NBER WORKING PAPER SERIES

THE LONG-TERM CONSEQUENCES OF TEACHER DISCRETION IN GRADING OF HIGH-STAKES TESTS

Rebecca Diamond
Petra Persson

Working Paper 22207
http://www.nber.org/papers/w22207

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
April 2016

A previous version of this paper was circulated under the title “The Long-term Consequences of Grade Inflation." We are grateful to Raj Chetty, Matt Gentzkow, Caroline Hoxby, Guido Imbens, Eddie Lazear, Paul Oyer, Luigi Pistaferri, Kathryn Shaw, Alan Sorensen, Chris Taber, and seminar and conference participants at Yale, NYU, Stanford, Santa Clara University, UC Berkeley, UC San Diego, UC Santa Cruz, University of Houston, Wisconsin, the Minneapolis Federal Reserve, SIEPR, Gothenburg University, Stockholm University, the Swedish Institute for Social Research, the University of Oslo, Uppsala University, the UCLA All California Labor Conference 2015, the Utah Winter Business Economics Conference 2016, and the 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. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

The Long-term Consequences of Teacher Discretion in Grading of High-stakes Tests
Rebecca Diamond and Petra Persson
NBER Working Paper No. 22207
April 2016, Revised June 2016
JEL No. I20,J24

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.

Rebecca Diamond
Graduate School of Business
Stanford University
655 Knight Way
Stanford, CA 94305
and NBER
diamondr@stanford.edu

Petra Persson
Department of Economics
Stanford University
579 Serra Mall
Stanford, CA 94305
and NBER
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 placed on standardized testing should be repealed.

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 estimate the width of the manipulation regions (i.e., the lowest test score at which *any* student gets bumped up) at the local level and find that it varies substantially: in some places, students' test scores are not manipulated at all; in others, students who lack as much

as eight test score points are bumped beyond the threshold. 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 to analyze *the distribution that is being manipulated*: distributions of reported incomes (Saez, 2010; Chetty et al., 2011b; 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 have been smooth.

The core idea in this paper is to further develop this literature 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. Intuitively, just like we can plot the test score density, we can plot the mean of a future outcome (say, 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 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. In the manipulation region around the Pass threshold, the difference between the average observed and counterfactual earnings captures the

reduced form impact of (potentially being exposed to) test score manipulation on earnings.[1] The interpretation around the PwD region is analogous.

Not all students whose test scores fall within the manipulation regions of the test score distribution actually get bumped up, however. We therefore also estimate a "first stage" effect of having a test score that falls in the manipulation region on the probability of receiving test score manipulation. The ratio of the reduced-form effect to the first stage effect identifies the local average treatment effect of receiving a higher grade *due to test score manipulation* on labor market earnings.[2]

More generally, this methodology could be used 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). We note that 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 where manipulation of the running variable precludes using RD methods for analyzing causal impacts (**?**). Our methodology essentially "fills this gap," by allowing for identification of causal effects of manipulation per se.

Our results suggest far-reaching consequences of receiving test score manipulation, at all subsequent stages of the student's life that we observe. Although inflating a student's test grade does not increase her knowledge, we find that it raises the student's performance in the immediate future. In particular, students who are inflated on the test, taken in February, perform better *in other subjects* during subsequent months, which raises their final grades (awarded in June), and thereby their GPA. These effects, which are particularly pronounced at the higher end of the ability distribution, 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 end of high school, three years after test score manipulation. Being graded up above the lower (higher) threshold raises the likelihood of high school graduation three years later by 20 (6) 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

[1] We combine the estimated relationship between earnings and un-manipulated test scores with the estimated distribution of un-manipulated test scores in the manipulation region of the test score distribution to calculate counterfactual average earnings.

[2] 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."

condition for admittance to *any* high school.[3] Moreover, inflated students perform better in high school: Those inflated above the lower (higher) threshold have 11% (7%) higher high school GPA – even though manipulation does not, per se, push these students into high schools with higher peer GPAs.

Inflated students continue to benefit eight years after test score manipulation. Students who are inflated above 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, inflated 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, inflated students earn 340-440 SEK more annual income.

In sum, despite that test score manipulation does not, per se, raise human capital, it has far-reaching consequences for the beneficiaries, raising their grades in future classes, high school graduation rates, and college initiation rates; lowering teen birth rates; and raising earnings at age 23.

The large immediate impacts are consistent with the fact that, in the context that we analyze, the "signaling value" of being inflated is large: First, students do not observe their numeric score on the test – only the letter grade – which means that bumping a student up creates a large signal change. Second, 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. 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; and the combination of higher effort and higher human capital ultimately "snowballs," generating large labor market gains.

This insight is related to the sheepskin literature, which analyzes the signaling value of education *in the labor market* (Cameron and Heckman, 1993; Kane and Rouse, 1995; Kane et al., 1999; Tyler et al., 2000; Clark and Martorell, 2014). This literature generally has found a small or zero signaling value of education in the labor market. Our results suggest that the signaling value of grades may be more important inside the educational system itself, by raising students' motivation or other teachers' perceptions.

The importance of signaling inside the educational system ties to the literature on the impact of receiving a failing grade (Jacob and Lefgren, 2006; Manacorda, 2012). This literature focuses on educational attainment within a few years of receiving a failing grade, and do not follow the students into the labor market. Our paper essentially marries this literature

[3]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.

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.

More generally, we contribute to the literature that documents long-term impacts of a range of school policies, including school desegregation (Billings et al., 2014); school choice (Ahlin, 2003; Sandström and Bergström, 2005; Björklund et al., 2006; Lavy, 2010; Deming et al., 2014; Edmark et al., 2014; Lavy, 2015); preschool programs (Garces et al., 2002); class size (Krueger and Whitmore, 2001; Chetty et al., 2011a; Fredriksson et al., 2012); and teacher value added (Chetty et al., 2011a, 2014). We assess the long-term consequences of test score manipulation, as well as to document which students stand to gain.

A few papers assess the potential *causes* of test score manipulation.[4] Previous work 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))[5], 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.[6] 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. They find that students whose answers were changed performed worse on future achievement tests and were more likely to drop out of school than students whose answers were left unchanged; however, selection of students into treatment (manipulation) could confound a causal interpretation. Indeed, in our context, teachers choose to manipulate students who perform idiosyncratically poorly on the test.

[4]While we focus on *ex post* test score manipulation, a related literature analyzes *ex ante* manipulation, including efforts to "teach to the test" (Jacob, 2005) and to target instruction to students who are believed to be at risk for falling close to target thresholds (Neal and Whitmore Schanzenbach, 2010).

[5]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.

[6]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 most closely related paper is concurrent work by Dee et al. (2016). They they build on Dee et al. (2011), which documented manipulation of test scores on the New York City's Regent's Exam and analyzed the causes of manipulation. The new paper studies effects of manipulation on high school graduation and college enrollment and find effects consistent with our study. Our paper differs from Dee et al. (2016), and from the rest of this literature, by studying key long term outcomes including labor market earnings, child bearing, and achievement outcomes in schooling over 7 years after test taking.

Our identification strategy is also very different from Dee et al. (2016). They rely on a combination of policy variation, which removed teachers abilities to manipulate test scores, and cross-sectional variation between schools that engaged in high versus low levels of manipulation. The policy variation exhibits strong pre-trends in treatment, making difference-in-difference estimation challenging. The cross-sectional specifications indicate treatment effects of manipulation on student race, suggesting unobservable sorting of students to schools, which could confound a causal interpretation. Our paper provides a new identification strategy building on the bunching literature methods.

Finally, there is a key distinction between our paper and the literature that analyzes the impact of passing a high-stakes test *in the absence of teacher manipulation*. Lavy et al. (2014) analyze the impact on earnings of performance on a high-stakes test in Israel, using pollution as an instrument. 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).[7] Our analysis is distinct in two key ways. First, receiving a high math test grade in our context does not immediately change access to future schooling. This allows us to separate the signaling value of a higher test grade to the student within the same school year from the mechanical effect of receiving a high score on gaining admittance to selective schools. Second, we analyze the impact of crossing an important proficiency threshold *when teachers have discretion in moving students' across the cut-off*. Depending on how teachers use this discretion, the impact of passing when selected by a teacher with discretion may be radically different from the impact of passing by chance – this, in fact, is one of the key motivations for potentially allowing for teacher discretion.

The remainder of the paper proceeds as follows. Sections 2 and 3 describe the institutional setting and data, respectively. Section 4 provides a model of teachers' grading behavior. In Section 5, we estimate where in the test score distribution manipulation takes place. Section 6 then uses data on long-term outcomes, coupled with our estimates from Section 5 of where

[7]Similarly, Papay et al. (2015) use an RD design to study the benefits of receiving a higher grade on a standardized high school math test. They find receiving the higher grade raises college enrollment for low income students in Massachusetts.

in the test score distribution manipulation takes place, to quantify the causal impacts of test score manipulation on subsequent schooling and adult labor market outcomes. Our results are presented in Section **??**.

2 Institutional background

2.1 Schooling in Sweden

In Sweden, schooling is mandatory until age 16, which corresponds to the first nine years of school (ages 7-16). During the time period that we analyze, grades are awarded for the first time in the Spring semester of eighth grade. One year later, at the end of ninth grade, the *final grades* are awarded. All subjects are graded. 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).

There is a range of high school programs, ranging from vocational (construction, hairdresser, etc.) to programs that serve to prepare individuals for further studies at the university level.[8] To be eligible for high school, the student must have a passing grade in math, English, and Swedish. Conditional on eligibility, the grade point average (GPA) when exiting ninth grade is the sole merit-based criterion used for acceptance to high school.[9] The GPA cutoff for admittance to a given program, in a given year, is determined by the lowest GPA among admitted individuals (a subset of those who applied). 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. The test consists of three sub-tests, which are administered approximately one week apart. The test dates are usually in the beginning of the Spring semester (February), approximately four months before the teacher sets the final grade in mathematics. Students are only informed of their letter grade on the test, but not their raw numeric score. They would be unaware how close they were to grade cutoffs.

[8] See Golsteyn and Stenberg (2015) for a detailed description of vocational and other programs and for comparison of earnings over the life cycle for students choosing vocational versus other programs.

[9] In particular, no other merit-based criterion (such as 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. The GPA reflects the average grade in the 16 subjects that are taught in grades 7-9. The maximum ninth grade GPA is 320., and the minimum GPA is zero.

The test is graded locally, either by the teacher or jointly by the teachers at a given school. The test is graded according to a grading manual provided by The Swedish National Agency for Education (Skolverket), which provides detailed instructions on how each question should be graded. While some points are awarded based on objective and non-manipulable criteria (e.g. providing a correct answer to a multiple choice question such as "which number is larger?"), others involve a subjective assessment: a subset of the points may be awarded for partially completed work, for "clarity," for "beautiful expression," and so on. This gives the teacher leeway in raising a student's test score somewhat, by grading up some questions.[10]

The grading sheet also provides a step function $t(r_i)$, which specifies exact cutoffs in the raw test score, r_i, for awarding the student the grades Fail, Pass, and PwD on the test. In addition, the grading sheet 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 B provides the exact step function from the grading sheet from 2004, along with more detailed information about r_i.

When writing the test, a student knows how many points each question is worth; however, the student does not know the step function $t(r_i)$. Thus, the student cannot write the test targeting the grade cutoffs. Further, these cutoffs vary over time; thus, there is no exact relationship between the cutoffs in one year and the corresponding cutoffs for earlier years. Any bunching that we observe in the test score distribution is thus attributable to teachers' grading leniency, and not to student sorting. In addition to the test in math, nationwide tests are administered in English and Swedish. The test grades obtained on these two language 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. We therefore exploit only the math test when recovering regions' respective grading leniency – this test is ideal for the purpose, as we observe the numeric score, and thereby can detect bunching in the test score distribution.

Final grades The test grade is not binding for the final grade, which is the one that counts towards the GPA. The final grade 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.

[10] In the context of Chicago, Jacob and Levitt (2003) document outright cheating among teachers. The manipulation that we document is distinct in the sense that, in the data, manipulation occurs on test points that are awarded based on subjective criteria, rather than on test points that are awarded based on objective criteria.

This suggests that teachers can engage in two different types of manipulation: first, as discussed above, the nationwide test can be graded leniently, so as to push the student above a test score threshold. Second, teachers can simply decide to set a final grade that is higher than the grade that the student deserves based on the student's un-manipulated test score and "everything else." This effectively corresponds to granting a grade that is higher than the deserved grade, which essentially can be thought of as inflating the student's true, underlying ability.

In practice, the final grade in math does deviate from the (potentially manipulated) test grade, in both directions. Moreover, these deviations do not occur in a uniform fashion; teachers are more likely to award a math grade that is higher than the (potentially manipulated) test grade than they are to award a lower final grade (Vlachos, 2010). This suggests that the nationwide test grade may be used as a "lower bound" on the final grade.

We focus on the first type of manipulation – of the nationwide test scores – and Section 4 formulates a simple but general theoretical framework that operationalizes teachers' incentives to engage in such test score manipulation. In Appendix C, we present a richer model that incorporates the second type of manipulation as well – of the final grade – and where teachers are allowed to trade off the two types of inflation. Ultimately, both models pinpoint the same key parameter of interest for our empirical analysis of long-term effects of test score manipulation; thus, restricting our attention to test score manipulation is innocuous.

2.2 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 inflation: 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.[11] 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. 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

[11] Nonvoucher tuition payments are forbidden in Sweden.

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.

Taken together, these two features of Sweden's schooling system provide an institutional context where 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 inflation, which artificially raises the school's reputation for quality. This gives school principals strong incentives to encourage their teachers to go easy on grading.[12]

3 Data

3.1 Swedish administrative data

We use administrative population-level data from Sweden. We start from the universe of students who attend ninth grade between 2004 to 2010.[13] For these children and their families, we obtain information from various data sources, which we link through unique individual identifiers. Taken together, these data sources provide information about each student's academic performance, subsequent medium- and longer-term outcomes, as well as detailed demographic and socio-economic characteristics.

Grade nine academic performance and schooling information We observe precise information on each student's performance on the nationwide test in mathematics, English, and Swedish. On the math test, we observe both the number of points obtained (after possible manipulation by the teacher), and the test grade. On the English and Swedish tests, we only 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.

In addition to results from the nationwide tests in grade nine, we observe the student's final grade in math. As explained in Section 2 above, this course grade partly reflects the result on the nationwide test, but also performance on homework, etc. In addition, we

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

[13] Note that our data includes both children who are born in Sweden and children who are born outside of Sweden but attended ninth grade in Sweden.

observe the grade point average (GPA) upon exit from grade nine.[14]

We observe the school that the student attends, as well as information about whether the school is a municipal school or a voucher school.

Demographic and socio-economic characteristics For each child in our sample, we have data on the exact date of birth; the municipality of residence when attending ninth grade; and whether the student has a foreign background.[15] We also have variables related to parental socio-economic status: we 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 and 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 the length of each individual's studies (i.e., 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 employments. Finally, we observe an indicator for teen birth by 2009, which corresponds to the age of 21 for the oldest cohort in our sample.

In sum, we create a unique data set that enables us to follow the students from ninth grade, throughout adolescence, and into adult life, all the while tracking key economic outcomes.

3.2 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

[14] See footnote 8 for more information about the GPA.

