In a Small Moment:

Class Size and Moral Hazard in the Italian Mezzogiorno*

Joshua D. Angrist

MIT, IZA and NBER

Erich Battistin

Queen Mary University of London, CEPR, IZA and IRVAPP

Daniela Vuri

University of Rome Tor Vergata, CEIS, CESifo and IZA

November 2016

Abstract

Instrumental variables (IV) estimates show strong class size effects in Southern Italy. But Italy's Mezzogiorno is distinguished by manipulation of standardized test scores as well as by economic disadvantage. IV estimates suggest small classes increase manipulation. We argue that score manipulation is a consequence of teacher shirking. IV estimates of a causal model for achievement as function of class size and score manipulation show that class size effects on measured achievement are driven entirely by the relationship between class size and manipulation. These results show how consequential score manipulation can arise even in assessment systems with few accountability concerns.

*Special thanks go to Patrizia Falzetti, Roberto Ricci and Paolo Sestito at INVALSI for providing the achievement data used here and to INVALSI staffers Paola Giangiacomo and Valeria Tortora for advice and guidance in our work with these data. Grateful thanks also go to Gianna Barbieri, Angela Iadecola, and Daniela Di Ascenzo at the Ministry of Education (MIUR) for access to and assistance with administrative schools data. Chiara Perricone provided expert research assistance. Our thanks to David Autor, Daniele Checchi, Eric Hanushek, Andrea Ichino, Brian Jacob, Michael Lechner, Steve Machin, Fabrizio Mattesini, Derek Neal, Parag Pathak, Daniele Paserman and Jonah Rockoff for helpful discussions and comments, and to seminar participants at NBER Education Fall 2013, the 2014 SOLE meeting, the University of California Irvine, Norwegian School of Economics (NHH), Padova University, IRVAPP, EUI, UCL, ISER (Essex), the CEP Labour Market Workshop, the Warwick 2014 CAGE conference, the 2014 Laax Labor Economics Workshop, the University of Rome Tor Vergata, EIEF, CEPR/IZA European Summer Symposium in Labour Economics, Tinbergen Institute, Oxford and George Washington for helpful comments. This research is supported by the Einaudi Institute of Economics and Finance (EIEF) - Research Grant 2011 and by the Fondazione Bruno Kessler. Angrist thanks the Institute for Education Sciences, the Arnold Foundation, and The Spencer Foundation for financial support. The views expressed here are those of the authors alone.

1 Introduction

School improvement efforts often focus on inputs to education production, the most important of which is staffing ratios. Parents, teachers, and policy makers look to small classes to boost learning. The question of whether changes in class size have a causal effect on achievement remains controversial, however. Regression estimates often show little gain to class size reductions, with students in larger classes sometimes appearing to do better (Hanushek, 1995). At the same time, a large randomized study, the Tennessee STAR experiment generated evidence of substantial learning gains in smaller classes (Krueger, 1999). An investigation of longer-term effects of the STAR experiment also suggests small classes increased college attendance (Chetty et al., 2011).

Standardized tests provide the yardstick by which school quality is most often assessed and compared. As testing regimes have proliferated, however, so have concerns about the reliability and fidelity of assessment results (Neal, 2013, lays out the issues in this context). Evidence on this point comes from Jacob and Levitt (2003), who documented substantial cheating on standardized tests in Chicago public schools, while a recent system-wide cheating scandal in Atlanta sent some school administrators and teachers to jail (Severson, 2011). Of course, students may cheat as well, especially on tests that have consequences for them. In many cases, however, the behavior of staff who administer and (sometimes) grade assessments is of primary concern. For example, Dee et al. (2016) show that New York's Regents exam scores are very likely manipulated by the school staff who grade them. Concerns regarding score manipulation have also been raised in discussions of Sweden's school choice reform (Böhlmark and Lindahl, 2012, Diamond and Persson, 2016) and in the United Kingdom and Israel, where important nationally administered tests are locally marked. In public school systems with few or no employee performance standards, such as the Italian public school system studied here, fidelity of school staff to test administration protocols may be especially

weak.[1]

Our investigation of the effect of score manipulation on the measurement of education production in Italy begins by applying the quasi-experimental research design introduced by Angrist and Lavy (1999). This design exploits variation in class size induced by rules stipulating a class size cutoff. In Israel, with a cutoff of 40, we expect to see a single class of size 40 in a grade cohort of 40, while with enrollment of 41, the cohort is typically split into two much smaller classes. Angrist and Lavy called this Maimonides' Rule, after the medieval scholar and sage Moses Maimonides, who commented on a similar rule in the Talmud. Maimonides-style instrumental variables (IV) estimates of the effects of class size on achievement for the population of Italian second and fifth graders, most of whom attend much smaller classes than those seen in Israel, suggest a statistically significant though modest return to decreases in class size. Importantly, however, our estimated returns to class reductions in Southern Italy are roughly three times larger than in the rest of the country.[2]

Why is there a large return to small classes in Southern Italy but not in the North? Differences in class size effects on learning may explain this, of course. Southern Italy is poorer and the returns to class size may be inversely related to family income, for example. Among other important distinctions, however, the Italian Mezzogiorno is characterized by widespread score manipulation on the standardized tests given in primary schools. This can be seen in Figure 1, which reproduces provincial estimates of score manipulation from the Italian Instituto Nazionale per la Valutazione del Sistema dell'Istruzione (INVALSI), a government agency charged with educational assessment. Classes in which scores are likely to have been manipulated are identified through a statistical model that looks for surprisingly high average scores, low within-class variability, and implausible missing data patterns.[3] Measured in this

[1]Local teachers grade the UK's Key Stage 1 assessments (given in year 2, usually at age 7). Key Stage 2 assessments given at the end of elementary school (usually at age 11) are locally proctored, with unannounced external visits, and are externally graded (documents and links at http://www.education.gov.uk/sta/assessment). See Battistin and Neri (2016) for evidence on UK manipulation. Lavy (2008) documents gender bias in the local grading of Israel's matriculation exams. De Paola et al. (2014) estimate the effects of workplace accountability on productivity in the Italian public sector. Ichino and Tabellini (2014) discuss possible benefits from organizational reform and increased choice in Italian public schools.

[2]The South, also known as Mezzogiorno, consists of the administrative regions of Basilicata, Campania, Calabria, Puglia, Abruzzo, Molise, and the islands of Sicily and Sardinia. Italy's 20 Administrative regions are further divided into over 100 provinces.

