))$ is log-concave if:

$$\begin{bmatrix} 1 & -2 & 1 & 0 & \dots & 0 \\ 0 & 1 & -2 & 1 & \dots & 0 \\ & & \ddots & & & \\ 0 & \dots & 0 & 1 & -2 & 1 \end{bmatrix} \begin{bmatrix} \theta_0 \\ \theta_1 \\ \vdots \\ \theta_N \end{bmatrix} \leq \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}.$$

- $\sum_{n=0}^{N} \theta_n f_n(r, N)$ is monotonically non-decreasing if:

$$\begin{bmatrix} -1 & 1 & 0 & \dots & 0 \\ 0 & -1 & 1 & \dots & 0 \\ & & \ddots & & \\ 0 & \dots & 0 & -1 & 1 \end{bmatrix} \begin{bmatrix} \theta_0 \\ \theta_1 \\ \vdots \\ \theta_N \end{bmatrix} \geq \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}.$$

Similarly, reversing the inequality sign would impose a monotonically non-increasing function.

We use Bernstein polynomials instead of standard Nth order polynomials because the shape of a standard polynomial cannot be constrained. Even if one were to check whether a polynomial was monotonic or log-concave at a set of arbitrary points, there is no guarantee that the polynomial would follow these shape restrictions at points not directly validated. The benefit of the Bernstein polynomial basis is that simple linear inequalities on the polynomial coefficients guarantee global monotonicity and log-concavity. See Wang and Ghosh (2012) for proofs and simulations showing the consistency and asymptotic properties of this style of shape-restricted sieve estimator.

We impose that all functions governing the manipulation and the un-manipulated distribution are log-concave. We further restrict the missing mass function to be montonically non-decreasing and the excess mass function to be montonically non-increasing. We estimate the model using constrained nonlinear-least squares and use k-fold (k=5) cross-validation select the orders of all the polynomials. We perform the following procedure for each region-year:[50]

For a given guess of the order of the polynomials, we perform a grid search over the possible width of the two manipulation regions. For each guess of the width of the manipulation region we use constrained non-linear least squares to estimate the un-manipulated distribution, and the deviation from it due to manipulation, to fit the observed test score distribution. We then select the width of the manipulation region and estimated polynomial

[50] Recall that we refer to one county*voucher as a "region."

coefficients that minimize the MSE. This is done on the 80% training sample, and then used to predict out-of-sample on the 20% hold-out sample. We then calculate the out-of-sample mean squared error (MSE) for the 20% hold-out sample, and the repeat this procedure on each of the five folds of data. We sum these out-of-sample MSEs as our measure of model fit for that particular guess of the orders of the polynomials. We then iterate this procedure for each guess of the orders of the polynomials.

We select the polynomial orders that have the smallest out-of-sample MSE. Note that when randomly binning the data into 5 groups for the cross-validation procedure, we sample data points from the histogram (e.g. treating a 51 point test score distribution as 51 data points), instead of binning the data by randomly sampling students. This allows there to be error in the model at the test score level, due to model misspecification or other quirks of the test that could lead to deviations from log-concavity randomly at each test score point for reasons other than manipulation.

Once we have selected the out-of-sample MSE-minimizing orders of the polynomials, we pool all the data back together and estimate the widths of the manipulation regions and the polynomial coefficients.

D.2 Estimating the causal impact of test score manipulation

First, we identify the impact of test score manipulation on students' age-23 earnings. This can be thought of as the reduced form regression of our Wald estimation. To do this, we proceed in two steps.