[15] We define foreign background as having a father who is born outside of Sweden; thus, this incorporates both first and second generation immigrants.

in math.

Table 1 presents summary statistics. The first column presents summary statistics for the full sample, 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.[16]

Our key longer-term outcomes of interest are whether the student graduates from high school, college 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.

4 A model of test score manipulation

To quantify the causal impact of test score manipulation (which we turn to in Section 6 below), we must know where in the test score distribution manipulation takes place. This section models teachers' incentives to manipulate students' test scores and provides a theoretical answer to this question. In a nutshell, the model shows, in a very general setting, that there exists a lowest test score where manipulation takes place (around each grade threshold). This is important for our estimation in Section 6, as it implies that, in the vicinity of each test score threshold, there exist test score ranges where students are left un-manipulated; thus, we can use these students to learn about how students who were manipulated would have fared in the absence of manipulation.

[16] See footnote 8 for more information about the GPA.

4.1 Set-up

For simplicity, we model test score manipulation around a single threshold, which we will refer to as Pass versus Fail. A richer but less general model, which more closely captures the precise institutional features of the grading system in Sweden and which permits a structural interpretation of the key parameter of interest (isolated below), is presented in Appendix C.

Student i is taught by teacher j. He attends class and has performed at level a_i on class assignments, other than the nationwide test. We refer to a_i as student $i's$ ability, and assume that it is observable to the teacher.[17]

Student i takes the nationwide test and, in the absence of any test score manipulation, receives a numeric test score $r_i = r(a_i, \varepsilon_i)$. We refer to this as the "raw" test score, to underscore that it is un-manipulated. The raw test score depends on the student's ability, but also on an error term $\varepsilon_i \tilde{} F(\varepsilon_i)$, which captures the fact that student i's performance on the test may deviate from her true ability; that is, the student can have a "good test day" (if $r_i > a_i$) or a "bad test day" (if $r_i < a_i$), with the magnitude of the deviation reflecting just how good or bad the test performance was relative to the student's innate ability. Because the teacher grades the test, she observes the raw test score r_i.

The teacher may choose to inflate the raw test score by awarding some amount of additional test points to student i, Δ_i, resulting in a manipulated test score of $r_i + \Delta_i$.

The test grade, t_i, is either Pass or Fail, and is given by the following indicator function (for Pass):

$$t_i = t(a_i, \varepsilon_i, \Delta_i) = \left\{ \begin{array}{c} 1 \text{ if } (r(a_i, \varepsilon_i) + \Delta_i \geq \bar{p}) \\ 0 \text{ o/w} \end{array} \right\}.$$

Intuitively, if student $i's$ test score $r(a_i, \varepsilon_i) + \Delta_i$ is higher than the passing threshold $\bar{p}$, then he passes the test; otherwise he fails. The teacher chooses the amount of manipulation of student i's test score, Δ_i, to maximize the per-student utility function:

$$u_{ij}(\Delta_i) = \beta_{ij} t(a_i, \varepsilon_i, \Delta_i) - c_{ij}(\Delta_i),$$

$$c'_{ij}(\Delta_i) > 0, c''_{ij}(\Delta_i) > 0.$$

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

[17] 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.

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 teacher to use her discretion both in a "corrective" and "discriminatory" fashion: 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). 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, $c_{ij}(\Delta_i)$. $c_{ij}(\Delta_i)$ is assumed to be strictly increasing and convex. This captures the fact that it is increasingly hard for a teacher to add an additional test point as she inflates the test score more and more.[18]

4.2 Teachers' grading behavior and the shape of the test score distribution

We now explore properties of the model above that will be useful for estimation. For now, we assume that when the teacher chooses Δ_i, she is free to pick any (positive) value that she wishes.[19] Before analyzing the teacher's decision to use her discretion to manipulate a student's test score, we discuss what happens if $\beta_{ij} = 0$. Then, trivially, there are no

[18] 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 are also 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.

[19] In reality, sometimes grading a question more generously may lead to lumpy amounts of test points (e.g. either the teacher must assign 3 extra points or 0, as she may not be able to give 1 point, given the structure of the test.)

incentives to engage in costly manipulation. Thus, all students with r_i below $\bar{p}$ fail the nationwide test ($t_i = 0$), and all students with r_i above $\bar{p}$ pass the nationwide test ($t_i = 1$).

Our understanding of the outcome in the absence of manipulation immediately highlights that, even if we were to raise β_{ij} from zero, the teacher would never inflate any student who obtains a test grade of Pass ($g_i = 1$) without manipulation. Now consider the case when $\beta_{ij} > 0$:

Lemma 1. *The teacher's manipulation of student i's test score satisfies $\Delta_i^* \in \{0, \bar{p} - r_i\}$. That is, the teacher either leaves student i's test score un-manipulated, or inflates the student's final numeric grade to exactly $\bar{p}$.*

If the teacher chooses to engage in any manipulation of student i's raw test score, then she puts the student's final numeric score exactly at $\bar{p}$, where the student (just) receives a passing final grade, $t_i = 1$. Intuitively, the teacher never inflates a student's test 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 ($t_i = 0$) to Pass ($t_i = 1$). Put differently, the teacher engages in manipulation only if it brings her an added utility of β_{ij}. Similarly, as $c_i(\Delta_i)$ is strictly increasing, the teacher never engages in more inflation than what puts the student's final numeric grade at $\bar{p}$.

This immediately implies that the teacher's decision of whether to inflate 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 $\bar{p}$. This required amount of manipulation is given by $(\bar{p} - r_i)$. Thus, the teacher inflates student i if and only if he would fail in the absence of manipulation and

$$\beta_{ij} \geq c_{ij}(\bar{p} - r_i). \tag{1}$$

The left-hand side of equation (1), β_{ij}, is a constant and the right-hand side, $c_{ij}(\bar{p} - r_i)$, is increasing in $\bar{p} - r_i$ (decreasing in r_i). Hence, equation (1) can equivalently be formulated as follows:

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 $\beta_{ij} = c_{ij}(\bar{p} - r_{ij,\min})$.*[20]

Proposition 1 highlights three key things:

[20] 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 $r_i \in [r_{ij,\min}, \bar{p})$.

- **Student-specific decision rules**: Because $r_{ij,\min}$ varies at the student level, the teacher has a set of student-specific decision rules: If the teacher has two students (say, Ben and Anna), and if the pass threshold $\bar{p}$ is 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.

- **Differential treatment**: An immediate implication of the student-specific rule is that a teacher who has two students with the same test score may bump one of them up, and leave the other behind. In the previous example, if both Ben and Anna got a test score of 17, the teacher would choose to bump up Anna, but not Ben.

- **Decision rules are independent of students' 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 $\bar{p}$).*[21] *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,

$$r_{j,min} = \min_i(r_{ij,min}). \tag{2}$$

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.

As we discuss in detail in Section 6 below, the existence of a minimum test score at which manipulation occurs will be at the heart of our estimation methods for identifying the long-run impact of receiving test score manipulation, as it implies that there exist test score ranges where students are left un-manipulated with probability one; thus, we can use these students to learn about how students who were manipulated would have fared in the absence of manipulation.

[21] 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.

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, $r_{g,\min}$.*

This result follows immediately from Proposition 2: the lowest test score at which any manipulation occurs in a school s, $r_{s,\min}$, is simply given by the minimum $r_{j,\min}$ among all teachers j at school s. In a similar vein, the lowest test score at which any manipulation occurs in a geographical region g, $r_{g,\min}$, is given by the minimum $r_{s,\min}$ among all schools s within geographical region g. This previews what we discuss in Section 6 in more detail; namely, that we can use the lowest test score at which any manipulation takes place to identify un-manipulated regions of the test score distribution *at any level of aggregation of our data* – at the teacher level, the school level, at the level of a geographical region (or even at the national level).

In Appendix C, we present a slightly less general model that places some restrictions on the teacher's utility function, but which more closely captures the precise institutional features of the grading system in Sweden. In this alternative setting, we show the same result, namely, that there exists a lowest test score at which manipulation occurs. In addition, the framework presented in Appendix C permits a structural interpretation of this parameter: The lowest test score at which test score manipulation occurs in a school, $r_{s,\min}$, identifies the school's desire to engage in test score manipulation. More precisely, if school A has a lower $r_{s,\min}$ than school B, then school A has a stronger inclination to engage in test score manipulation (regardless of whether, say, the underlying ability distributions of school A and school B differ). The simpler and more general framework presented in this Section, however, highlights that the existence of $r_{s,\min}$ requires only minimal assumptions. Moreover, for quantifying the impact of test score manipulation on long-term outcomes, we do not need to interpret the parameter $r_{s,\min}$ as capturing grading leniency.

5 Demarcating the regions of test score manipulation

The central insight from the framework presented in the previous Section is that there exists a minimum test score at which manipulation occurs (around each grade threshold). This minimum test score is a crucial parameter as it demarcates the manipulation region (in the vicinity of a given grade threshold), and thus divides the test score range into manipulated and un-manipulated regions – the key information that we need in Section 6 to quantify the causal impacts of test score manipulation. In this section, we turn to the data and estimate this key parameter (for the Pass and PwD thresholds, respectively).

Since the parameter $r_{s,\min}$ ($r_{g,\min}$) is closely related to a school's (region's) overall desire to manipulate, we will henceforth use the term "grading leniency" to refer to this parameter.

5.1 Estimation of grading leniency

We estimate grading leniency parameters for each county, separately for voucher and non voucher schools, in each year from 2004 to 2010. By aggregating many schools together at the county level, we identify the minimum test score where manipulation occurs across all these schools (i.e., a min of mins, as formalized in the previous section). This is a lower bound, across all (voucher or public) schools in the region, for where in the test score distribution manipulation occurs.[22]

In theory, our method for quantifying causal impacts of test score manipulation, presented in Section 6, would work equally well by estimating this cut-off in the aggregate nationwide test score distribution (for a given year); this would identify the minimum test score that gets inflated across all schools in that year. We discuss how the level of aggregation impacts our treatment effect estimation in Section 6.1.

For each county*voucher*year, we estimate two grading leniency parameters, capturing the width of the manipulation window around the Pass and PwD thresholds, respectively.

To do this, we analyze the histogram of the test score distribution for each county-voucher school-year. We want to estimate the point, below each of the two test grade thresholds, at which there begins to be missing mass in the test score distribution (and where this missing mass instead is shifted above the test grade threshold, into the "bunching region").[23] Let the two thresholds – for Pass and PwD – be indexed by k.

5.1.1 Refining existing bunching methodologies

Previous bunching estimators have relied on visual inspection for determining where the manipulation region begins and/or ends (Saez, 2010; Chetty et al., 2011b; Kleven and Waseem, 2012). 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

[22] We are unable to estimate separate grading leniency parameters at the individual school level since there are too few students in a school to give us sufficient statistical power. We instead estimate grading leniency in each county, separately for voucher and non voucher schools, in each year from 2004 to 2010. 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.

[23] Some bunching may occur above the Pass threshold, instead of exactly at the Pass threshold, due to teachers being imprecise with their grading or the test points being structured in a way where the points awarded for questions are lumpy, forcing teachers to sometimes give more points than what is needed to pass the test.

around a notch begins. KW's method relies on visual inspection of where manipulation of the analyzed distribution ends. In our setting, 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 relying on visual inspection tedious. Second, and most importantly, we want to allow for the possibility that there is no manipulation in some locations. We therefore refine the methods of KW to create a "fully automatic" estimator, which does not require picking any parameters using visual inspection.

More specifically, because we want an estimator that can identify zero manipulation, we first must make some assumption on the shape of the un-manipulated density. This is because, without a restriction on the shape of the un-manipulated test score distribution, one could not reject that any observed bunching simply represents an unusual looking test score distribution *without* manipulation. To rule out this possibility, we assume that the test score distribution is log concave in the absence of grade inflation. 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. Further, it allows for considerable generality: many commonly used probability distributions are log concave (normal, gumbel, gamma, beta, logistic).

Having specified the shape of the un-manipulated distribution 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.[24] In addition, we make a number of assumptions that are common for our estimator and that of KW, such as imposing that the missing mass below the test grade threshold must equal the excess mass above the test grade threshold.

5.1.2 Our estimator

Specifying the counterfactual (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.[25] $h_{jt}(r)$ is defined as:

[24] 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.

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

$$h_{jt}(r) = exp(\delta_{rjt}) \tag{3}$$

$$s.t.$$

$$\delta_{rjt} - \delta_{r-1jt} \leq \delta_{r-1jt} - \delta_{r-2jt}. \tag{4}$$

The frequency of each test score r in region j and year t is represented by a dummy variable, $exp(\delta_{rjt})$. Note that without any constraints, this is a completely non-parametric specification, as the number of dummy variables in each region-year is equal to the number of test points. We constrain this non-parametric specification of the test score distribution to be log-concave, which is captured by $\delta_{rjt} - \delta_{r-1jt} \leq \delta_{r-1jt} - \delta_{r-2jt}$. This constraint ensures that the log of the test score distribution is concave (the change in δ_{rjt} is weakly decreasing as r increases.)[26] δ_{rjt} are parameters to be estimated.

Additionally, we impose that the estimated un-manipulated distribution sums to one, to ensure that it is a valid probability distribution:

$$\sum_r exp\left(\delta_{rjt}\right) = 1. \tag{5}$$

Missing and excess mass. We assume that, in the regions where some test scores are inflated above the test grade threshold, the missing mass below the test grade threshold must equal the excess mass above the test grade threshold. This is simply an adding up condition.[27] Let $m_{jt}^{low,k,p_{kjt}^{low}}(r)$ equal the amount of missing mass below grade threshold k at test score r in region j and year t. Similarly define $m_{jt}^{high,k,p_{kjt}^{high}}(r)$ as the amount of excess mass above grade threshold k at test score r in region j in year t. We parameterize $m_{jt}^{high,k}(\theta_{jt}^{high,k}, r)$ and $m_{jt}^{low,k}(\theta_{jt}^{low,k}, r)$ each as polynomials, where $(\theta_{jt}^{high,k}, \theta_{jt}^{low,k})$ are the coefficients of the polynomials and $(p_{kjt}^{high}, p_{kjt}^{low})$ are the orders of the polynomials. These functions are constrained by:

$$\sum_r m_{jt}^{high,k,p_{kjt}^{high}}(\theta_{jt}^{high,k}, r) = \sum_r m_{jt}^{low,k,p_{kjt}^{low}}(\theta_{jt}^{low,k}, r). \tag{6}$$

[26] To see that this restriction implies log concavity, note that δ_{rjt} is equal to the log of the share of students who would receive a test score of r in the absence of manipulation. To verify concavity, we ensure that the change in the log share of students receiving the un-manipulated test score r ($\delta_{rjt} - \delta_{r-1jt}$) is weakly decreasing in r.

[27] 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, must be located above the threshold in the manipulated distribution.

In addition to the adding up constraint, 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 non-negative at all test scores r:

$$\text{For all } r : m_{jt}^{high,k,p_{kjt}^{high}}(\theta_{jt}^{high,k}, r) \geq 0, \tag{7}$$

$$\text{For all } r : m_{jt}^{low,k,p_{kjt}^{low}}(\theta_{jt}^{low,k}, r) \geq 0. \tag{8}$$

This guarantees that there can only be missing mass below each threshold k, and not excess mass. Similarly, there can only be excess mass above the test score threshold, and not missing mass.