[3]The INVALSI testing program is described below and in INVALSI (2010). The INVALSI score manipulation variable identifies classes with substantially anomalous score distributions, imputing a probability of manipulation for each (see Quintano et al., 2009). Figure 1 uses this variable for the 2009-2011 scores of

way, roughly 5 percent of Italian scores are compromised, about the same rate reported for Chicago elementary schools by Jacob and Levitt (2003). In Southern Italy, however, the proportion of compromised exams averages about 14 percent (see Table 1) and reaches 25 percent in some provinces. Further evidence suggesting extensive score manipulation on the South comes from Bertoni et al. (2013), who analyze data generated by the random assignment of external monitors sent to observe test administration.

The purpose of this paper is to document and explain the effects of class size on score manipulation, with a special focus on how manipulation distorts estimates of class size effects on learning. IV estimates show that large classes reduce manipulation, especially in the South. We argue that manipulation of INVALSI scores reflects teacher behavior - specifically, dishonest transcription of hand-written answer sheets onto machine-readable score report forms. Dishonest score reporting appears to be largely a form of shirking, that is, moral hazard in grading effort, rather than cheating motivated by accountability concerns. The theoretical and institutional case for a link between teacher shirking in score transcription and class size is made with the aid of a simple model of teacher behavior. A likely factor in this model is the social constraint imposed by peers: just as randomly assigned monitors inhibit manipulation, score sheets for larger classes are likely to be transcribed by a team of teachers rather than only one.

Motivated by empirical and theoretical results linking class size and external monitoring with score manipulation, we develop an empirical model for achievement as a function of two endogenous variables, class size and score manipulation. The model is identified by a combination of Maimonides' Rule and random assignment of external monitors. The resulting estimates suggest that the relationship between class size and INVALSI test scores is explained entirely by score manipulation: class size is unrelated to student learning in Italy, at least insofar as learning is measured by standardized tests.

The fact that score manipulation explains class size effects in Italy should be of interest to policy makers and to researchers studying the causal effects of school inputs. The Maimonides' Rule research design is not guaranteed to work. Urquiola and Verhoogen (2009) show how systematic sorting induces selection bias in comparisons across class size caps in Chilean private schools. By contrast, our analysis uncovers a new substantive problem in-

second and fifth graders.

herent in analyses of the causal effects of class size, a problem that arises independently of the research design. Class size has a causal effect on *measured achievement,* but these measurements are compromised. Even when the research design is uncompromised, statistically significant and credibly identified class size effects need not signal increased learning in smaller classes.

Our behavioral model suggests class size can affect manipulation in any setting where exams are marked with discretion. The findings reported here also provide evidence of a previously unrecognized source of moral hazard in school assessments. In contrast with teacher and administrator cheating in response to high-stakes testing, the manipulation problem uncovered here emerges in a low-stakes assessment program meant to guide national education policy rather than through specific school and personnel decisions. Italian teachers work in a highly regulated public sector, with little risk of termination, and are subject to a pay and promotion structure largely independent of their performance. Although employees might not like to be seen by their colleagues as slouches or free riders, regulation and employment protection make formal disciplinary actions costly and unlikely. Manipulation appears to arise in the Mezzogiorno in part because worker performance standards are weak*;* in fact, it seems fair to say that moral hazard arises here from diminished rather than excessive accountability pressures. Finally, it bears emphasizing that concerns with teacher shirking are not unique to Italy. For example, Clotfelter et al. (2009) discusses distributional and other consequences of American teacher absenteeism, while teacher absenteeism and other forms of public sector shirking are a perennial concern in developing countries (see, e.g., Banerjee and Duflo, 2006, and Chaudhury et al., 2006).

The rest of the paper is organized as follows. The next section presents institutional background on Italian schools and tests. Section 3 describes our data and documents the Maimonides' Rule first stage. Following a brief graphical analysis, Section 4 reports Maimonides-style estimates of effects of class size on achievement and score manipulation. Section 5 explores the nature of score manipulation by linking score distributions and response patterns with class size and item difficulty. Finally, Section 6 uses the monitoring experiment and Maimonides' Rule to jointly estimated class size and manipulation effects. This section also reviews possible threats to validity in our research design. Section 7 concludes.

2 Background and Context

Italian Schools and Tests

Primary schooling (*scuola elementare*) in Italy is compulsory from ages 6 to 11. Schools are administrated as single- or multi-unit institutions, a distinction that's important to us because some of the instrumental variables used below are defined at the school level and some are defined at the institution level. Families apply for school admission in February, well before the beginning of the new academic year in September. Parents or legal guardians typically apply to a school in their province, located near their homes. In (rare) cases of over-subscription, distance usually determines who has a first claim on seats. Rejected applicants are assigned other schools, mostly nearby. School principals group students into classes and assign teachers over the summer, but parents learn about class composition only in September, shortly before or at school starts. At this point, parents who are unhappy with a teacher or classroom assignment are likely to find it difficult to change schools.

Italian schools have long used matriculation exams for tracking and placement in the transition from elementary to middle school and throughout high school, but standardized testing for evaluation purposes is a recent development. In 2008, INVALSI piloted voluntary assessments in elementary school; in 2009 these became compulsory for all schools and students. INVALSI assessments cover mathematics and Italian language skills in a national administration lasting two days in the Spring. INVALSI reports school and class average scores to schools, but not to students. School leaders may choose to release this information to the public.[4]

Test administration protocols play an important role in our story. INVALSI tests include multiple choice questions and open-response items, for which some grading is required. Proctoring and grading are done by local teachers. In addition, teachers are expected to copy students' original responses onto machine-readable answer sheets (called *scheda risposta*, illustrated in Appendix Figure A1), a burdensome clerical task that's meant to be completed shortly within a few days of testing. Teachers tasked with grading and transcription can en-

[4]INVALSI regulations state that folders containing students' answer sheets must identify students using a code unrelated to student names. Only school administrators (and the external monitor, if any) can link these codes with student identities. Individual test scores are never reported or released to students or the public (see http://www.invalsi.it/snv1011/documenti/Informativa_privacy_SNV2010_2011.pdf).

list colleagues for help. Specifically, INVALSI memos on grading protocols allow for multiple teachers to be involved in grading and transcription. It seems likely that multi-teacher grading and transcription appears are the norm for larger classes, as anecdotal evidence and our discussions with administrators suggest. Peer monitoring may therefore reduce manipulation in larger classes. All test-related clerical tasks must be completed at the institution, typically after school, but this is not paid overtime work. Once transcription onto *scheda risposta* is accomplished, the original student test sheets remain at school while the transcribed answer sheets are sent to INVALSI. These procedures, combined with the extra uncompensated work they require, open the door to score manipulation.