Recall that g_{ijt} is student $i's$ age-23 earnings (who is enrolled in region j in year t). We estimate:

$$g_{ijt} = \hat{g}_{kjt}\left(r_{ijt}, \theta_{kjt}^{grade}\right) + \alpha_{kjt} * (r_{ijt} \geq k) + \epsilon_{ijt}^{g}, \tag{11}$$

$$\text{where: } (r_{ijt} < k - \beta_{kjt} \text{ or } r_{ijt} > k + \beta_{kjt} - 1)$$

$$\text{and}$$

$$r_{ijt} > (k-1) + \beta_{k-1jt} + 1 \text{ and } r_{ijt} < (k+1) - \beta_{k+1jt}.$$

$\hat{g}_{kjt}\left(r_{ijt}, \theta_{kjt}^{grade}\right)$ is a third order polynomial with coefficients θ_{kjt}^{grade} that captures the smooth relationship between students' un-manipulated test scores, r_{ijt}, and their expected earnings. $(r_{ijt} < k - \beta_{kjt}$ or $r_{ijt} > k + \beta_{kjt} - 1)$ ensures that the data used to estimate equation (2) is outside of the test score inflated region around test grade threshold k. $r_{ijt} > (k-1) + \beta_{k-1jt} + 1$ and $r_{ijt} < (k+1) - \beta_{k+1jt}$ ensures that the data is also not within the test score inflated regions around the higher $(k+1)$ or lower $(k-1)$ test grade thresholds.

We allow there to be a discrete jump in students' expected earnings at the test grade cut-off k, represented by $\alpha_{kjt} * (r_{ijt} \geq k)$.

For a few region-years for which there are few students, this extrapolation inwards using the polynomial causes predictions outside of the range of the outcome variables. To limit the impact of these outliers on our overall estimates, we trim the predicted outcomes inside the manipulation region to never be above the polynomial predicted values just outside either side of the manipulation window. This preserves monotonicity of the relationship between the outcome variable and un-manipulated test scores.

In our estimation of long-term effects, we exclude regions where the manipulation region is estimated to be of width 7, as this is the highest width that we searched over in our grid search described above. Thus, in these regions, 7 could be an under estimate of the true width of the manipulated region. Further, the ability to extrapolate inward with a polynomial becomes more challenging as the width of the manipulation region widens. (Hence the decision not to search over regions wider than 7.)

We use our estimates to calculate the expected average earnings for students within the manipulation region of the test score distribution *had there been no test score manipulation*:

$$\bar{g}_{jt}(k) = \int_{k-\beta_{kjt}}^{k+\beta_{kjt}-1} \left[\hat{g}_{kjt}\left(r, \hat{\theta}_{kjt}^{grade}\right) + \hat{\alpha}_{kjt} * (r \geq k)\right] * \frac{\hat{h}_{jt}(r)}{\int_{k-\beta_{kjt}}^{k+\beta_{kjt}-1} \hat{h}_{jt}(r)dr} dr. \tag{12}$$

For students inside the manipulation region, we now compare the estimated counterfactual average earnings, had there been no test score manipulation, calculated in (12), with the actual average earnings for students in the manipulation region (observed in the data), g_{ijt}. Thus, this difference is our "intent-to-treat" estimate of the average increase in a student's earnings due to the student having a raw test score that falls within the manipulated region of the test score distribution:

$$\begin{aligned} ITT = \quad & E\left(\text{earnings}|\text{teacher can manipulate}\right) - E\left(\text{earnings}|\text{teacher can't manipulate}\right) \\ = \quad & \underbrace{\frac{\sum_{jt}\left(\sum_{i\epsilon\text{manip region k}} g_{ijt}\right)}{\sum_{jt}\left(N_{kjt}^{\text{manip}}\right)}}_{\text{Average observed earnings across all students in manipulation region across all j regions and t years}} - \underbrace{\frac{\sum_{jt} N_{kjt}^{\text{manip}} \bar{g}_{jt}(k)}{\sum_{jt}\left(N_{kjt}^{\text{manip}}\right)}}_{\text{Average predicted earnings for students in manipulation region, had there been no manipulation across all j regions and t years}}, \end{aligned}$$

where N_{jt}^{manip} is the number of students in the manipulation region around threshold k in region j in year t.

The “reduced form” and LATE estimates. The procedure above can be repeated with math test grade instead of earnings. This yields the first stage effect of falling into the manipulation region on the test grade. The ratio of the reduced-form effect on earnings to the first-stage effect, in turn, identifies the local average treatment effect (LATE) of receiving an inflated exam math grade on future income. We block bootstrap the entire procedure to calculate standard errors, sampling at the county by voucher by year level. This is the same level at which we estimated the widths of the manipulation regions.