Width of the manipulation region Third, we define β_{kjt} as the difference between the test grade threshold k and the minimum test score to ever receive inflation in region j in year t. Test scores below $k - \beta_{kjt}$ will never be inflated up to k or higher. 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. β_{kjt} is our 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. We also use β_{kjt} as an upper bound on how far beyond the grade cutoff a test score can be manipulated. If β_{kjt} is the most points a test score can be inflated to reach a grade cutoff, this should also bound how many points beyond the cutoff a score could be manipulated.[28]

$$\text{If } r < (k - \beta_{kjt}) : m_{jt}^{low,k,p_{kjt}^{low}}(\theta_{jt}^{low,k}, r) = 0, \tag{9}$$

$$\text{If } r > (k - 1) : m_{jt}^{low,k,p_{kjt}^{low}}(\theta_{jt}^{low,k}, r) = 0, \tag{10}$$

$$\text{If } r > (k + \beta_{kjt} - 1) : m_{jt}^{high,k,p_{kjt}^{high}}(\theta_{jt}^{high,k}, r) = 0, \tag{11}$$

$$\text{If } r < (k) : m_{jt}^{high,k,p_{kjt}^{high}}(\theta_{jt}^{high,k}, r) = 0. \tag{12}$$

Full test score distribution. Combining these gives us the full test score distribution, including test score manipulation. 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:

[28] A teacher might manipulate scores beyond the exact threshold because they are engaging in manipulation as they go through the grading and end up with more manipulated points than necessary once they finish. The teachers may also try to overshoot the cutoff for some student in order to obfuscate that they are manipulating scores.

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

such that equations (4), (5), (6), (7), (8), (9), (10), (11), and (12) hold.

Estimation. We estimate the model using constrained nonlinear-least squares and use k-fold (k=5) cross-validation to prevent overfitting.[29] The parameters to estimate are:

$(\beta_{1jt}, \beta_{2jt}, \theta_{jt}^{low,1}, \theta_{jt}^{high,1}, \theta_{jt}^{low,2}, \theta_{jt}^{high,2}, \delta_{1jt-1}, ..., \delta_{R^{max}jt-1})$. We estimate the model separately for voucher and municipal schools, within each county, in each year. This allows us to recover the maximum amount of test score manipulation that occurs around each of the test grade thresholds: $(\beta_{1jt}, \beta_{2jt})$. See Appendix D for additional technical details.

Intuition for identification of key parameters of interest. To give some intuition behind how our estimator identifies β_1 and β_2, Figure 1a 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, Figure 1b 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 $\hat{\beta}_2 = 1$.

5.2 Estimates of grading leniency

The estimation strategy outlined above yields estimates of grading leniency at the county by voucher level, for each year between 2004 and 2010. Figure 2 displays histograms of $\hat{\beta}_{1jt}$ (upper panel) and $\hat{\beta}_{2jt}$ (lower panel) – that is, of the estimated widths of the manipulation region around the thresholds for Pass and PwD, respectively. Note that we do not use this cross-sectional variation when identifying the impact of test score manipulation on future

[29] 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 β_1 and β_2 estimates, as well as the orders of the polynomials, to minimize this out-of-sample mean-squared error. We do this separately for each county-voucher-year.

outcomes – we simply show this distribution to give an idea of how much manipulation takes place in Swedish schools.

The estimates show considerable heterogeneity across counties (by voucher) in grading leniency: some students' test scores are virtually un-manipulated, while other (marginal) students' test scores are inflated by as much as 7 test score points, which amounts to about 10% of the total number of test points. While there is more manipulation around the Pass threshold, there is considerable manipulation around the higher threshold as well.

Figure 3 illustrates the estimated county*voucher counterfactuals, aggregated up to the national level, in year $t = 2010$. The blue connected line plots the observed (manipulated) distribution of test scores, and the red connected line shows the estimated (un-manipulated) counterfactual density. This "eyeball" check shows that our log-concavity assumption appears reasonable and highlights where manipulation begins and ends in the aggregate distribution of test scores.

5.3 Towards estimation of causal impacts of test score manipulation

The estimation procedure in this section yielded two pieces of information that we, in the subsequent Section, will use to estimate the causal effects of test score manipulation: (i) The width of the manipulation region, around the Pass and PwD thresholds, respectively, in each county*voucher*year. Thus, in the test score distribution for each county*voucher*year, we now know where test score manipulation occurs, and where it does not occur. (ii) The counterfactual test score distribution for each county*voucher*year. This tells us, for each test score point inside the two manipulation regions, how many students would have obtained this test score in the absence of manipulation.

Of course, these two pieces of information are simply "two sides of the same coin," as is illustrated in Figure 3. The blue connected line plots the observed data, i.e., the manipulated distribution of test scores, at the national level in 2010. The red connected line shows our estimated county*voucher counterfactuals, aggregated up to the national level, for the year $t = 2010$; that is, the second "piece of information" that we discussed above. Comparing the counterfacutal density with the actual data also immediately demarcates where manipulation begins and ends in the aggregate distribution of test scores; that is, the first piece of information that we discussed above.

Next, we describe our methodology for estimation of causal effects of test score manipulation, and how we use each of these two pieces of information.

6 The impacts of teacher discretion

To analyze the long-term consequences of test score manipulation, we develop a new bunching identification strategy that harnesses the fact – shown in the previous Section – that teacher manipulation only occurs in two regions of the test score distribution.

To see how our methodology works, it is helpful to recall some common features of bunching strategies. 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., 2011b; Kleven and Waseem, 2012), dates of marriage (Persson, 2014), or, as in the previous section of this paper, distributions of test scores. The underlying idea of these methodologies is to (i) figure out which parts of the distribution is manipulated, and then (ii) use the un-manipulated parts of the distribution to help "fill in" the shape inside the manipulated regions of the distribution.

In this Section, we further develop this literature and 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. 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 (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 relationship between mean earnings and test scores captures the underlying relationship between (un-manipulated) test scores and earnings. Thus, we can use this observed relationship between mean earnings and test scores outside of the two manipulation regions to predict a counterfactual relationship between *earnings* and test scores inside the test score ranges where manipulation occurred. In the manipulation region around the Pass threshold, the difference between the average observed and counterfactual earnings captures the reduced form impact of (potentially being exposed to) test score manipulation on earnings. The interpretation around the PwD region is analogous.

We describe this in detail in the below subsections. Here, we provide an intuitive description, which also provides a roadmap of the remainder of this Section. To implement our estimator, we proceed in several steps:

1. For each county*voucher*year, we flexibly estimate the relationship between test scores and an outcome of interest (say earnings), using only students whose test scores fall outside of the two manipulation regions. In this step, we use one piece of information from Section 5 above, namely, the estimates of where test score manipulation occurs

and where it does not occur (in each county*voucher*year).

2. We predict inwards into the two manipulation regions, assuming that the parameters that we estimated using data only outside of these regions also govern the relationship between test scores and earnings inside the manipulation regions; the only exception is that we also allow for discrete jumps in earnings exactly at the cutoffs for Pass and PwD, respectively. This step produces an estimate of the counterfactual *earnings* at each test score point inside the test score ranges where manipulation occurred.

3. We then focus only on the part of the test score distribution where manipulation occurs. In particular, from (2) we can construct a prediction of the *average* earnings of students whose un-manipulated test scores fell in each of the two manipulation regions. (In this step, in addition to using (2) above, we use a piece of information from Section 5 above, namely, the estimated counterfactual densities of the *test score* distribution.)

4. We calculate, from the raw data, the average *actual* earnings of the students whose un-manipulated test scores fell in each of the two manipulation regions.

5. In the manipulation region around the Pass threshold, the difference between (4) and (3) captures the reduced form impact of (potentially being exposed to) test score manipulation on earnings. The interpretation around the PwD region is analogous.

6. Not all students whose un-manipulated test scores fall within the manipulation regions of the test score distribution actually get bumped up. We therefore also estimate a "first stage" effect of having a test score that falls in the manipulation region on the probability of receiving a higher grade. To estimate this first stage, we repeat steps (1) - (5) above, with the outcome being "receiving a higher math grade." (This, in fact, is the estimation that we show in detail below.)

7. Finally, the ratio of the reduced-form effect to the first stage effect identifies the local average treatment effect of receiving a higher grade *due to test score manipulation* on labor market earnings.

To provide some intuition for (6) above – the fact that not all students whose un-manipulated test scores fall in a manipulation region actually gets bumped up, which creates the need for a "first stage" – Figure 4 illustrates how students in the the manipulated regions of the test score distribution can be thought of in terms of the potential outcomes framework, in an example where the Pass threshold is 21 and the manipulation region starts at 14. 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 The key identifying assumption that we rely on for identifying causal impacts of test score manipulation is that, in the absence of manipulation, the distribution of outcomes that we consider *would have been smooth across the thresholds to the manipulation windows*. This ascertains that data from outside of the manipulation regions is informative of the counterfactual distribution within these regions. This is the standard assumption that underlie all bunching methodologies. Moreover, recall from Section 4 that the teacher-specific (and, hence school-specific and region-specific) thresholds are pre-determined, and do not depend on the realization of student test scores.

Hence, it is worth emphasizing that we *do not* need to assume that teachers allocate test score manipulation to a random subset of the students. In contrast, we know from our theoretical framework in Section 4 that the students who are bumped up – i.e., the compliers in Figure 4 – are a select subset. Our reliance on bunching methods overcomes the identification challenge that this would pose to a simple comparison of students who are bumped up with students who are not.

6.1 Identifying the causal impact of test score manipulation

The "first stage." We formally present steps (1)-(5) of the above procedure with the outcome being the student's final grade in math (obtained in June). This yields our "first stage" of our Wald estimation.[30]

Steps one and two. The first step is to estimate the relationship between students' final math grades and their un-manipulated test scores. We start by estimating this relationship, by fitting a third order polynomial, using data only from the un-manipulated parts of the test score distribution. We then predict this relationship inwards into the manipulation regions.

[30] In other words, we implement step (6) of the above procedure, which is to implement steps (1)-(5) with the outcome being the student's final grade in math.

Specifically, recall that g_{ijt} is student $i's$ observed final math grade (who is enrolled in region j in year t). We estimate:

$$g_{ijt} = \hat{g}_{kjt}\left(r_{ijt}, \theta^{grade}_{kjt}\right) + \alpha_{kjt} * (r_{ijt} \geq k) + \epsilon^{g}_{ijt}, \tag{13}$$

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

$$\text{and}$$

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

$\hat{g}_{kjt}\left(r_{ijt}, \theta^{grade}_{kjt}\right)$ is a third order polynomial with coefficients θ^{grade}_{kjt}, which captures the smooth relationship between students' un-manipulated test scores, r_{ijt}, and their expected final grades. $(r_{ijt} < k - \beta_{kjt}$ or $r_{ijt} > k + \beta_{kjt} - 1)$ ensures that the data used to estimate equation (13) 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 final grade 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 (13) yields the expected final math grade at each point in the manipulation region *had students not received test score manipulation*. This "constructed control group" comes from students *within the same region/school* as the students exposed to grade manipulation. (We do not rely on the variation across regions and years in the width of the manipulation region.)

Example. Figure 5 illustrates the outcome of steps one and two of our first stage estimation in the context of the example setting described in Figure 4. The solid red vertical lines mark the contours of the manipulation region (around the Pass threshold), and the dark (blue) solid line shows the average observed grades at each test score (in the data). Using data only from the un-manipulated parts of the test score distribution (below 14; and above 26 but below the start of the manipulation region around PwD), we then predict inwards into the manipulation region around the Pass threshold. The light gray line illustrates the expected final math grade at each point in the manipulated region *had students not received test score manipulation* – i.e., if all students were never-takers or always-takers.

Step three. We now compute the expected final grades across the *entire* set of students in the manipulation region of the test score distribution. We do this because we are unable to compute the counterfactual final grades only for the students who are bumped up (i.e.,

only for the compliers); this, in turn, is because we cannot separate the compliers from the always-takers in the data.

From equation (13), we have the expected final grade at each test score inside the manipulation region, in the counterfactual scenario with no test score manipulation. We now combine this with our estimates of the counterfactual test score distribution, that is, the share of students who would have received each test score, had there been no test score manipulation, $\hat{h}_{jt}(r)$. We recovered this counterfactual distribution of test scores during the estimation of the grading leniency parameters in Section 5 above. We combine these to calculate the expected average final math grade for students within the manipulation region of the test score distribution *had there been no test score manipulation*:

$$\begin{aligned} &E\left(\text{grade in jt}|\text{teacher can't manipulate, r in manipulation region k}\right) \equiv \bar{g}_{jt}\left(k\right) \\ &= \int_{k-\beta_{kjt}}^{k+\beta_{kjt}-1} E\left(g_{kjt}|r, \text{No manip}\right) * \frac{Pr\left(r|\text{No manip, jt}\right)}{\int_{k-\beta_{kjt}}^{k+\beta_{kjt}-1} Pr\left(r|\text{No Manip, jt}\right) dr} dr. \\ &= \int_{k-\beta_{kjt}}^{k+\beta_{kjt}-1} \left[\hat{g}_{kjt}\left(r, \hat{\theta}_{kjt}^{grade}\right) + \hat{\alpha}_{kjt} * \left(r \geq k\right)\right] * \frac{\hat{h}_{jt}(r)}{\int_{k-\beta_{kjt}}^{k+\beta_{kjt}-1} \hat{h}_{jt}(r) dr} dr. \end{aligned} \tag{14}$$

Steps four and five. For students inside the manipulation region, we now compare the estimated counterfactual average grade, had there been no test score manipulation, calculated in (14), with the actual average final math grade for students in the manipulation region (observed in the data), g_{ijt}. This difference is entirely driven by the fact that compliers 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 final grade due to the student having a raw test score that falls within the manipulation region of the test score distribution:

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

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

in region(*voucher) j in year t. Figure 6 illustrates this first stage estimate in the context of the example used in Figure 5.

The "reduced form" and LATE estimates. The five-step procedure above can be repeated with a different outcome variable, such as income at age 23, to identify the reduced-form effect of falling into the manipulation region on future income. The ratio of this reduced-form effect to the first-stage effect, in turn, identifies the local average treatment effect (LATE) of receiving an inflated final math grade on future income.[31]

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 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.[32]

6.2 Identifying the beneficiaries of test score manipulation

We develop new methods to recover observable summary statistics of the types of students that teachers select to grade up, i.e., the compliers. These methods can be used more generally for any type of bunching estimation to recover observable characteristics of those responding to the incentive to bunch at the threshold.

[31] 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.

[32] 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.