In an effort to reduce score manipulation, INVALSI randomly assigns external monitors to about 20% of institutions in the country. Monitors supervise test administration, encouraging compliance with INVALSI testing standards. Monitors are also responsible for score sheet transcription in some (non-randomly) selected classes. Regional education offices select monitors from a pool consisting of retired teachers and principals who have not worked in the towns or at the schools they are assigned to monitor for the preceding two years. Monitors are paid for their work and are required to complete transcription by the end of the test day.

Related Work

Maimonides-style empirical strategies have been used to identify class size effects in many countries, including the US (Hoxby, 2000), France (Piketty, 2004 and Gary-Bobo and Mahjoub, 2013), Norway (Bonesronning, 2003 and Leuven et al., 2008) and the Netherlands (Dobbelsteen et al., 2002). On balance these results point to modest returns to class size reductions, though mostly smaller than those reported by Angrist and Lavy (1999) for Israel. A natural explanation for this finding is the relatively large class size in Israeli elementary schools. In line with this view, Woessmann (2005) finds a weak association between class size and achievement in a cross-country panel covering Western European school systems in which classes tend to be small. Results in Sims (2008) suggest class size reduction obtained throught combination classes has a negative effect on students' achievement.

The returns to class size in Italy have received little attention from researchers to date, in large part because test score data have only recently become available. One of the few Italian micro-data studies we've seen, Bratti et al. (2007), reports estimates showing an

insignificant class size effect. In an aggregate analysis, Brunello and Checchi (2005) look at the relationship between staffing ratios and educational attainment for cohorts born before 1970; they find that lower pupil-teacher ratios at the regional level are associated with higher average schooling. We haven't found other quantitative explorations of Italian class size, though Ballatore et al. (2014) use a Maimonides-type identification strategy to estimate the effects of the number of immigrants in the classroom on native students' achievement.

As noted above, many scholars have documented manipulation in standardized tests. The (natural) experiment used here to identify the effects of Italian score manipulation and class size jointly was first analyzed by Bertoni et al. (2013), who focus on the effects of external classroom monitors on scores. Our analysis of this experiment looks at monitoring effects by region, while also adjusting for features of the scheme that INVALSI uses to assign monitors not fully accounted for in earlier work.

A final closely related set of findings documents a range of economic and behavioral differences across Italian regions. Southern Italy is characterized by low levels of social capital (Guiso et al., 2004; Guiso et al., 2010) and relatively widespread opportunistic behavior and public corruption (Ichino and Ichino, 1997; Ichino and Maggi, 2000). Differences along these dimensions have been used to explain persistent regional differentials in economic outcomes (Costantini and Lupi, 2006) and differences in the quality of governance and civic life (Putnam et al., 1993). Finally, as noted in the Introduction, our work connects with research on teacher shirking around the world.

3 Data and First Stage

Data and Descriptive Statistics

The standardized test score data used in this study come from INVALSI's testing program in Italian elementary schools in the 2009-11 academic school years. Raw scores indicate the number of correct answers. We standardized these by subject, year of survey, and grade to have zero mean and unit variance. Data on test scores were matched to administrative information describing institutions, schools, classes, and students. Class size is measured by administrative enrollment counts at the beginning of the school year. Student data include gender, citizenship, and parents' employment status and educational background. These

data are collected as part of test administration and supposed to be provided by school staff when scores are submitted. Italian students attending private primary schools are omitted from this study (these account for less than 10 percent of enrollment).

Our statistical analysis focuses on class-level averages since this is the aggregation level at which the regressor of interest varies. The empirical analysis is restricted to classes with more than the minimum number of students set by law (10 before 2010 and 15 from 2011). This selection rule eliminates classes in the least populated areas of the country, mostly mountainous areas and small islands. We also drop schools with more than 160 students in a grade, as these are above the threshold where Maimonides' Rule is likely to matter (this trims classes above the 99th percentile of the enrollment-weighted class size distribution).

The matched analysis file includes about 70,000 classes in each of the two grades covered by our three-year window (these are repeated cross-sections; the data structure doesn't follow the same classes over time). Table 1 shows descriptive statistics for the estimation sample by grade. Statistics are reported at the class level in Panel A, at the school level in Panel B, and at the institution level in Panel C. Class size averages around 20 in both grades, and is slightly lower in the South. Although our statistical analyses use standardized scores, the score means reported in Panel A give the class average percent correct. Scores are higher in language than in math and higher in grade 5 than in grade 2. The table also shows averages for an indicator of score manipulation (the construction of this variable is detailed below). Manipulation rates are higher in the South and in math.

Maimonides in Italy

Our identification strategy for class size effects exploits minimum and maximum class sizes (these rules are a laid out in a regulation known as *Decreto Ministeriale 331/98*). Until the 2008 school year, primary school classes had to be between 10 and 25. Grade enrollment beyond 25 or a multiple thereof usually prompted the addition of a class. The rule allows exceptions, however. Principals can reduce the size of any class attended by one or more disabled students, and schools in mountainous or remote areas are allowed to open classes with fewer than 10 students. The law allows a 10% deviation from the maximum in either direction (that is, the Ministry of Education may fund an additional class when enrollment exceeds 22 and typically requires a new class when average enrollment would otherwise exceed 28). A

2009 reform changed size limits to 15 and 27, again with a tolerance of 10% (promulgated through *Decreto del Presidente della Repubblica 81/2009*). This reform was rolled out one grade per year, starting with first grade. In our data, second graders in 2009 and fifth graders in any year were subject to the old rule, while second graders in 2010 and 2011 were subject to the new rule.

Ignoring discretionary deviations near class size cutoffs, Maimonides' Rule predicts class size to be a non-linear and discontinuous function of enrollment. Writing f_{igkt} for the predicted size of class i in grade g at school k in year t, we have

$$f_{igkt} = \frac{r_{gkt}}{[int\left(\left(r_{gkt}-1\right)/c_{gt}\right)+1]}, \tag{1}$$

where r_{gkt} is beginning-of-the-year grade enrollment at school k, c_{gt} is the relevant cap (25 or 27) for grade g, and $int(x)$ is the largest integer smaller than or equal to x. Figures 2 and 3 plot average class size and f_{igkt} against enrollment in grade, separately for pre- and post-reform periods. Plotted points show the average actual class size at each level of enrollment. Actual class size follows predicted class size reasonably closely for enrollments below about 75, especially in the pre-reform period. Predicted discontinuities in the class size/enrollment relationship are rounded by the soft nature of the rule. Many classes are split before reaching the theoretical maximum of 25. Earlier-than-mandated splits occur more often as enrollment increases. In the post-reform period, class size tracks the rule generated by the new cap of 27 poorly when enrollment exceeds about 70.

Measuring Manipulation

Our score manipulation variable is a function of implausible score levels, the within-class average and standard deviation of test scores, the number of missing items, and a Herfindahl index of the share of students with similar response patterns. These indicators are used as inputs for a cluster analysis that flags as suspicious classes with abnormally high performance, an unusually small dispersion of scores, an unusually low proportion of missing items, or a high concentration in response patterns. This procedure yields class-level indicators of compromised scores, separately for math and language. The resulting manipulation indicator is similar to the manipulation variable used in Quintano et al. (2009) and INVALSI publications (e.g., INVALSI, 2010). The INVALSI version generates a continuous class-level

probability of manipulation. The procedure used here generates a dummy variable indicating classes where score manipulation seems likely. Methods and formulas used to identify score manipulation are detailed in the Appendix. A section on threats to validity considers the consequences of possible misclassification of manipulation for our empirical strategy.[5]

4 Class Size Effects: Achievement and Manipulation

Graphical Analysis

We begin with non-parametric RD plots that capture class size effects near enrollment cutoffs. The first in this sequence, Figure 4, documents the relationship between cutoffs (multiples of 25 or 27) and class size. This figure was constructed from a sample of classes at schools with enrollment that falls in a [-12,12] window around the first four cutoffs shown in Figures 2 and 3. Enrollment values in each window are centered to be zero at the relevant cutoff. The y-axis shows average class size conditional on the centered enrollment value shown on the x-axis, reported as a 3-point moving average. Figure 4 also plots fitted values generated by local linear regressions (LLR) fits to class-level data. In this context, the LLR smoother uses data on one side of the cutoff only, smoothed with an edge kernel and Imbens and Kalyanaraman (2012) bandwidth.[6]

In view of the 2-3 student tolerance around the cutoff for the addition of a class, enrollment within two points of the cutoff is excluded from the local linear fit. As a result of this tolerance, class size can be expected to decline at enrollment values shortly before the cutoff and to continue to decline thereafter. Consistent with this expectation, the figure shows a clear drop at the cutoff, with the sharpness of the break moderated by values near the cutoff. Class size is minimized at about 3-5 students to the right of the cutoff instead of immediately after, as we would expect were Maimonides' Rule to be tightly enforced. The parametric identification strategy detailed below exploits both the discontinuous variation

[5]Our procedure also follows Jacob and Levitt (2003) in inferring score manipulation from patterns of answers within and across tests in a classroom. Jacob and Levitt (2003) also compare test scores over time, looking for anomalous changes. Values in the upper tail of the Jacob-Levitt suspicious answer index are highly predictive of their cheating variable in the cross section. Our main results are unchanged when manipulation is measured continuously. A binary indicator leads to parsimonious models and easily interpreted estimates, however, while also facilitating the discussion of misclassification bias.

[6]The figures here plot residuals from a regression of class size on the controls included in equation (2), below.

in class size generated when enrollment moves across cutoffs, changes in slope as a cohort is divided into classes more finely, and the change in the nominal maximum introduced by the 2009/10 reform. Looking only at points immediately adjacent to the cutoff, the change in size generated by moving across a cutoff is on the order of 2-3 students.

When plotted as a function of enrollment values near Maimonides cutoffs, test scores in the South show a jump that mirrors the drop in class size seen at Maimonides cutoffs. By contrast, there's little evidence of such a jump in schools outside the South. These patterns are documented in Figure 5, which plots math and language scores against enrollment in a format paralleling that of Figure 4. The reduced-form achievement drop for schools in Southern Italy is about 0.02 standard deviations (hereafter, σ). Assuming this reduced-form change in test scores in the neighborhood of Maimonides cutoffs is driven by a causal class size effect, the implied return to a one-student reduction in class size is about 0.01σ in Southern Italy (this comes from dividing 0.02 by a rough first stage of about 2). The absence of a jump in scores at cutoffs in data from schools elsewhere in the country suggests that outside the South class size reductions leave scores unchanged.

Score manipulation also varies as a function of enrollment in the neighborhood of class size cutoffs, with a pattern much like that seen for achievement. This is apparent in Figure 6, which puts the proportion of classes identified as having compromised scores on the y-axis, in a format like that used for Figures 4 and 5. Mirroring the pattern of achievement effects, a discontinuity in score manipulation rates emerges most clearly for schools in Southern Italy. This pattern suggests that the achievement gains generated by class size in Figure 5 may reflect the manipulation behavior captured in Figure 6.

A possible caveat here is the role of mismeasured manipulation might have in generating this pattern. The implications of misclassification for 2SLS estimates of class size effects are explored in detail in a separate section, below. We note here, however, that classification error is unlikely to change discontinuously at Maimonides class size cutoffs. Moreover, the fact that manipulation is essentially smooth through the cutoff for schools outside the South weighs against purely mechanical explanations of the pattern in Figure 6 (mechanical in the sense that components of the manipulation variable might be determined by class size through channels other than changing teacher or student behavior).

Empirical Framework for Class Size Effects

Figure 5 suggests that variation in class size near Maimonides cutoffs can be used to identify class size effects in a non-parametric fuzzy regression discontinuity (RD) framework. In what follows, however, we opt for parametric models that exploit variation in enrollment arising from changes in the slope of the relationship between enrollment and class size, as well as discontinuities. The parametric strategy gains statistical power by combining features of both RD and regression kink designs, while easily accommodating a setup with multiple endogenous variables and covariates.