For any observable characteristic of the students, Y, we can apply the same type of method as in equation (13) to use students outside of the manipulation region to estimate $E(Y|r)$ at any test score r inside the manipulation region.

$$Y_{ijt} = \hat{g}^{Y}_{kjt}\left(r_{ijt}, \theta^{grade}_{kjt}\right) + \epsilon^{g}_{ijt}, \tag{15}$$

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

$$\text{and}$$

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

For example, if Y were a dummy variable for being an immigrant, we could estimate the expected share of immigrant children at each test score, 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, $\bar{Y}^{down\_all}$.[33]

$$\bar{Y}^{up\_all} = \frac{1}{N^{tot}_{up}} \sum_{it} Y_{ijt}, \tag{16}$$

$$\text{where: } k \leq t_{ijt} \leq k + \beta_{kjt} - 1,$$

$$\bar{Y}^{down\_all} = \frac{1}{N^{tot}_{down}} \sum_{it} Y_{ijt}, \tag{17}$$

$$\text{where: } k - \beta_{kjt} \leq t_{ijt} \leq k - 1.$$

$\bar{Y}^{up\_all}_{t}$ is an average of those who were inflated to these scores ("compliers"), as well as students who naturally received a passing test score absent manipulation ("always-takers"):

$$\bar{Y}^{up\_all} = \frac{N_{up}}{N_{up} + N_{compliers}} * \bar{Y}^{up} + \frac{N_{compliers}}{N_{up} + N_{compliers}} * \bar{Y}^{compliers}.[34] \tag{18}$$

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

$$\bar{Y}^{down\_all} = \frac{N_{down}}{N_{down} - N_{compliers}} * \bar{Y}^{down} - \frac{N_{compliers}}{N_{down} - N_{compliers}} * \bar{Y}^{compliers}.[35] \tag{19}$$

[33] 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.

[34] N_{up} is the number of students who earned an un-manipulated test score above the grade cutoff, within the manipulation region. These are the always-takers. $\bar{Y}^{up}$ is the average level of Y for the always-takers.

[35] 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

We can recover the expected share of immigrant students within these regions of the distribution, using our extrapolation from equation (15) and the estimated un-manipulated distribution, $\hat{h}_{jt}(r)$:

$$\bar{Y}^{up} = \sum_j \left(N_j \int_k^{k+\beta jt-1} \hat{g}^Y_{kjt}\left(r, \theta^{grade}_{kjt}\right) * \hat{h}_{jt}(r)dr \right) \tag{20}$$

$$\bar{Y}^{down} = \sum_j \left(N_j \int_{k-\beta jt}^{k-1} \hat{g}^Y_{kjt}\left(r, \theta^{grade}_{kjt}\right) * \hat{h}_{jt}(r)dr \right). \tag{21}$$

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

$$N^{tot}_{up} = N_{up} + N_{compliers},$$

$$N^{tot}_{down} = N_{down} - N_{compliers},$$

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

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

Plugging these into equations (20) and (21) and solving for the mean immigrant share of the compliers gives:

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

$$+0.5 * \left(\frac{N_{down}}{N^{down} - N^{tot}_{down}} \bar{Y}^{down} - \frac{N^{tot}_{down}}{N^{down} - N^{tot}_{down}} * \bar{Y}^{down\_all} \right).^{36}$$

Intuitively, if teachers are disproportionately choosing 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 region 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 un-manipulated test scores fell within the manipulation region of the test

the never-takers and the compliers.

[36] $\bar{Y}^{compliers}$ 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.

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 inflate 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 the previous section to analyze the observable characteristics of students who are actually chosen to be graded up, and compare them to all students who fall right below the relevant test grade cutoff and *could* be chosen by teachers to receive an inflated grade.

For a set of predetermined outcomes, Table 2 compares the average among all students who are eligible for inflation (Column one) with the average among the complier students (those inflated up; 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 they 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 student eligible for inflation) to have passed their national test in English. Around the the PwD margin, inflated students are 33 percentage points more likely than those eligible for inflation 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: Inflated 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 inflate students who have higher grades on other, pre-determined tests than the average student who is eligible for inflation. This is consistent with the selection of students for inflation 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' inflation choices are related to students' demographics, the next row in Table 2 compares the male share of students eligible for inflation with the male share of inflated students. We see a very precisely estimated zero effect around both thresholds, showing that teachers treat boys and girls equally when choosing whom to inflate. Similarly, the next row of Table 2 shows that inflated students are not selected based on whether they come from an immigrant household. These results are reassuring in that teachers do not appear to bias their math test grading based on race or gender.

Next, we turn to whether inflated students come from disadvantaged backgrounds. Around the Pass margin, inflated students are positively selected on household income: inflated students' household income is 3.9 percent higher than the household income of the average student who is eligible for inflation. Around the PwD margin, in contrast, the point estimate implies that household incomes of inflated students are 3.2 percent *lower* than the household income of the average student who is eligible for inflation; 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. Inflated 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 the 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 inflate students whose parents may have more free time to pressure teachers into giving their children high grades, we look at whether inflation is selected on whether students have a stay at home parent. The last row of Table 2 shows a zero effect of selectively inflating 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.[37] 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 these 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 the 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 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

[37] 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.

(awarded in February). Appendix Table A1 estimates the impact of receiving an inflated math test grade on the final math grade. Around the Pass margin, receiving an inflated test grade leads to a 35 percentage point increase in the probability of receiving a passing final grade, and around the PwD margin, an inflated 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 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 grade inflation cannot change their grades on previous tests. We expect to see that our estimator finds this to hold. Appendix Table A2 shows that for both the Pass and PwD thresholds, there is no causal effect of grade inflation 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 A3). 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 inflation 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, inflation 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 inflation raises the student's performance substantially over and above the mechanical effect induced by the 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. This highlights that their is a strong signaling value from the math test grade, especially 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 bigger, showing 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 how endogenous math grades are 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, grade inflation 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 inflation over the Pass threshold causes a 1.4 point higher high school GPA, relative to a base of 11.9. Inflation 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 inflated 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 A4 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 inflation 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 inflated 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 340 SEK, relative to a mean of 1580. This is a large, 20% increase in earnings at age 23, and it is quite similar to the OLS relationship of 370. 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 448 SEK higher age 23 income, relative to a base mean of 1461. This is quite different than the OLS relationship which shows an income *decrease* of 193 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.. 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

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.

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).

Billings, Stephen B., David J. Deming, and Jonah Rockoff, "School Segregation, Educational Attainment, and Crime: Evidence from the End of Busing in Charlotte-Mecklenburg," *The Quarterly Journal of Economics*, 2014, *129* (1), 435–476.

Björklund, Anders, Melissa A. Clark, Per-Anders Edin, Peter Fredriksson, and Alan B. Krueger, *The Market Comes to Education in Sweden: An Evaluation of Sweden's Surprising School Reforms* UPCC book collections on Project MUSE, Russell Sage Foundation, 2006.

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.

Chetty, Raj, John N. Friedman, and Jonah E. Rockoff, "Measuring the Impacts of Teachers II: Teacher Value-Added and Student Outcomes in Adulthood," *American Economic Review*, 2014, *104* (9), 2633–79.

_ , _ , Nathaniel Hilger, Emmanuel Saez, Diane Whitmore Schanzenbach, and Danny Yagan, "How Does Your Kindergarten Classroom Affect Your Earnings? Evidence from Project Star," *The Quarterly Journal of Economics*, 2011, *126* (4), 1593–1660.

_ , 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.

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.

Deming, David J., Justine Hastings, Tom Kane, and Doug Staiger, "School Choice, School Quality and Postsecondary Attainment," *American Economic Review*, 2014, *104*, 991–1013.

Edmark, Karin, Markus Frölich, and Verena Wondratschek, "Sweden's school choice reform and equality of opportunity," *Labour Economics*, 2014, *30* (2), 129 – 142.

Fredriksson, Peter, Björn Öckert, and Hessel Oosterbeek, "Long-Term Effects of Class Size," *The Quarterly Journal of Economics*, 2012.

Garces, Eliana, Duncan Thomas, and Janet Currie, "Longer-Term Effects of Head Start," *American Economic Review*, 2002, *92* (4), 999–1012.

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.

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

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.

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 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.

Krueger, Alan B. and Diane M. Whitmore, "The Effect of Attending a Small Class in the Early Grades on College-test Taking and Middle School Test Results: Evidence from Project Star," *The Economic Journal*, 2001, *111* (468), 1–28.

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

_ , "Effects of Free Choice Among Public Schools," *The Review of Economic Studies*, 2010, *77* (3), 1164–1191.

_ , "Long Run Effects of Free School Choice: College Attainment, Employment, Earnings, and Social Outcomes at Adulthood," *NBER Working Paper No. 20843*, 2015.

_ **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.

_ **, Avraham Ebenstein, and Sefi Roth**, "The Long Run Human Capital and Economic Consequences of High-Stakes Examinations," *NBER Working Paper No. 20647*, 2014.

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.

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.

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.

Åsa Ahlin, "Does School Competition Matter? Effects of a Large-Scale School Choice Reform on Student Performance," *Uppsala University Working Paper Series*, 2003, (2).

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

Sandström, F. Mikael and Fredrik Bergström, "School vouchers in practice: competition will not hurt you," *Journal of Public Economics*, 2005, *89* (2).

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

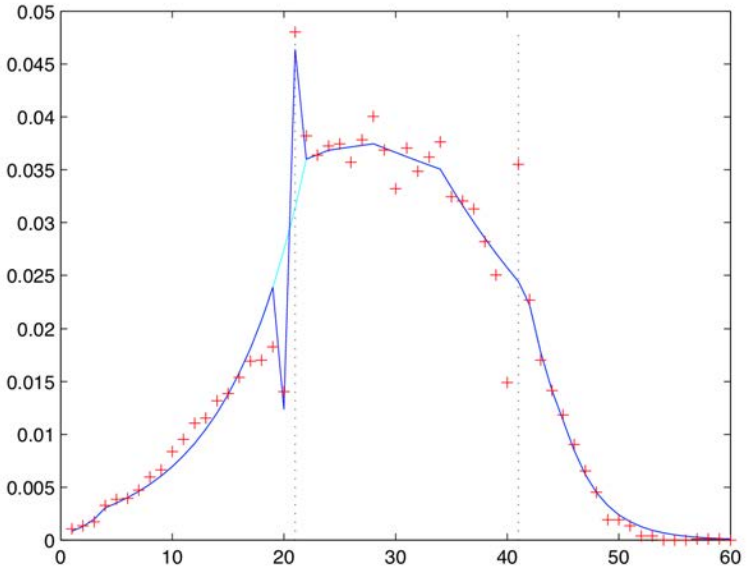
Tyler, John H., Richard J. Murnane, and John B. Willett, "Estimating the Labor Market Signaling Value of the GED," *The Quarterly Journal of Economics*, 2000, *115* (2), 431–468.

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

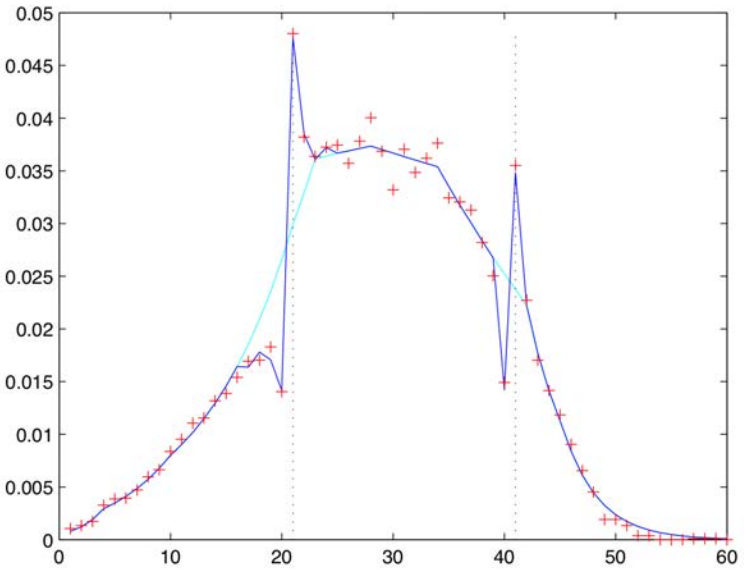
9 Figures and Tables

Figure 1: Examples of Estimates of Unmanipulated Distributions for Different Guesses of β_1 and β_2