Our parametric framework models y_{igkt}, the average outcome score in class i in grade g at school k in year t, as a polynomial function of the running variable, r_{gkt}, and class size, s_{igkt}. With quadratic running variable controls, the specification pooling grades and years can be written

$$y_{igkt} = \rho_0(t, g) + \beta s_{igkt} + \rho_1 r_{gkt} + \rho_2 r_{gkt}^2 + \epsilon_{igkt}, \tag{2}$$

where $\rho_0(t, g)$ is shorthand for a full set of year and grade effects. This model also controls for the demographic variables described in Table 1, as well as the stratification variables used in the monitoring experiment to increase precision in the estimates. Standard errors are clustered on school and grade.[7]

The instrument used for 2SLS estimation of equation (2) is f_{igkt}, as defined in equation (1). In addition to estimates of equation (2), results are also reported from models that include a full set of cutoff-segment (window) main effects, allowing the quadratic control function to differ across segments (we refer to this as the interacted specification).[8] The corresponding OLS estimates for models without interacted running variable controls are shown as a benchmark. As can be seen in columns 1-3 of Table 2, these show a negative correlation between class size and achievement for schools in the Northern and Central regions, but not in the South (class size effects are scaled for a 10-student change). Larger classes are asso-

[7]Control variables include percent female in the class, the proportion of immigrants, the proportion of students whose father is a high school graduate, have unemployed mothers, have mothers not in the labor force, have employed mothers, and dummies for missing values for these variables. Stratification controls consist of total enrollment in grade, region dummies, and the interaction between enrollment and region. Results with linear control in the running variable only, estimated on samples limited to the 12-student bandwidth, are similar.

[8]Pre-reform segments cover the intervals 10-37, 38-62, 63-87, 88-112, 113-137, and 138-159; post-reform segments cover the intervals 15-40, 41-67, 68-94, 95-121, and 122-159. These segments cover intervals of width +/- 12 in the pre-reform period and +/-13 in the post-reform period, with modifications at the lower and upper segments to include a few larger and smaller values.

ciated with somewhat higher language scores in the South while Southern class sizes appear to be unrelated to achievement in math.

2SLS estimates using Maimonides' Rule, reported in columns 4-9 of Table 2, suggest that larger classes reduce achievement in both math and language. The associated first stage estimates, which can be seen in Appendix Table A1, show that predicted class size increases actual class size with a coefficient around one-half when regions are pooled, with a first stage effect of 0.43 in the South and 0.55 elsewhere. 2SLS estimates for Southern schools, implying something on the order of a 0.10σ achievement gain for a 10-student reduction, are 2-3 times larger than the corresponding estimates for schools outside the South. The 2SLS estimates are reasonably precise; only estimates of the interacted specification for language scores from non-Southern schools fall short of conventional levels of statistical significance. On balance, the results in Table 2 indicate a substantial achievement payoff to class size reductions, though the gains here are not as large as those reported by Angrist and Lavy (1999) for Israel. A substantive explanation for this difference in findings might be concavity in the relationship between class size and achievement, combined with Italy's much smaller average class sizes.

Class Size and Manipulation

The estimates in Table 3 suggest that the causal effect of class size on measured achievement reported in Table 2 need not reflect more learning in smaller classes. This table reports estimates from specifications identical to those used to construct the estimates in Table 2, with the modification that a class-level score manipulation indicator replaces achievement as an outcome. The 2SLS estimates in columns 4-9 show a large and precisely-estimated negative effect of class size on manipulation rates, with effects on the order of 4-6 percentage points for a 10-student class size increase in the South. Estimates for schools outside the South also show a negative relationship between class size and score manipulation, though here the estimated effects are much smaller and significantly different from zero in only one case (language scores from the non-interacted specification). OLS estimates of effect of class size on score manipulation, though smaller in magnitude, reflect the same negative effects as 2SLS.

5 The Anatomy of Manipulation

INVALSI's randomized monitoring policy provides key evidence on the nature and consequences of score manipulation. Institutions are sampled for monitoring with a probability proportional to grade enrollment in the year of the test. Sampling is also stratified by regions.

Table 4 documents balance across institutions with and without randomly assigned monitors. Specifically, this table shows regression-adjusted treatment-control differences from models that control for strata in the monitoring sample design. These specifications include a full set of region dummies and a linear function of institutional grade enrollment that varies by regions. Administrative variables - generated as a by-product of school administration and INVALSI testing - are well-balanced across groups, as can be seen in the small and insignificant coefficient estimates reported in Panel A of the table.[9] Demographic data and other information provided by school staff, such as parental information, show evidence of imbalance. This seems likely to reflect the influence of monitoring on data quality, rather than a problem with the experimental design or implementation. The hypothesis that monitors induced more careful data reporting by staff is supported by the large treatment-control differential in missing data rates documented at the bottom of the table. Among other salutary effects, randomly assigned monitors reduce item non-response by as much as three percentage points, as can be seen in Panel C of Table 4. Monitoring effects on data quality at class size cutoffs are discussed in Section 6.

The presence of institutional monitors reduces score manipulation considerably. This is apparent from the estimated monitoring effects shown in columns 1-3 of Table 5. Specifically, monitoring reduces manipulation rates by about 3 percentage points for Italy (column 1), with effects twice as large in the South (column 3). These estimates come from models similar to those used to check covariate balance with a score manipulation indicator replacing covariates as the dependent variable. Monitoring also reduces language scores by 0.08σ, while the estimated monitoring effect on math scores is about -0.11σ. Effects of monitoring in the South range from -0.13σ for language to -0.18σ for math, estimates that appear in column

[9] One class in each grade is selected for monitoring in sampled institutions with grade enrollment below 100. Two classes are selected in remaining institutions (randomness of within-institution monitoring appears to have been compromised in practice). Bertoni et al. (2013) mistakenly treated institutions as schools. Their identification strategy also presumes random assignment of classroom monitors within institutions, but we find that monitors are much more likely to be assigned to large classes, probably a consequence of that fact that in most institutions only one class is monitored.

6 of the table. The fact that monitoring matters shows teachers prefer not to be identified as manipulators.[10]

The estimates reported in columns 1-3 and 4-6 of Table 5 constitute the first stage and reduced form for a model that uses the assignment of monitors as an instrument for the effects of score manipulation on test scores. Dividing reduced form estimates by the corresponding first stage estimates produces second stage manipulation effects of about 3σ for the South, with even larger second stage estimates for the North. These effects seem implausibly large, implying a boost in scores that exceeds the range of the dependent variable in some cases. Because classification error attenuates first stage estimates in this context, the resulting second stage estimates may be proportionally inflated. This and other implications of miss-classification are discussed in Section 6.