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

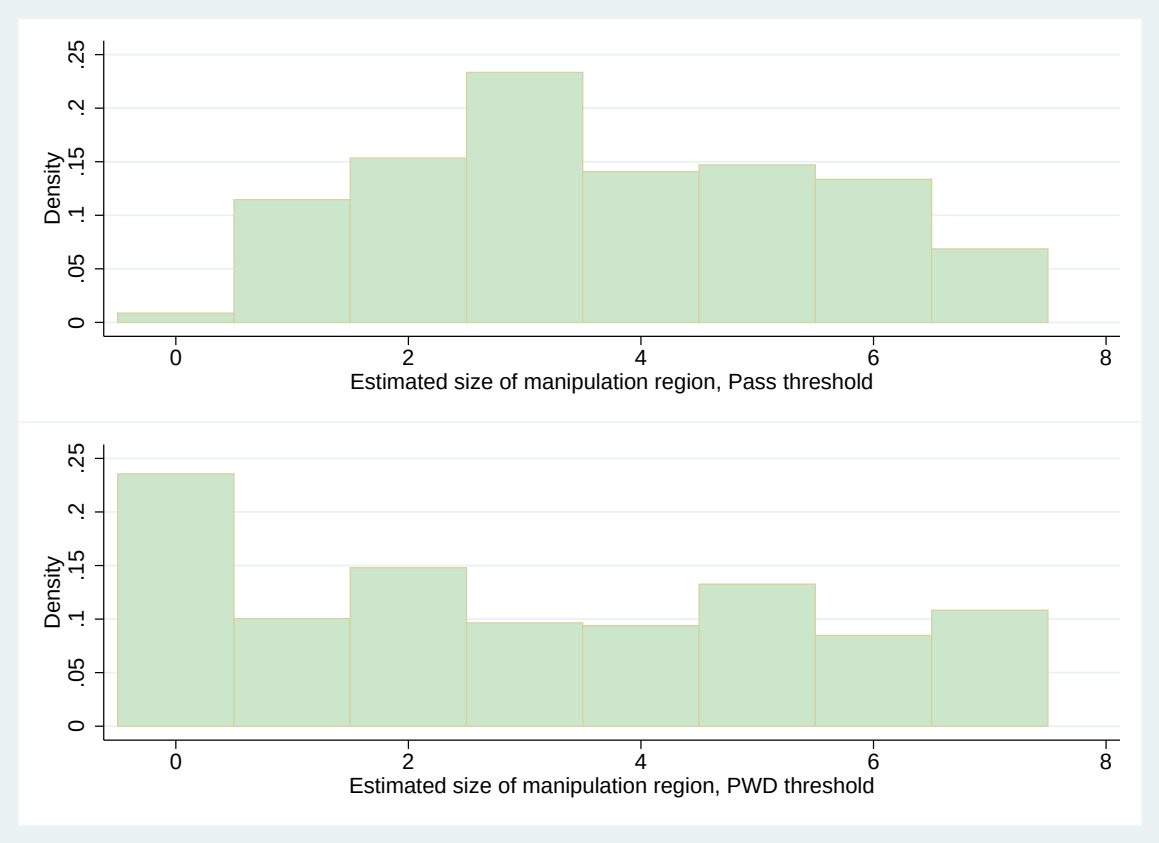


(b) $\hat{\beta}_1 = 4, \hat{\beta}_2 = 1$



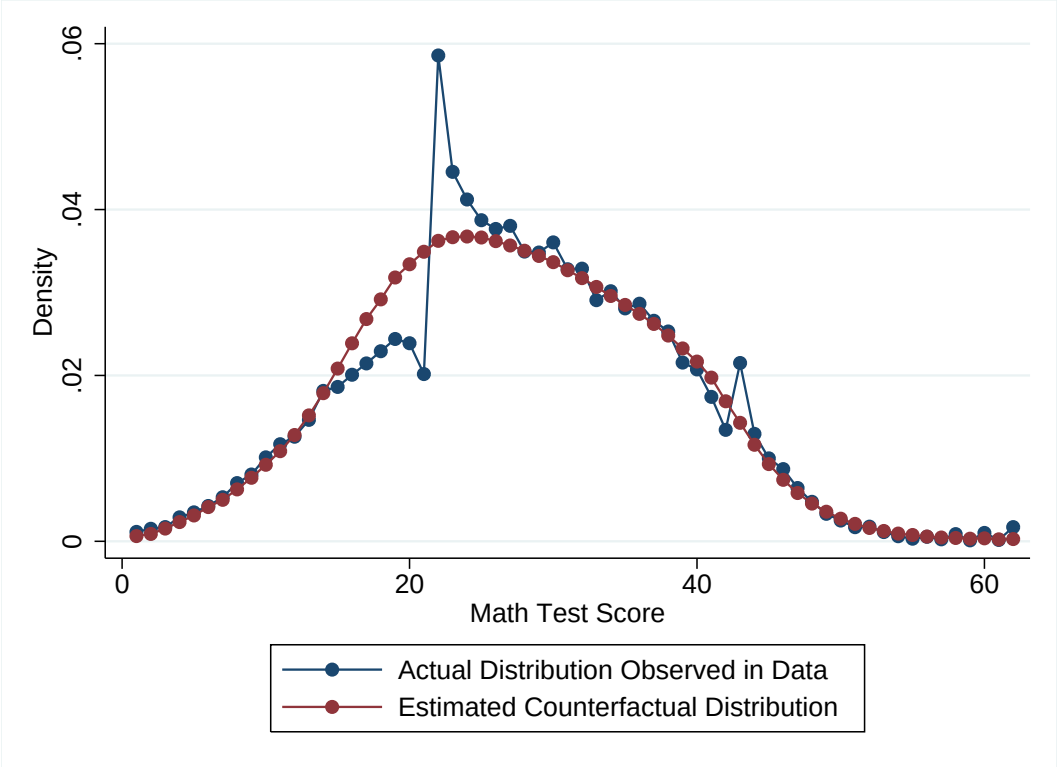
Note: In each subfigure, the red plus signs display the raw data; the blue solid line displays the estimated distribution including manipulation; and the turquoise solid line displays the estimated counterfactual (un-manipulated) distribution (which only deviates from the blue solid line where manipulation 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 1a, we display our estimate of the manipulated and un-manipulated distribution under the hypotheses that $\beta_1 = 1$ and $\beta_2 = 0$. In Figure 1b , 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 2: Distribution of Grading Leniency around the Two Thresholds



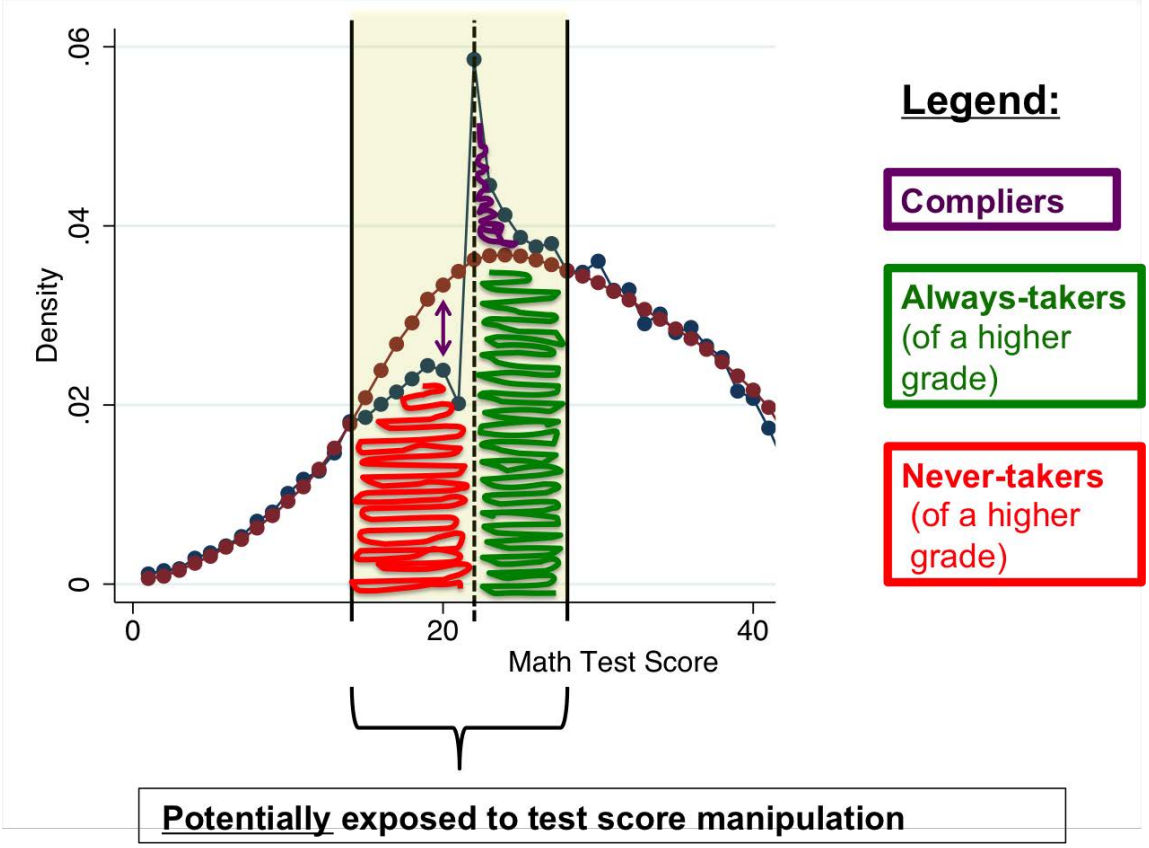
Note: The figure illustrates the distribution of the estimated sizes of the manipulation region, around the thresholds for Pass and PwD, respectively, by county*voucher*year, from 2004 to 2010.

Figure 3: National Test Score Distribution and Estimated Counterfactual, 2010



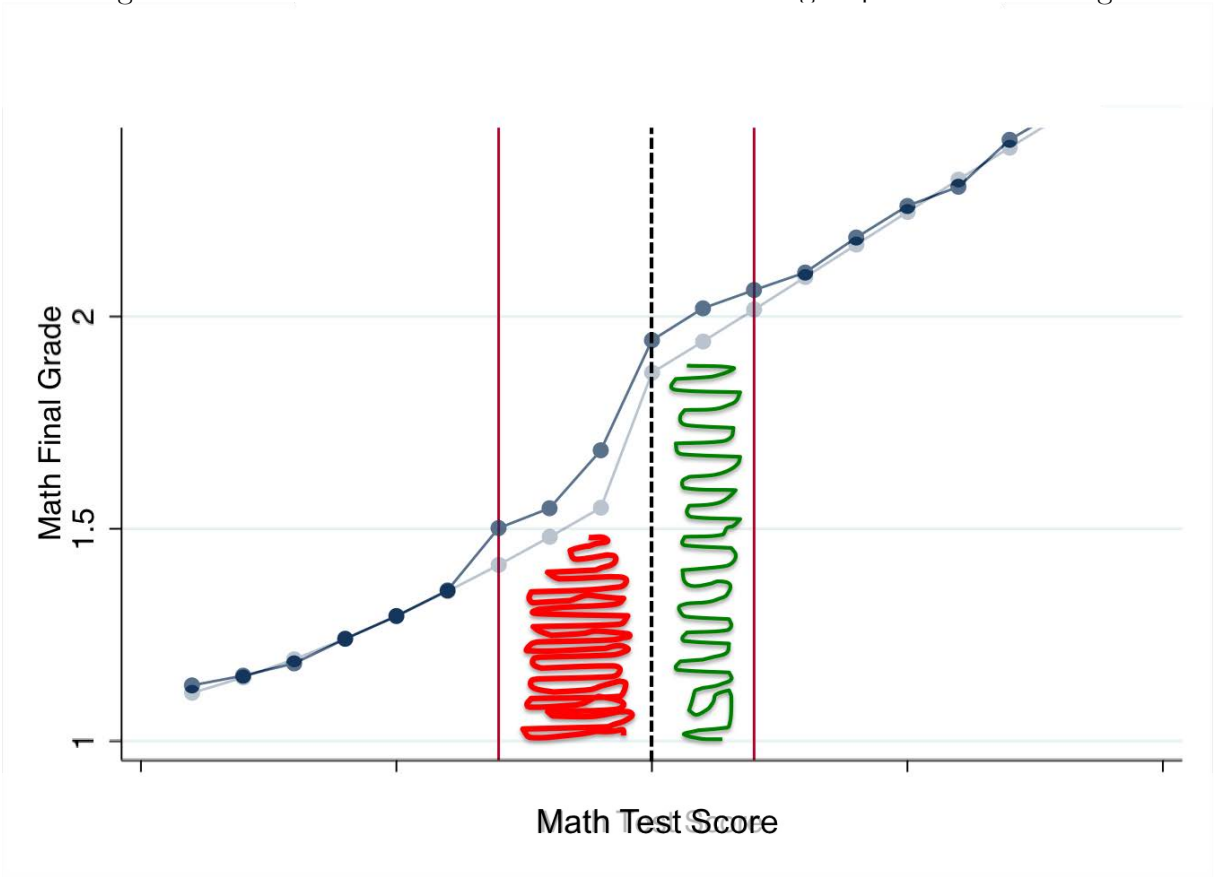
Note: The figure illustrates the national test score distribution and the estimated counterfactual (aggregated from the county*voucher estimated counterfactuals) in 2010. The estimation of the counterfactual density is described in Section 5. The blue connected line plots the actual distribution of test scores, and the red connected line shows the estimated counterfactual density in the absence of manipulation.

Figure 4: Wald estimator



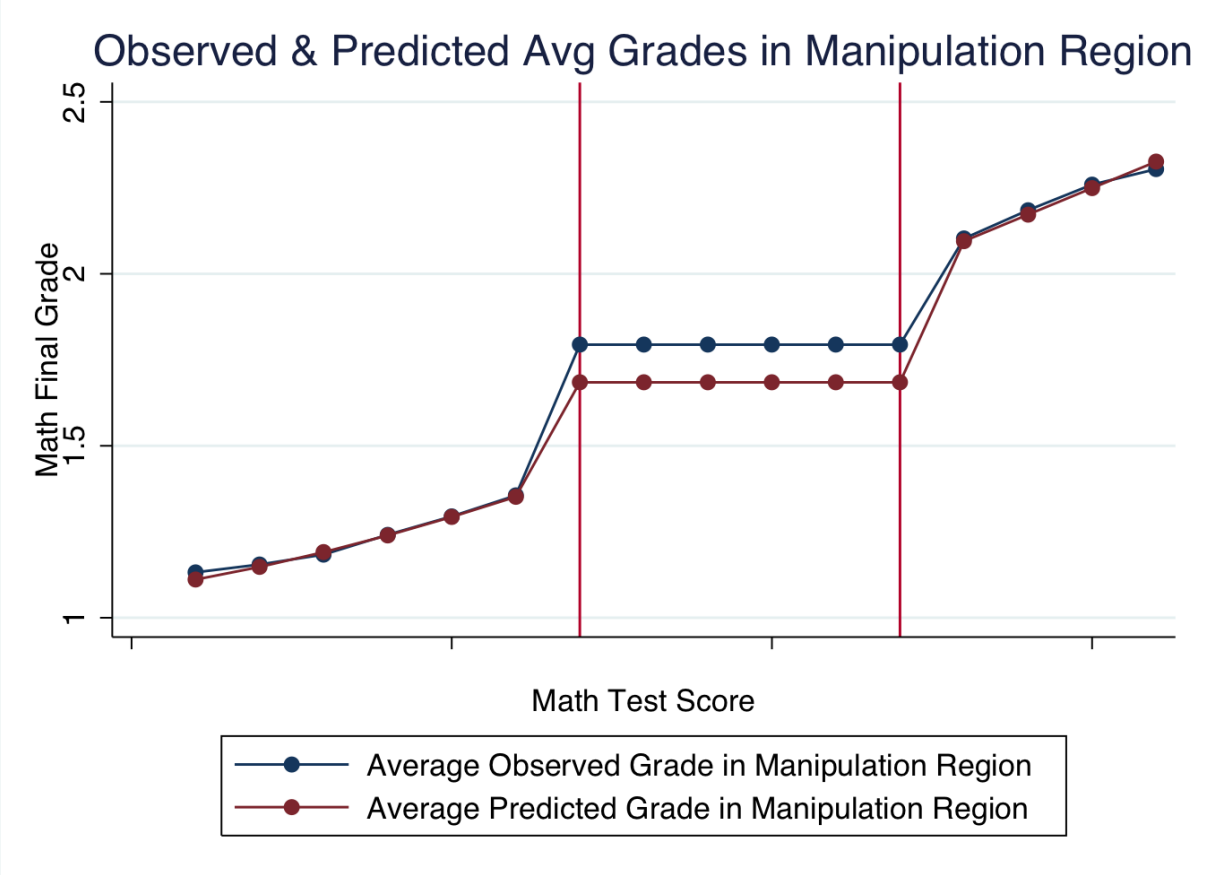
Note: The figure illustrates the test score regions of relevance to our Wald estimator in an example where the Pass threshold is 21 and the manipulation region starts at 14. Students who receive a raw (un-manipulated) test score of 13 face a zero probability of being graded up, whereas students who receive 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.

Figure 5: Wald estimator: Construction of "control group" in the first stage



Note: We estimate the relationship between students' final math grades and un-inflated test grades by first estimating the relationship using data only from the un-manipulated parts of the test score distribution. We then predict the test score inwards into the "donut" (manipulated region), extrapolating from the predicted grades at un-manipulated test scores to the left and right of the manipulated region. This Figure illustrates this in the context of the example setting described in Figure 4. The solid, red vertical lines mark the contours of the manipulation region (around the Pass threshold), and the dark (blue) solid line shows the average observed grades at each test score. Using data only from the un-manipulated parts of the test score distribution (below 14; and above 26 but below the start of the manipulation region around PwD), we then predict inwards into the manipulated Pass region. The expected final math grade at each point in the manipulated region *had students not received test score manipulation* – i.e., if all students were never-takers or always-takers – is illustrated by the bright (gray) line.

Figure 6: Wald estimator: First stage (intent-to-treat effect)



Note: Figure 6 illustrates the intent-to-treat (first stage) estimate. Inside the manipulation region, the blue solid line displays the average observed grade (in the data). Inside the manipulation region, the red solid line displays the average predicted grade, had there been no test score manipulation. This prediction is obtained by using two pieces of information: First, the expected final math grade at each point in the manipulation region *had students not received test score manipulation*, displayed by the gray line in Figure 5. Second, our estimates of the counterfactual test score distribution, that is, the share of students that had received each test score within the manipulation region if there had been no test score manipulation. We recovered this counterfactual distribution of test scores during the estimation of the grading leniency parameters in Section 5. The difference between the blue and red solid lines inside the manipulation region is entirely driven by the fact that compliers inside the manipulation region received test score manipulation. Thus, this difference is our "intent-to-treat" estimate, capturing the average increase in a student's final grade due to the student having a raw test score within the manipulated region of the test score distribution.

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 Varaible 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***	0.055*
	(0.044)	(0.034)
F Stat	308.6	185.4
Dep Varaible 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***	1.01*
	(0.49)	(0.61)
F Stat	308.6	185.4
Dep Varaible 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 Varaible 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	-0.035**
	(0.019)	(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*	448.5**
	(183.0)	(215.0)
F Stat	67.3	57.9
Dep Varaible 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

A Supplemental Tables

Table A1: Impact of Inflated Test Grade on Final Math Grade

	Panel A. Causal Impact Estimate	
	Pass	PWD
Δ Math Test Grade	0.35***	0.87***
	(0.020)	(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 A2: "Sanity Check": Impact of Grade Inflation on English Test Grade (LATE)

	Panel A. Causal Impact Estimate	
	Pass	PWD
Δ Final Math Grade	-0.019	-0.021
	(0.062)	(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 A3: "Sanity Check": Impact of Grade Inflation on Swedish Test Grade (LATE)

	Panel A. Causal Impact Estimate	
	Pass	PWD
Δ Final Math Grade	0.036	0.072
	(0.055)	(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 A4: Impact of Grade Inflation on High School *Peers'* GPA (LATE)

	Panel A. Causal Impact Estimate	
	Pass	PWD
Δ Final Math Grade	-0.14	0.012
	(0.14)	(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.

B 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:

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

$S_i \geq 23$ 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.[38] In contrast to

[38]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.

C A model of teachers' grading behavior

All proofs are presented in Online Appendix Section C.3 below.

C.1 A Model of Grade Inflation

Student i is enrolled in school j. He attends class and has performed at level a_i on class assignments, other than the nationwide test. We will refer to a_i as student $i's$ ability. Student i takes the nationwide test and receives a numeric test score r_i and a test grade of t_i as defined by:

$$r_i = r\left(a_i, \varepsilon_i, \Delta_{i1}\right) = a_i + \varepsilon_i + \Delta_{i1},$$

$$t_i = t\left(a_i, \varepsilon_i, \Delta_1\right) = \left\{ \begin{array}{c} \bar{p} \text{ if } \left(r\left(a_i, \varepsilon_i, \Delta_{i1}\right) \geq \bar{p}\right) \\ 0 \text{ o/w} \end{array} \right\}.$$

If the student does not receive any grade inflation, student i earns a test score equal to his true performance on the test: $a_i + \varepsilon_i$, where $\varepsilon_i \tilde{} F\left(\varepsilon_i\right)$ and $E(\varepsilon_i) = 0$. We refer to $a_i + \varepsilon_i$, as student $i's$ raw test score, as it is what he would receive if there was no grade inflation. ε_i represents that student i could have a "good day" or a "bad day" on the test. The teacher may also choose to inflate the test score by awarding some amount of additional test points, Δ_{i1}. If student $i's$ numeric test score is above $\bar{p}$, then he passes the test and receives a grade of $\bar{p}$, otherwise he fails the test and receives a grade of 0.

The teacher also assign student $i's$ final grade g_i for the class. g_i is defined as:

$$g_i = g\left(a_i, \varepsilon_i, \Delta_{i1}, \Delta_{i2}\right) = \left\{ \begin{array}{c} 1 \text{ if } \left[wt\left(a_i, \varepsilon_i, \Delta_{i1}\right) + (1-w)\left(a_i + \Delta_{i2}\right)\right] \geq \bar{p} \\ 0 \text{ o/w} \end{array} \right\}.$$

Student $i's$ final numeric grade is a weighted average of his test grade, $t\left(a_i, \varepsilon_i, \Delta_1\right)$, and his grade on all other class assignments, $\left(a_i + \Delta_{i2}\right)$. w measures the weight placed on the test grade in computing the final numeric grade. We refer to the grade on all class assignments

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 #.

excluding the nationwide test, $(a_i + \Delta_{i2})$, as student $i's$ homework grade. The teacher may also choose to inflate the homework grade, as measured by Δ_{i2}. Student i passes the class (and receives a final grade of 1) if his numeric final grade is above $\bar{p}$, otherwise he fails the class.

The teachers assign test grade t_i and final grade g_i to maximize the school's utility function:

$$u(\Delta_{i1}, \Delta_{i2}) = \beta_j g\left(a_i, \varepsilon_i, \Delta_{i1}, \Delta_{i2}\right) - c_1\left(\Delta_{i1}\right) - c_2\left(\Delta_{i2}\right),$$

$$c_1'\left(\Delta_{i1}\right) > 0, c_2'\left(\Delta_{i2}\right) > 0,$$

$$c_1''\left(\Delta_{i1}\right) > 0, c_2''\left(\Delta_{i2}\right) > 0.$$

Schools are heterogenous in their desire to inflate grades, as measured by β_j. β_j could represent pressure from parents to give higher grades or competitive pressures between schools to attract students to enroll in school j. In order to inflate a student's test grade or homework grade, the teacher must pay a cost $c_1\left(\Delta_{i1}\right)$ or $c_2\left(\Delta_{i2}\right)$, respectively. $c_1\left(\Delta_{i1}\right)$ and $c_2\left(\Delta_{i2}\right)$ are assumed to be increasing and convex. This captures the fact that it is increasingly hard for a teacher to justify the higher grade as she inflates the grade more and more.[39] The teacher chooses Δ_1 and Δ_2 to maximize the school's utility function.

We now explore properties of the model above that will be useful for estimation. For now, we assume that when the teacher chooses Δ_1 and Δ_2, she is free to pick any (positive) value that she wishes. In reality, sometimes grading a question more generously may lead to lumpy amounts of test points (e.g. either the teacher assigns 3 extra points or 0, as she may not be able to give 1 point, given the structure of the test.)

Before analyzing the teacher's decision to inflate, we illustrate what happens if $\beta_j = 0$. Then, there are no incentives to engage in any type of manipulation (neither of Type I or of Type II). Figure C1 illustrates the outcome when the distribution of student ability a_i, displayed on the x-axis, is assumed to be Uniform over [0,1] and the distribution of errors ε_i, displayed on the y-axis, is assumed to be Uniform over [-0.5,0.5]. In the Figure, a diagonal "line" distinguishes the two lower, blue fields from the two upper, green and yellow fields. Along this diagonal line, all combinations of a_i and ε_i yield the same test score, $r_i = a_i + \varepsilon_i$, which is assumed to be the required score for passing the test, $\bar{p}$. Thus, all students with (a_i, ε_i) that yield test scores that fall below $\bar{p}$ are in the light blue and dark blue regions; they

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

fail the nationwide test ($t_i = 0$). Similarly, all students with (a_i, ε_i) that yield test scores above $\bar{p}$ are in the green and yellow regions; they pass the nationwide test ($t_i = \bar{p}$).

Among the students that fail the test (i.e., those with (a_i, ε_i) in the light and dark blue areas), the subset of students with sufficiently high innate ability obtains a passing final grade in math *even though they failed the nationwide test*. Specifically, all students in the right, lower region (colored light blue) fail the nationwide test ($t_i = 0$) but obtain a passing final grade ($g_i = 1$)). In contrast, students in the left, lower region (colored dark blue) fail both the nationwide test ($t_i = 0$) and obtain a failing final grade ($g_i = 0$).

Similarly, among the students that pass the test (i.e., those with (a_i, ε_i) in the green and yellow areas), the subset with sufficiently high innate ability (yellow region) pass both the nationwide test and obtain a passing final grade, whereas students with insufficient ability (in the green region) pass the nationwide test but nonetheless obtain a failing final grade.

Our understanding of the outcome in the absence of manipulation immediately highlights that, even if we were to raise β_j from zero, the teacher would never inflate any student who obtains a final grade of Pass ($g_i = 1$) without manipulation. In Figure C1, regardless of the value of β_j, the teacher would never engage in any type of manipulation of students in the yellow and light blue regions; they obtain a passing final grade (and yield a utility of β_j to the teacher) even without the teacher engaging in any costly inflation. Now consider the case when $\beta_j > 0$:

Proposition C.1. *The teacher plays one of four actions,* $(\Delta_{i1}^*, \Delta_{i2}^*) \in \left\{(0,0), (\bar{p} - a_i - \varepsilon_i, 0), \left(0, \frac{\bar{p} - wt(a_i, \varepsilon_i, \Delta_{i1})}{(1-w)} - \alpha_i\right), \left(\bar{p} - a_i - \varepsilon_i, \frac{\bar{p} - wt(a_i, \varepsilon_i, \Delta_{i1})}{(1-w)} - \alpha_i\right)\right\}$.

Proposition C.1 states that if the teacher chooses to engage in any manipulation, she puts the student's final numeric grade exactly at $\bar{p}$, where the student (just) receives a passing final grade, $g_i = 1$. Intuitively, inflation is costly to the teacher, so she only engages in manipulation if this alters the student's final grade from fail ($g_i = 0$) to Pass ($g_i = 1$). Put differently, the teacher only engages in manipulation if it brings her an added utility of β_j. Clearly, the teacher never engages in more inflation than what puts the student's final numeric grade at $\bar{p}$.

The teacher's decision of whether to inflate a given student hinges on whether β_j, the teacher's utility from raising the final grade from Fail to Pass, (weakly) exceeds the cost of the cheapest combination of test score and homework inflation that enables the student to pass. Depending on the student's (a_i, ε_i), the cost-minimizing strategy is one of the following three strategies: (i) use only test score manipulation by raising the test grade to $\bar{p}$, $(\Delta_{i1}^*, \Delta_{i2}^*) = (\bar{p} - a_i - \varepsilon_i, 0)$; (ii) use only homework grade manipulation by inflating the homework grade such that the final grade is $\bar{p}$, $(\Delta_{i1}^*, \Delta_{i2}^*) = \left(0, \frac{\bar{p} - wt(a_i, \varepsilon_i, \Delta_{i1})}{(1-w)} - \alpha_i\right)$; or (iii)

use a combination of both types of inflation, $(\Delta^*_{i1}, \Delta^*_{i2}) = \left(\bar{p} - a_i - \varepsilon_i, \frac{\bar{p} - wt(a_i, \varepsilon_i, \Delta_{i1})}{(1-w)} - \alpha_i\right)$.

Figure C2a illustrates the teacher's strategy space when we maintain the same assumptions on the distributions of a_i and ε_i as in Figure C1, but assume that β_j takes on a strictly positive value. The teacher's strategy is $(\Delta_{i1}, \Delta_{i2}) = (0, 0)$ unless indicated otherwise. In Figure C2b, we display the corresponding distribution of raw test scores $r_i = a_i + \varepsilon_i$ in the absence of manipulation (lower subgraph) as well as the observed test score distribution after manipulation of the test scores (upper subgraph). In Figure C2a, we see that test score manipulation occurs in two regions, colored brown and orange, respectively. In both regions, the (a_i, ε_i) pairs are such that the un-manipulated test score $r_i = a_i + \varepsilon_i$ lies close to, but below, the passing threshold, $\bar{p}$ (in one of the two regions, homework grade manipulation occurs as well). In the upper subgraph of Figure C2b, we see that it is precisely the test scores in these brown and orange regions that are bunched at the passing threshold (we assume $\bar{p} = 62$). Finally, Figure C2a also indicates the regions where homework grade manipulation occurs; this manipulation is not visible in the test score distribution (Figure C2b).

Finally, Figures C2a and C2b illustrate that not all students with a given raw test score r_i close to $\bar{p}$ are inflated to $\bar{p}$. There are many different types of students who earn the same raw test score $r = a_i + \varepsilon_i$. Some students had a "bad day" when taking the nationwide test (drew a low ε), but have very high homework scores, a. These students would be able to pass the class even if they failed the nationwide test, even in the absence of manipulation. As discussed above, these students do not receive grade inflation on their test grade – and consequently $(\Delta_{i1}, \Delta_{i2}) = (0, 0)$ – although they pass the class overall. In Figure C2a, these students are located in the lower, far right, part of the area plot. Other students might have had a very "good day" when taking the nationwide test (drew a high ε). However, if their homework grade (a_i) is very low, the amount of grade inflation they would need to pass the class it too costly, and they would receive no grade inflation on their test grade. These students would fail the class overall. In Figure C2a, students that pass the nationwide test (due to "luck") but fail the class are in the upper, left corner.

Identification of β_j We now turn to the question of identification of β_j. In this context, we cannot identify β_j from the amount of excess mass at the passing test score, $\bar{p}$. To see this intuitively, again consider Figure C2a. It displays the teacher's strategy for any pair (a_i, ε_i). Thus, if we were to assume another distribution of student ability a_i, or of the student error distribution ε_i, we would obtain more mass in the bunching region in Figure C2a, *even if* β_j *were held constant*. In other words, the amount of excess mass not only varies with β_j, but also with the distributions of student ability and test taking errors. Consequently, we cannot quantify a school's leniency by the magnitude of excess mass at the Pass threshold. If we

were to use the amount of excess mass, we would risk to erroneously infer that schools have different grading leniency when, in fact, it is their student ability distributions that differ.

Instead, we theoretically show that the *lowest test score at which test score manipulation occurs* in a school identifies the school's inclination to grade leniently.

Proposition C.2. *Let $r_{j,\min}$ be the minimum raw test score which school j gets inflated to $\bar{p}$. $r_{j,\min}$ is strictly decreasing in β_j.*

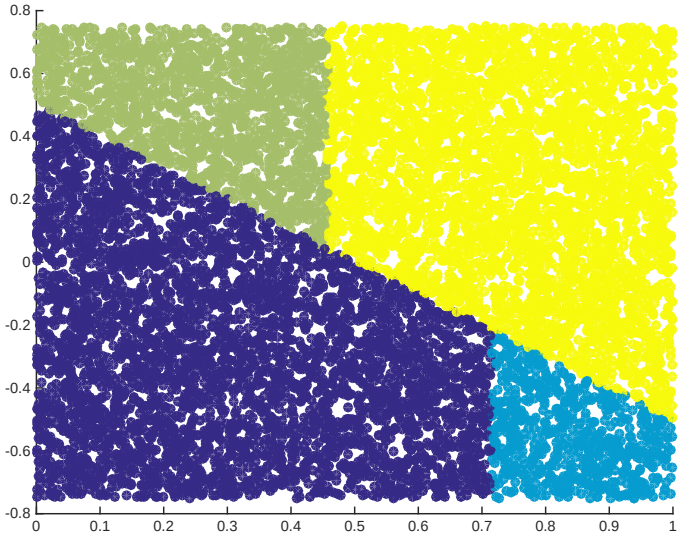
Proposition C.2 refers to to students with the minimum raw test score at which the teacher would ever inflate their test score. At this minimum raw test score which receive test grade inflation, the teacher is indifferent to inflating the test score and not. Assuming that the costs of grade inflation are the same across schools, if one school's threshold for test grade inflation is lower than that of another school, then the school with the lower threshold for grade inflation must have a higher desire to inflate grades.