Manipulation is Curbstoning

The fact that monitoring reduces score manipulation and that manipulation *decreases* with class size suggests that teachers are the source of manipulation and not students. Honest teacher-proctors should have the same deterrent effect as external monitors on cheating students: both are likely to catch cheaters, perhaps teachers even more so if they recognize cheating more readily. Moreover, any class size effect on student cheating is likely to be positive, that is, larger classes should facilitate student cheating by making cheating harder to detect. Results in Table 3 showing that score manipulation decreases with class size therefore weigh against student cheating. Finally, because individual test scores are never disclosed even to those tested, its hard to see why students might care to cheat (students are informed of disclosure limits when testing begins). At the same time, the fact that teachers must transcribe scores - except when monitors do it for them - provides a natural opportunity for manipulation and misreporting.

The nature of score manipulation is revealed in part by estimation of the difficulty gradient, that is, average reported scores as a function of item difficulty. Reported scores are assumed to reflect two underlying potential score distributions for each item, j, one revealed in the presence of manipulation, denoted $y^j_{igkt}(1)$, and one revealed otherwise, denoted $y^j_{igkt}(0)$.

[10] We find stronger monitoring effects than those reported in Table 5 in a sample where institutions were monitored two years in a row. These results provide further evidence of a social constraint on manipulation.

Observed scores in class i on item j, denoted by y^j_{igkt}, are determined by

$$y^j_{igkt} = (1 - m_{igkt})y^j_{igkt}(0) + m_{igkt}y^j_{igkt}(1),$$

where m_{igkt} is the class-level manipulation indicator (there are about 45 items per year, grade and subject).

The mean of the underlying potential scores determining y^j_{igkt} is identified adapting methods developed by Abadie (2002) (an application of this approach to school reform treatment effects appears in Angrist, Pathak, and Walters, 2013). Specifically, we compute 2SLS estimates of the parameters β^j_1 and β^j_0 in models of the form

$$y^j_{igkt}m_{igkt} = \rho_1(t, g) + \beta^j_1 m_{igkt} + \epsilon_{igkt},$$
$$y^j_{igkt}(1 - m_{igkt}) = \rho_0(t, g) + \beta^j_0(1 - m_{igkt}) + \epsilon_{igkt},$$

using data from the South, where manipulation is prevalent. Manipulation indicators, m_{igkt} and $1-m_{igkt}$, are treated as endogenous and instrumented by randomly assigned institutional monitoring, M_{igkt}. The resulting estimates of β^j_1 capture potential scores on item j under manipulation for complying classes, that is, for classes in which we can expect manipulation in the absence of monitoring and honest scoring otherwise. Similarly, the parameter β^j_0 is the average potential score on item j without manipulation for the same classes. The two potential scores are then plotted against item difficulty, proxied using percent correct on item j for monitored institutions in Veneto (a province where manipulation rates are very low). This follows INVALSI practice, which benchmarks official reports of score manipulation rates using Veneto as a non-manipulating standard (see, e.g., Falzetti, 2013).

Manipulation indeed changes the relationship between item difficulty and test scores markedly, pushing an otherwise steep difficulty gradient up to a high level, with scores uniformly close to 100 percent correct. This can be seen in Figure 7, which also shows a least squares fit to the relationship, weighted by the precision of the item-level estimates. When accountability concerns are paramount, manipulation of difficult items generates the largest payoff: selective manipulation maximizes gains and minimizes risk if the goal is solely to boost measured achievement. As an empirical matter, selective manipulation should flatten the score gradient at high levels of difficulty, leaving the gradient unchanged for easy items. In other words, selective manipulation of difficult items makes the overall relationship

between manipulated scores and item difficulty convex. By contrast, copying entire answer sheets should push scores on all items up to the same high (near-perfect) level, as in the figure.

The figure also distinguishes items by the level of effort required for transcription. Some items are transcribed quickly and easily onto the machine-readable *scheda risposta*, but others require thought and judgment; transcription of these items is more of a grading exercise than a copying task. Examples of high-grading-effort items are given in the Appendix. In view of this difference in effort, teachers might target high-grading-effort items for manipulation. If manipulators focus on high-grading-effort items, we should see large score differences by manipulation status for such items only. A comparison of the left and right panels in Figure 7, however, offers little evidence of such targeted manipulation behavior: conditional on difficulty, the difference in scores between manipulators and non-manipulators is similar for high- and low-grading-effort items.

The item-level analysis offers little evidence of selective manipulation of difficult or harder-to-grade items. The fact that manipulated scores are well above honest scores also makes pervasive random transcription unlikely. What sort of behavior is consistent with the patterns apparent in the figure? In this case, the simplest story seems most likely: manipulating teachers would appear to forgo honest transcription entirely, copying entire answer sheets, without regard to item characteristics. In other words, manipulation reflects a form of dishonest reporting akin to "curbstoning" in survey research.

Why Small Classes Increase Manipulation

The mediating role of monitoring in the link between class size and measured achievement is supported by Table 6, which reports 2SLS estimates of class size effects on test scores for institutions with and without INVALSI monitors. Specifically, the table reports 2SLS estimates of coefficients on $M_{igkt}s_{igkt}$ and $(1-M_{igkt})s_{igkt}$ in models like those used to construct the estimates reported in Table 2. These estimates reveal a strong negative effect of class size on achievement, but much more so in the absence of monitoring (and, again, in the South). These findings are consistent with the view that in the absence of monitors, smaller class sizes increase reported scores because they facilitate or encourage manipulation.

The link between teacher manipulation and class size can be explained using a stylized

model of grading behavior. Consider a testing system in which scores vary across items and as a result of manipulation, but not otherwise. Without manipulation, the score on item j is $L_j \in (0,1)$. As suggested by Figure 7, manipulation boosts scores to one. The average score on item j in a class of size s is therefore

$$y_j = L_j + \tau_j p_j,$$

where $p_j = \frac{n_j}{s}$ is the manipulation rate for item j, n_j is number of score sheets manipulated and $\tau_j \equiv 1 - L_j \in (0,1)$. The score gain from manipulation is $\widetilde{p}_j \equiv \tau_j p_j$, implying $\frac{\partial \widetilde{p}_j}{\partial n_j} = \frac{\tau_j}{s}$. This reflects the fact that the value of a single exam manipulated declines with class size, and that the gains from manipulation are larger for more difficult items (that is, for large τ_j).

Teachers decide to manipulate in view of grading costs, the risk of discovery and score gains. Although teachers in the Italian public sector are unlikely to be fired for manipulation, we expect there is still a social constraint; this explains, for example, lower manipulation rates in the North. Teachers maximize a risk-adjusted utility of class performance minus grading costs. The latter are assumed to increase linearly in the number of score sheets to be transcribed, hence in class size, while manipulation reduces grading costs to zero. Assuming that the risk of disclosure increases linearly across items manipulated, and that utility is linear in score gains, the teacher's problem can be written

$$\max_{\widetilde{\mathbf{p}}} \underbrace{\left(1 - \gamma(s) \sum_j n_j\right)}_{disclosure\ risk} \quad \underbrace{\alpha \sum_j \widetilde{p}_j}_{utility\ of\ score\ gain} \quad - \quad \underbrace{\beta \sum_j (s - n_j)}_{honest\ grading\ effort},$$

where $\alpha \sum_j \widetilde{p}_j$ is the utility of overall exam performance, $\gamma(s) \sum_j n_j$ is discovery risk, $\beta \sum_j (s - n_j)$ is the disutility of honest grading, and utility falls to zero when manipulation is discovered. Parameters α and β reflect the relative weight teachers place on grading effort and discovery-weighted score gains. Consistent with the idea of increased peer monitoring in large classes, the risk of disclosure increases with class size through the function $\gamma(s)$.

A single manipulated exam yields a disclosure-weight utility gain that decreases with class size, specifically a gain of $\alpha \frac{\tau_j}{s}$, with the addition utility of a constant reduction in grading effort, β. Utility gains from manipulation are offset by increased disclosure risk of amount $\gamma(s)$. Disclosure risk is presumably lower in small classes where teachers transcribe score

sheets unassisted by peers. A distribution of manipulation behavior can be generated by modeling utility and aversion to honest grading as teacher-specific. Even so, when $\gamma(s) \approx 0$, this model predicts manipulation of entire score sheets for entire classes. This behavior produces the pattern seen in Figure 7, which shows near-perfect exams on all items in classes identified as having manipulated scores.

Manipulation effects on achievement are given by $\frac{dy_j}{ds} = \frac{\tau_j}{s}\left[\frac{dn_j}{ds} - p_j\right]$ with $\frac{dn_j}{ds} < 0$,[11] so $\frac{dy_j}{ds}$ must be negative, while increasing class size reduces scores more for large p_j than for small. This pattern arises even when teachers care little about measured achievement per se, that is, when the utility of overall exam performance is flat, say, at $\bar{u}$. If we imagine $\alpha \sum_j \widetilde{p}_j$ is constant, the objective function above amounts to a comparison of β, the utility gained by not having to grade an item, and $\gamma(s)\bar{u}$, the utility cost of disclosure, which is more likely to exceed β in large classes. Allowing for convex disutility of effort moderates the negative effect of increasing class size on manipulation. For example, when the cost of honest grading is

$$c(s, n_j) = \sum_j \beta_1 (s - n_j) + \beta_2 (s - n_j)^2,$$

(with positive β_1 and β_2) the gains from score manipulation (that is, the reduction in costs associated with an increase in n_j) are larger in larger classes. This can be seen by writing the marginal cost reduction as

$$\frac{\partial c}{\partial n_j} = -(\beta_1 + 2\beta_2 s) + 2\beta_2 n_j.$$

The bottom line is still unclear, however; what matters is the contrast with the disclosure risk parameterized by $\gamma(s)$. Costs might also be concave if honest graders become more efficient when they grade more.

The Appendix gives a more general version of these results, relaxing linear utility. We're

[11] The FOC for optimal n_j yields

$$\alpha\left(1 - \gamma(s)\sum_j n_j\right) - \frac{\gamma(s)}{\tau_j}\alpha\sum_j \tau_j n_j + \frac{\beta s}{\tau_j} = 0,$$

and comparative statics shows

$$\frac{dn_j}{ds} = -\frac{\beta}{2\tau_j \alpha \gamma(s)} < 0.$$

especially interested in comparative statics predictions for $\frac{dy_j}{ds}$, the effect of class size on the score on item j in a world where class size is unrelated to actual learning. Assuming log-linear preferences (see, for example, Blundell and McCurdy, 1999) we show that

$$\frac{dy_j}{ds} = -\frac{\Delta_j}{s},$$

for a positive quantity, Δ_j. The Appendix also shows that if teachers are largely indifferent to achievement, but seek only to reduce effort without discovery, increasing class size reduces manipulation rates similarly across all items. On the other hand, because $\frac{dy_j}{ds} = \frac{\tau_j}{s}\left[\frac{dn_j}{ds} - p_j\right]$, curbstoning does not eliminate item difficulty as a mediating factor in the relationship between class size and item-level achievement.

6 Score Manipulation Explains Class Size Effects

6.1 Estimates with Two Endogenous Variables

The discussion in the previous section motivates a causal model in which achievement depends on class size (s_{igkt}) and score manipulation (m_{igkt}), both treated as endogenous variables to be instrumented. This model can be written

$$y_{igkt} = \rho_0(t, g) + \beta_1 s_{igkt} + \beta_2 m_{igkt} + \rho_1 r_{gkt} + \rho_2 r_{gkt}^2 + \eta_{igkt}, \tag{3}$$

where $\rho_0(t, g)$ is again a shorthand for year and grade effects. We interpret equation (3) as describing the average achievement that would be revealed by alternative assignments of class size, s_{igkt}, in an experiment that holds m_{igkt} fixed. This model likewise describes causal effects of changing score manipulation rates in an experiment that holds class size fixed. In other words, equation (3) represents a model for potential outcomes indexed against two jointly manipulable treatments.

We estimate equation (3) by 2SLS in a setup that includes the same covariates that appear in the models used to construct the estimates reported in Table 2. The instrument list contains Maimonides' Rule (f_{igkt}) and a dummy indicating classes at institutions with randomly assigned monitors, M_{igkt}. The first-stage equations associated with these two

instruments can be written:

$$s_{igkt} = \lambda_{10}(t,g) + \mu_{11} f_{igkt} + \mu_{12} M_{igkt} + \lambda_{11} r_{gkt} + \lambda_{12} r_{gkt}^2 + \xi_{ik}, \quad (4)$$

$$m_{igkt} = \lambda_{20}(t,g) + \mu_{21} f_{igkt} + \mu_{22} M_{igkt} + \lambda_{21} r_{gkt} + \lambda_{22} r_{gkt}^2 + \upsilon_{ik}, \quad (5)$$

where $\lambda_{10}(t,g)$ and $\lambda_{20}(t,g)$ are shorthand for first-stage year and grade effects. First stage estimates, reported in Table 7, show both a monitoring and a Maimonides' Rule effect on score manipulation, both of which are considerably more pronounced in the South. The Maimonides first stage for class size remains at around one-half, while the presence of a monitor at institution is unrelated to class size. This is consistent with the hypothesis that monitors are randomly assigned to institutions.