To illustrate this, Figures C3a and C3b display the teacher's strategy space when we maintain the same assumptions on the distributions of a_i and ε_i as in Figure C1, but allow β_j to take on two different positive values. Figure C3a

This result will be at the heart of our estimation methods, as it implies that we can identify a school's desire to grade inflate by measuring *the minimum test score at which manipulation occurs*. This is illustrated in Figures C4a and Figure C4b. In particular, Figure C4a reproduces Figure C3a, and Figure C4b displays the corresponding distribution of raw test scores $r_i = a_i + \varepsilon_i$ in the absence of manipulation (lower subgraph) as well as the observed test score distribution after test score manipulation (upper subgraph). Both Figures C4a and C4b illustrate the lowest test score at which manipulation occurs. Figure C4b illustrates that, to quantify grading leniency, we must estimate, from the manipulated test score distribution, the lowest test score at which manipulation takes place.

C.2 Figures

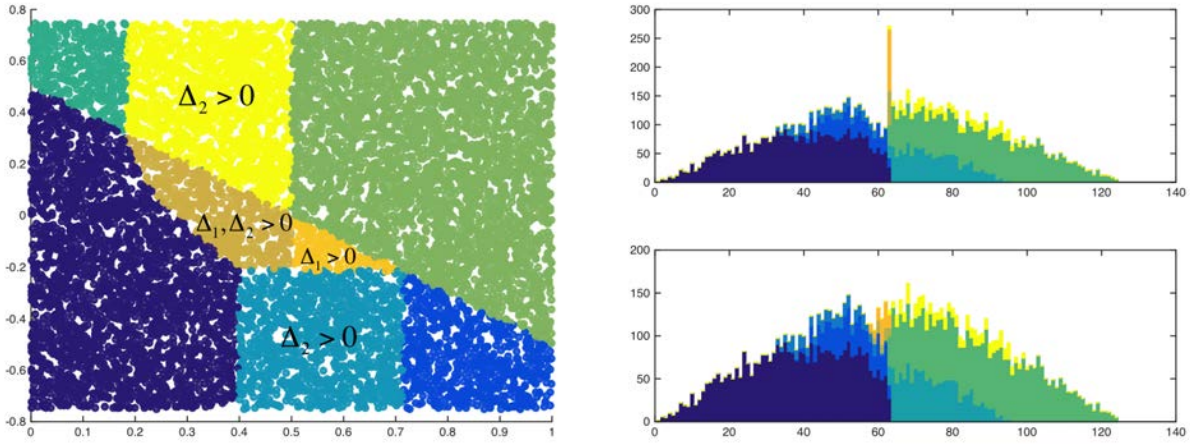
Figure C1: Student outcomes in the absence of manipulation



Note: The figure displays the teacher's strategy when $\beta_j = 0$, i.e., in the absence of any manipulation. The student ability distribution, displayed on the x-axis is assumed to be Uniform over [0,1] and the student error distribution, displayed on the y-axis, is assumed to be Uniform over [-0.5,0.5]. A diagonal "line" distinguishes the two lower, blue fields from the two upper, green and yellow fields. Along this diagonal line, all combinations of a_i and ε_i yield the same test score, $r_i = a_i + \varepsilon_i$, which is assumed to be the passing score. Thus, all students with (a_i, ε_i) that yield test scores that fall below the Pass threshold are in the light blue and dark blue regions; they fail the nationwide test. Similarly, all students with (a_i, ε_i) that yield test scores above the Pass threshold are in the green and yellow regions; they pass the nationwide test. Among students that fail the nationwide tests (i.e., those with (a_i, ε_i) in the light blue and dark blue areas), students with sufficiently high innate ability will obtain a passing final grade in math *even though they failed the nationwide test.* Specifically, all students in the right, lower region (colored light blue) will fail the nationwide test but obtain a passing final grade, whereas students in the left, lower region (colored dark blue) will fail both the nationwide test and obtain a failing final grade. Similarly, among students that pass the nationwide tests (i.e., those with (a_i, ε_i) in the green and yellow areas), students with sufficiently high innate ability (yellow region) will pass both the nationwide test and obtain a passing final grade, whereas students with insufficient ability (in the green region) will pass the nationwide test but nonetheless obtain a failing final grade.

Figure C2: Student outcomes and the teacher's strategy in the presence of manipulation

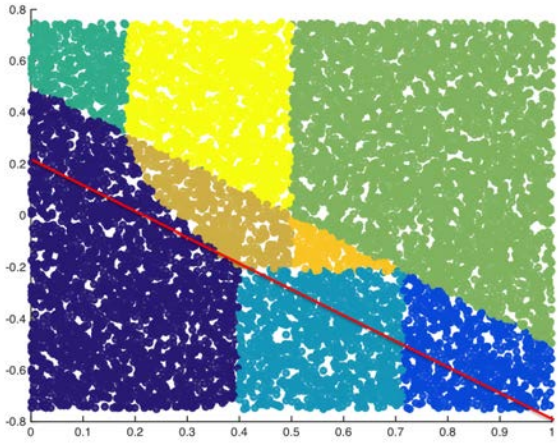
(a) Outcomes for various (a_i, ε_i) (b) Distribution of $r_i = a_i + \varepsilon_i$



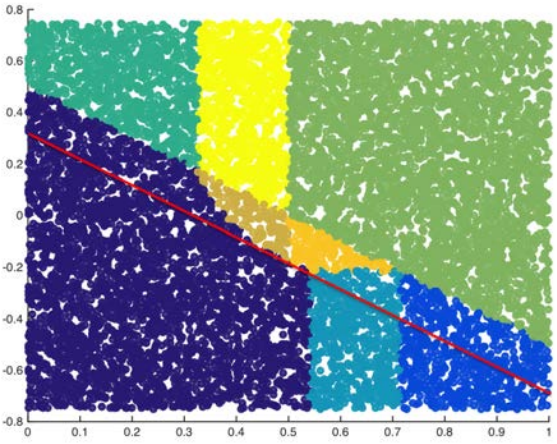
Note: In Figure C2a, we display student outcomes and the teacher's strategy for all possible pairs (a_i, ε_i). The teacher's strategy is $(\Delta_{i1}, \Delta_{i2}) = (0, 0)$ unless indicated otherwise. In Figure C2b, we display the distribution of raw test scores $r_i = a_i + \varepsilon_i$ in the absence of manipulation (lower subgraph) as well as the observed test score distribution after manipulation of the test scores (upper subgraph). As in Figure C1, we assume that the student ability distribution is Uniform over [0,1] and that the student error distribution is Uniform over [-0.5,0.5]. In Figure C2a, we see that test score manipulation occurs in two regions, colored brown and orange, respectively. In both regions, the (a_i, ε_i) pairs are such that the un-manipulated test score $r_i = a_i + \varepsilon_i$ lies close to, but below, the passing threshold, $\bar{p}$ (in one of the two regions, homework grade manipulation occurs as well). In the upper subgraph of Figure C2b, we see that it is precisely the test scores in these brown and orange regions that are bunched at the passing threshold (we assume $\bar{p} = 62$). Finally, Figure C2a also indicates the regions where homework grade manipulation occurs; this manipulation is not visible in the test score distribution (Figure C2b).

Figure C3: Student outcomes for different levels of grading leniency, β_j

(a) Outcomes for various (a_i, ε_i), higher β_j



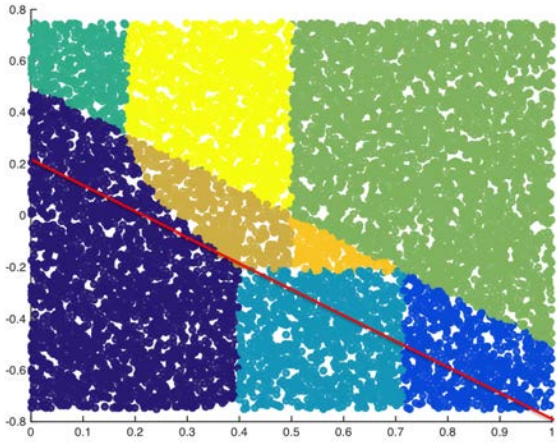
(b) Outcomes for various (a_i, ε_i), lower β_j



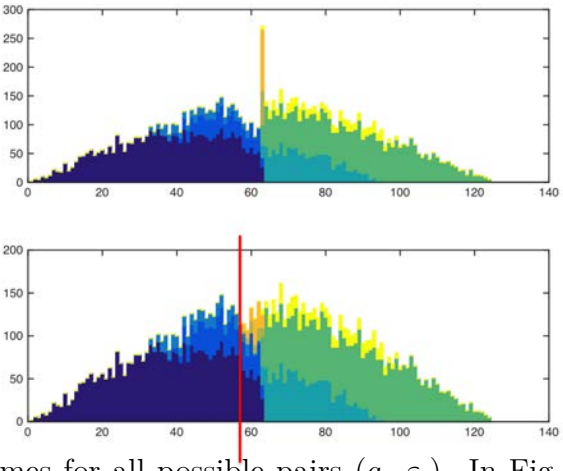
Note: The Figures display student outcomes and the teacher's strategy for all possible pairs (a_i, ε_i), for two different levels of grading leniency, β_j. All other assumptions are as in Figure C1. We observe that the lowest test score where manipulation occurs is lower in Figure C3a than in Figure C3b, consistent with the fact that teachers grade more leniently in Figure C3a.

Figure C4: Identifying grading leniency, β_j, from the empirical test score distribution

(a) Outcomes for various (a_i, ε_i)



(b) Distribution of $r_i = a_i + \varepsilon_i$



Note: In Figure C4a, we display student outcomes for all possible pairs (a_i, ε_i). In Figure C4b, we display the distribution of raw test scores $r_i = a_i + \varepsilon_i$ in the absence of manipulation (lower subgraph) as well as the observed test score distribution after manipulation of the test scores (upper subgraph). All assumptions are as in Figure C1. Both Figures illustrate the lowest test score at which manipulation occurs. Figure C4b illustrates that, to quantify grading leniency, we must estimate, from the manipulated test score distribution, the lowest test score at which manipulation takes place.

C.3 Mathematical proofs

C.3.1 Proof of Proposition A1

We derive results under a general cost function $c(\Delta_1, \Delta_2)$, which is increasing and convex in each argument.

- If inflate test grade, then inflate to exactly $\bar{p}$. $\Delta_1 = \bar{p} - a_i - \varepsilon_i$, otherwise, $\Delta_1 = 0$.
- If inflate final grade, inflate exactly to $\bar{p} : \Delta_2 = \frac{\bar{p} - wt(a_i, \varepsilon_i, \Delta_1)}{(1-w)} - \alpha_i$, otherwise $\Delta_2 = 0$.

Thus, possible strategies are:
$(\Delta_1^*, \Delta_2^*) \in \left\{(0,0), (\bar{p} - a_i - \varepsilon_i, 0), \left(0, \frac{\bar{p} - wt(a_i, \varepsilon_i, \Delta_1)}{(1-w)} - \alpha_i\right), \left(\bar{p} - a_i - \varepsilon_i, \frac{\bar{p} - wt(a_i, \varepsilon_i, \Delta_1)}{(1-w)} - \alpha_i\right)\right\}$.
This defines a set of inequalities where each strategy is optimal. First partition the (a_i, ε_i) space where the student passes the test and/or class without the help of inflation. These are:

1. $t(a_i, \varepsilon_i, 0) = \bar{p}$ & $g(a_i, \varepsilon_i, 0, 0) = 1$. In other words, $a_i + \varepsilon_i \geq \bar{p}$, $a_i \geq \bar{p}$
2. If $t(a_i, \varepsilon_i, 0) = 0$ & $g(a_i, \varepsilon_i, 0, 0) = 1$. In other words, $a_i + \varepsilon_i < \bar{p}$, $a_i(1-w) \geq \bar{p}$
3. If $t(a_i, \varepsilon_i, 0) = \bar{p}$ & $g(a_i, \varepsilon_i, 0, 0) = 0$. In other words, $a_i + \varepsilon_i \geq \bar{p}$, $a_i < \bar{p}$
4. If $t(a_i, \varepsilon_i, 0) = 0$ & $g(a_i, \varepsilon_i, 0, 0) = 0$. In other words, $a_i + \varepsilon_i < \bar{p}$, $a_i(1-w) < \bar{p}$

We now partition the space into where each possible strategy is optimal:

1. $u(0,0) \leq u\left(0, \frac{\bar{p} - wt(a_i, \varepsilon_i, \Delta_1)}{(1-w)} - \alpha_i\right)$

   (a) If $t(a_i, \varepsilon_i, 0) = \bar{p}$ & $g(a_i, \varepsilon_i, 0, 0) = 1$. In other words, $a_i + \varepsilon_i \geq \bar{p}$, $a_i \geq \bar{p}$:

   $$c\left(0, \frac{\bar{p} - wt(a_i, \varepsilon_i, \Delta_1)}{(1-w)} - \alpha_i\right) \leq 0$$

   This region does not exist.

   (b) If $t(a_i, \varepsilon_i, 0) = 0$ & $g(a_i, \varepsilon_i, 0, 0) = 1$. In other words, $a_i + \varepsilon_i < \bar{p}$, $a_i(1-w) \geq \bar{p}$:

   $$c\left(0, \frac{\bar{p} - wt(a_i, \varepsilon_i, \Delta_1)}{(1-w)} - \alpha_i\right) \leq 0$$

   This region does not exist.

   (c) If $t(a_i, \varepsilon_i, 0) = \bar{p}$ & $g(a_i, \varepsilon_i, 0, 0) = 0$. In other words, $a_i + \varepsilon_i \geq \bar{p}$, $a_i < \bar{p}$:

   $$\beta \geq c(0, \bar{p} - a_i).$$

(d) If $t(a_i, \varepsilon_i, 0) = 0$ & $g(a_i, \varepsilon_i, 0, 0) = 0$. In other words, $a_i + \varepsilon_i < \bar{p}$, $a_i(1-w) < \bar{p}$:

$$\beta \geq c\left(0, \frac{\bar{p}}{(1-w)} - a_i\right)$$

2. $u(0,0) \leq u(\bar{p} - a_i - \varepsilon_i, 0)$.

(a) If $t(a_i, \varepsilon_i, 0) = \bar{p}$ & $g(a_i, \varepsilon_i, 0, 0) = 1$. In other words, $a_i + \varepsilon_i \geq \bar{p}$, $a_i \geq \bar{p}$:

$$c(\bar{p} - a_i - \varepsilon_i, 0) \leq 0$$

This region does not exist.

(b) If $t(a_i, \varepsilon_i, 0) = 0$ & $g(a_i, \varepsilon_i, 0, 0) = 1$. In other words, $a_i + \varepsilon_i < \bar{p}$, $a_i(1-w) \geq \bar{p}$:

$$c(\bar{p} - a_i - \varepsilon_i, 0) \leq 0$$

This region does not exist.

(c) If $t(a_i, \varepsilon_i, 0) = \bar{p}$ & $g(a_i, \varepsilon_i, 0, 0) = 0$. In other words, $a_i + \varepsilon_i \geq \bar{p}$, $a_i < \bar{p}$:

$$c(\bar{p} - a_i - \varepsilon_i, 0) \leq 0$$

This region does not exist.

(d) If $t(a_i, \varepsilon_i, 0) = 0$ & $g(a_i, \varepsilon_i, 0, 0) = 0$. In other words, $a_i + \varepsilon_i < \bar{p}$, $a_i(1-w) < \bar{p}$:

$$\beta \geq c(\bar{p} - a_i - \varepsilon_i, 0) \text{ if } a_i \geq \bar{p}.$$

$$0 \geq c(\bar{p} - a_i - \varepsilon_i, 0) \text{ if } a_i < \bar{p},$$

This region where $a_i < \bar{p}$ does not exist.

3. $u(0,0) \leq \left(\bar{p} - a_i - \varepsilon_i, \frac{\bar{p} - wt(a_i, \varepsilon_i, \Delta_1)}{(1-w)} - \alpha_i\right)$

(a) If $t(a_i, \varepsilon_i, 0) = \bar{p}$ & $g(a_i, \varepsilon_i, 0, 0) = 1$. In other words, $a_i + \varepsilon_i \geq \bar{p}$, $a_i \geq \bar{p}$:

$$0 \geq c\left(\bar{p} - a_i - \varepsilon_i, \frac{\bar{p} - wt(a_i, \varepsilon_i, \Delta_1)}{(1-w)} - \alpha_i\right).$$

This region does not exist.

(b) If $t(a_i, \varepsilon_i, 0) = 0$ & $g(a_i, \varepsilon_i, 0, 0) = 1$. In other words, $a_i + \varepsilon_i < \bar{p}$, $a_i(1-w) \geq \bar{p}$:

$$0 \geq c\left(\bar{p} - a_i - \varepsilon_i, \frac{\bar{p} - wt(a_i, \varepsilon_i, \Delta_1)}{(1-w)} - \alpha_i\right).$$

This region does not exist.

(c) If $t(a_i, \varepsilon_i, 0) = \bar{p}$ & $g(a_i, \varepsilon_i, 0, 0) = 0$. In other words, $a_i + \varepsilon_i \geq \bar{p}$, $a_i < \bar{p}$:

$$\beta \geq c(\bar{p} - a_i - \varepsilon_i, \bar{p} - \alpha_i).$$

(d) If $t(a_i, \varepsilon_i, 0) = 0$ & $g(a_i, \varepsilon_i, 0, 0) = 0$. In other words, $a_i + \varepsilon_i < \bar{p}$, $a_i(1-w) < \bar{p}$:

$$\beta \geq c(\bar{p} - a_i - \varepsilon_i, \bar{p} - \alpha_i).$$

4. $u\left(0, \frac{\bar{p} - wt(a_i, \varepsilon_i, 0)}{(1-w)} - a_i\right) \leq u(\bar{p} - a_i - \varepsilon_i, 0)$

(a) Region where $u\left(0, \frac{\bar{p} - wt(a_i, \varepsilon_i, \Delta_1)}{(1-w)} - a_i\right)$ beats no inflation, and $t\left(0, \frac{\bar{p} - wt(a_i, \varepsilon_i, \Delta_1)}{(1-w)} - a_i\right) = p, g\left(a_i, \varepsilon_i, 0, \frac{\bar{p} - wt(a_i, \varepsilon_i, \Delta_1)}{(1-w)} - a_i\right) = 1$. In other words: $a_i + \varepsilon_i \geq \bar{p}$, $a_i < \bar{p}, \beta \geq c(0, \bar{p} - a_i)$. This regions does not intersect with any regions where $(\bar{p} - a_i - \varepsilon_i, 0)$ was preferred to $(0, 0)$. Thus, this region does not exist.

(b) Region where $u\left(0, \frac{\bar{p} - wt(a_i, \varepsilon_i, \Delta_1)}{(1-w)} - a_i\right)$ beats no inflation, and $t\left(0, \frac{\bar{p} - wt(a_i, \varepsilon_i, \Delta_1)}{(1-w)} - a_i\right) = 0, g\left(a_i, \varepsilon_i, 0, \frac{\bar{p} - wt(a_i, \varepsilon_i, \Delta_1)}{(1-w)} - a_i\right) = 1$. $a_i + \varepsilon_i < \bar{p}$, $a_i(1-w) < \bar{p}, \beta \geq c\left(0, \frac{\bar{p}}{(1-w)} - a_i\right)$.

$$c\left(0, \frac{\bar{p}}{(1-w)} - a_i\right) \geq c(\bar{p} - a_i - \varepsilon_i, 0).$$

5. $u(\bar{p} - a_i - \varepsilon_i, 0) \leq u\left(\bar{p} - a_i - \varepsilon_i, \frac{\bar{p} - wt(a_i, \varepsilon_i, \Delta_1)}{(1-w)} - \alpha_i\right)$. This could only be the case if $u(\bar{p} - a_i - \varepsilon_i, 0)$ resulted in a failing grade.

$$a \leq \bar{p}$$

6. $u\left(0, \frac{\bar{p} - wt(a_i, \varepsilon_i, 0)}{(1-w)} - \alpha_i\right) \leq u\left(\bar{p} - a_i - \varepsilon_i, \frac{\bar{p} - wt(a_i, \varepsilon_i, \Delta_1)}{(1-w)} - \alpha_i\right)$.

(a) Region where $\left(0, \frac{\bar{p} - wt(a_i, \varepsilon_i, 0)}{(1-w)} - \alpha_i\right)$ beats no inflation and $t\left(0, \frac{\bar{p} - wt(a_i, \varepsilon_i, \Delta_1)}{(1-w)} - \alpha_i\right) = 0$. That is: $a_i + \varepsilon_i < \bar{p}$, $a_i(1-w) < \bar{p}, \beta \geq c\left(0, \frac{\bar{p}}{(1-w)} - a_i\right)$.

$$\beta \geq c\left(0, \frac{\bar{p}}{(1-w)} - a_i\right) \geq c(\bar{p} - a_i - \varepsilon_i, \bar{p} - a_i).$$

Combining the inequalities above:

- The region where where the teacher inflates both the test and final grade, $u\left(\bar{p}-a_i-\varepsilon_i, \frac{\bar{p}-wt(a_i,\varepsilon_i,\Delta_1)}{(1-w)}-\alpha_i\right)$ is played is where:

$$\begin{aligned} \beta &\geq c\left(\bar{p}-a_i-\varepsilon_i, \bar{p}-a_i\right), \\ a_i+\varepsilon_i &\leq \bar{p}, \\ a_i &\leq \bar{p}, \\ \beta &\geq c\left(0, \frac{\bar{p}}{(1-w)}-a_i\right) \geq c\left(\bar{p}-a_i-\varepsilon_i, \bar{p}-a_i\right). \end{aligned}$$

- The region where the teacher only inflates the test grade, $u\left(\bar{p}-a_i-\varepsilon_i, 0\right)$, is played when:

$$\begin{aligned} a_i+\varepsilon_i &< \bar{p}, \\ a_i\left(1-w\right) &< \bar{p}, \\ a &\geq \bar{p} \\ \beta &\geq c\left(\bar{p}-a_i-\varepsilon_i, 0\right), \\ c\left(0, \frac{\bar{p}}{(1-w)}-a_i\right) &\geq c\left(\bar{p}-a_i-\varepsilon_i, 0\right) \end{aligned}$$

- The region where the teacher only inflates the final grade, $u\left(0, \frac{\bar{p}-wt(a_i,\varepsilon_i,\Delta_1)}{(1-w)}-\alpha_i\right)$, is played when either:

  1. Students who natually pass the test:

$$\begin{aligned} a_i+\varepsilon_i &\geq \bar{p}, \\ a_i &\leq \bar{p}, \\ \beta &\geq c\left(0, \bar{p}-a_i\right) \end{aligned}$$

  2. Students who natually fail the test:

$$\begin{aligned} a_i+\varepsilon_i &\leq \bar{p}, \\ a_i\left(1-w\right) &\leq \bar{p}, \\ \beta &\geq c\left(0, \frac{\bar{p}}{1-w}-a_i\right), \\ c\left(0, \frac{\bar{p}}{1-w}-a_i\right) &\geq c\left(\bar{p}-a_i-\varepsilon_i, \bar{p}-a_i\right) \end{aligned}$$

C.3.2 Proof of Proposition A2

First, consider the region where $u\left(\bar{p}-a_i-\varepsilon_i,0\right)$ is optimal. Define $c_2^{-1}\left(b,\Delta_2\right)$ as the inverse type 1 cost function where b is total cost and Δ_2 is amount of type 2 inflation. The minimum r within this region is:

$$\begin{aligned} r_{1,\min} &= \bar{p}-c_2^{-1}\left(\beta,0\right) \text{ if } c\left(0,\frac{w\bar{p}}{1-w}\right)\geq c\left(c_2^{-1}\left(\beta,0\right),0\right) \\ &= \bar{p}-c_2^{-1}\left(c\left(0,\frac{w\bar{p}}{1-w}\right),0\right) \text{ otherwise.} \end{aligned}$$

Now consider the region where $\left(\bar{p}-a_i-\varepsilon_i,\bar{p}-\alpha_i\right)$ is optimal. The minimum r within this region is:

$$r_{2,\min}=\bar{p}-c_2^{-1}\left(\beta,0\right) \text{ if } c\left(0,\frac{w\bar{p}}{1-w}\right)\geq c\left(c_2^{-1}\left(\beta,0\right),0\right).$$

If $c\left(0,\frac{w\bar{p}}{1-w}\right)<c\left(c_2^{-1}\left(\beta,0\right),0\right)$, then

$$c\left(0,\frac{\bar{p}}{(1-w)}-a_i\right)= c\left(\bar{p}-r,\bar{p}-a\right)=\beta.$$

Implicity define $a^*\left(r\right)$ as the function which satisfies:

$$c\left(0,\frac{\bar{p}}{(1-w)}-a^*\left(r\right)\right)= c\left(\bar{p}-r,\bar{p}-a^*\left(r\right)\right).$$

Implicitly differentiating the expression above and rearranging, we get:

$$\frac{da}{dr}=\frac{c_1\left(\bar{p}-r,p-a\right)}{c_2\left(0,\frac{\bar{p}}{1-w}-a\right)-c_2\left(\bar{p}-r,p-a\right)}.$$

Now, implicitly differentiate $c\left(\bar{p}-r,\bar{p}-a\right)=\beta$ with respect to β :

$$\begin{aligned} \frac{dr}{d\beta} &= \frac{-1}{c_1\left(\bar{p}-r,p-a\right)+c_2\left(\bar{p}-r,p-a\right)*\frac{da}{dr}} \\ &= \frac{c_2\left(\bar{p}-r,p-a\right)-c_2\left(0,\frac{\bar{p}}{1-w}-a\right)}{c_1\left(\bar{p}-r,p-a\right)*c_2\left(0,\frac{\bar{p}}{1-w}-a\right)} \end{aligned}$$

Since the cost function is weakly increasing, the denominator is negative. For the numerator

to be negative, we need: $c_2\left(0, \frac{\bar{p}}{1-w} - a\right) \geq c_2\left(\bar{p} - r, p - a\right)$. Specifically, we need:

$$c_2\left(0, \frac{\bar{p}}{1-w} - a\right) \geq c_2\left(c_1^{-1}\left(c\left(0, \frac{\bar{p}}{1-w} - a\right), p - a\right), p - a\right).$$

If we assumed the cross partial of $c\left(\Delta_1, \Delta_2\right)$ to be zero, then $c_2\left(\bar{p} - r, p - a\right) = c_2\left(0, p - a\right) \leq c_2\left(0, \frac{\bar{p}}{1-w} - a\right)$ because $c\left(\Delta_1, \Delta_2\right)$ is convex and $\frac{\bar{p}}{1-w} - a > p - a$. Thus, the minimun inflated testscore is strictly decreasing in β, the payoff from grade inflation.

D Estimation Details

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

We estimate the model using constrained nonlinear-least squares and use k-fold (k=5) cross-validation to prevent overfitting. We perform the following procedure for each region-year:[40]

For a given guess of the width of the two manipulation regions in the region-year, $(\beta_{1jt}, \beta_{2jt})$, and of the order of the polynomials that determine the deviation from log-concavity due to manipulation, $(p_{kjt}^{high}, p_{kjt}^{low})$, 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. 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 a given guess of the widths of the manipulation regions and the orders of the polynomials. We then iterate this procedure using a grid search over all possible combinations of manipulation region widths from 0 to 7 and polynomial orders from 0 to 4.

We select the set of $(\beta_{1jt}, \beta_{2jt})$ and $(p_{kjt}^{high}, p_{kjt}^{low})$ 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 deviation from log-concavity randomly at each test score for reasons other than manipulation.

Once we have selected the out-of-sample MSE minimizing combination of $(\beta_{1jt}, \beta_{2jt})$ and $(p_{kjt}^{high}, p_{kjt}^{low})$, we pool all the data back together and estimate the parameters of the

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

manipulated and un-manipulated distributions, $(\theta_{jt}^{high,1}, \theta_{jt}^{low,2}, \theta_{jt}^{high,2}, \delta_{1jt-1}, ..., \delta_{R^{max}jt-1})$.

D.2 Estimating the causal impact of test score manipulation

The "first stage." First, we identify the impact of test score manipulation on students' final grades in math. This can be thought of as the first stage regression of our Wald estimation. To do this, we proceed in two steps.

Recall that g_{ijt} is student $i's$ observed final math grade (who is enrolled in region j in year t). We estimate:

$$g_{ijt} = \hat{g}_{kjt}\left(r_{ijt}, \theta_{kjt}^{grade}\right) + \alpha_{kjt} * (r_{ijt} \geq k) + \epsilon_{ijt}^{g}, \tag{22}$$

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

$$\text{and}$$

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

$\hat{g}_{kjt}\left(r_{ijt}, \theta_{kjt}^{grade}\right)$ 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 final grades. $(r_{ijt} < k - \beta_{kjt}$ or $r_{ijt} > k + \beta_{kjt} - 1)$ ensures that the data used to estimate equation (13) 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 final grade 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 final math grade for students within the manipulation region of the test score distribution *had there been no test score manipulation*:

$$\bar{g}_{jt}(k) = \int_{k-\beta_{kjt}}^{k+\beta_{kjt}-1} \left[\hat{g}_{kjt}\left(r, \hat{\theta}_{kjt}^{grade}\right) + \hat{\alpha}_{kjt} * (r \geq k)\right] * \frac{\hat{h}_{jt}(r)}{\int_{k-\beta_{kjt}}^{k+\beta_{kjt}-1} \hat{h}_{jt}(r)dr} dr. \tag{23}$$

For students inside the manipulation region, we now compare the estimated counterfactual average grade, had there been no test score manipulation, calculated in (23), with the actual average final math grade 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 final grade due to the student having a raw test score that falls within the manipulated region of the test score distribution:

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

where N_{jt}^{manip} is the number of students in the manipulation region around threshold k in region j in year t. Figure 6 illustrates this first stage estimate in the context of the example used in Figure 5.

The "reduced form" and LATE estimates. The procedure above can be repeated with a different outcome variable, such as income at age 23, to identify the reduced-form effect of falling into the manipulation region on future income. The ratio of this reduced-form effect to the first-stage effect, in turn, identifies the local average treatment effect (LATE) of receiving an inflated final math grade on future income. We block bootstrap the entire procedure (including the estimation of the manipulation width and the shape of the un-manipulated distribution) to calculate standard errors, sampling at the region by year level. This is the same level at which we estimated the widths of the manipulation regions.