The 2SLS estimates of β_2 in equation (3), reported in Table 8, show large effects of manipulation on test scores. At the same time, this table reports small and mostly insignificant estimates of β_1, the coefficient on class size in the multivariate model. In an effort to boost the precision of these estimates, we estimate over-identified models that add four dummies for values of the running variable that fall within 10% of each cutoff, a specification motivated by the non-parametric first stage captured in Figure 4.[12] The most precise of the estimated zeros reported in Table 8, generated by the over-identified specification for Italy as a whole, run no larger than 0.022, with an estimated standard error of 0.015 (for a 10-student increase in class size); these appear in column 4. It's also worth noting that the over-identification p-values associated with these estimates are far from conventional significance levels.

Table 8 also reports 2SLS estimates computed by adding an interaction term, $s_{igkt} m_{igkt}$, to equation (3), using $f_{igkt} M_{igkt}$ and the extra dummy instruments interacted with M_{igkt} as excluded instruments. This specification is motivated by the idea that class size may matter only in a low-manipulation sub-sample, while an additive model like equation (3) may miss this. There is little evidence for interactions, however: the estimated interaction effects, reported in columns 7-9 of Table 8 are not significantly different from zero.

The most important findings in Table 8 are the small and insignificant class size effects for the Italian Mezzogiorno, a result that contrasts with the much larger and statistically significant class size effects for the same area reported in Table 2. In column 9 of the latter table, for example, a 10-student reduction in class size is estimates to boost achievement

[12] First stage estimates for the over-identified model appear in Appendix Table A2.

by 0.10σ or more. The corresponding multivariate estimates in Table 8 are of the opposite sign, showing that larger classes increase achievement, though not by very much. The over-identified estimates come with estimated standard errors ranging from about 0.02 to 0.04, so that the estimated class size effects in Table 2 fall well outside the estimated confidence intervals associated with the multivariate estimates. It seems reasonable, therefore, to interpret the estimated class effects in Table 8 as precise zeros. This in turn aligns with an interpretation of the return to class size in Italy as due entirely to the causal effect of class size on score manipulation, most likely by teachers.

6.2 Threats to validity

We briefly consider three possible threats to validity relevant for the causal interpretation of the estimates in Table 8. An initial concern comes from the fact that one of the four indicators used to construct the score manipulation dummy, that for unusually high average scores, may be connected to score outcomes for reasons unrelated to manipulation. RD estimates of the relationship between class size, score manipulation, and achievement, however, are largely unaffected by substitution of a manipulation variable that ignores score levels.

Two other concerns relate to measurement error in score manipulation and potentially endogenous sorting around class size cutoffs.

Score manipulation with misclassification

The large 2SLS estimates of manipulation effects in Table 8 reflect attenuation bias in first stage estimates if score manipulation is misreported. We show here that, as long as misclassification rates are independent of the instruments, mismeasurement of manipulation leaves 2SLS estimates of *class size effects* in the multivariate model unaffected. This result is derived using a simplified version of the multivariate model, which can be written with a class subscript as

$$y_i = \rho_0 + \beta_1 s_i + \beta_2 m_i^* + \zeta_i, \tag{6}$$

where instruments are assumed to be uncorrelated with the error, ζ_i, as in equation (3). Here, m_i^* is an accurate score manipulation dummy for class i, while m_i is observed score manipulation as before.

Let $z_i = [f_i \; M_i]'$ denote the vector of instruments. Assuming that classification rates are

independent of the instruments conditional on m_i^*, we can write

$$m_i = (1 - \pi_0) + (\pi_0 + \pi_1 - 1)m_i^* + \omega_i, \tag{7}$$

where the residual, ω_i, is defined by

$$\omega_i = m_i - E[m_i|z_i, m_i^*],$$

and π_d, the probability that score manipulation is correctly detected, satisfies

$$P[m_i = d|z_i, m_i^* = d] = P[m_i = d|m_i^* = d] = \pi_d, \tag{8}$$

for $d = 0, 1$. Note that $E[z_i\omega_i] = 0$ by definition of ω_i. Using (7) to substitute for m_i^*, equation (6) can be rewritten

$$y_i = \left[\rho_0 - \frac{\beta_2(1 - \pi_0)}{\pi_0 + \pi_1 - 1}\right] + \beta_1 s_i + \left[\frac{\beta_2}{\pi_0 + \pi_1 - 1}\right] m_i + \left[\zeta_i - \beta_2 \frac{\omega_i}{\pi_0 + \pi_1 - 1}\right]. \tag{9}$$

We assume that the π_d's are strictly greater than 0.5, so that reported score manipulation is a better indicator of actual manipulation than a coin toss. This ensures that the coefficient on m_i in (9) is finite and has the same sign as β_2.

The 2SLS estimate of the coefficient on reported score manipulation is therefore biased upward, since $\pi_0 + \pi_1 - 1$ is strictly between 0 and 1 given these assumptions. This implies that estimates of β_2 for the North/Centre region (columns 2, 5 and 8 of Table 8), where score manipulation is lower and therefore misclassification is higher, are more inflated than in the South. Most importantly, because the feasible estimating equation (9) has a residual uncorrelated with the instruments and the coefficient on class size is unchanged in this model, misclassification of the sort described by (8) leaves estimates of the class size coefficient, β_1, unchanged. Similar results for the consequences of classification error under the same assumptions appear in Kane et al. (1999), Mahajan (2006), and Lewbel (2007), among others, though our work focuses on the consequences for the coefficient on a variable subject to error rather than implications for other regressors in the model.[13]

[13]We can learn whether 2SLS estimates of the coefficient on m_i, that is, the size of the estimated manipulation effects, are plausible by experimenting with data from an area where manipulation rates are low and assuming that true manipulators earn perfect scores. We use data from Veneto, the region with the lowest score manipulation rate in Italy, to estimate β_2 in this scenario by picking 20% of classes at random and re-coding scores for this group to be 100**.** The resulting estimates of β_2 come out at around 2.25σ. Taking this as a benchmark, the manipulation effects in Table 8 are consistent with values of π_j around .8 for Italy (since $\frac{2.25}{2\times.8-1} = 3.75$), though the implied π_j's are closer to .65 for math scores outside the South. These

Sorting near cutoffs

The Maimonides research design identifies causal class size effects assuming that, after adjusting for secular effects of the running variable, predicted class size (f_{igkt}) is unrelated to student or school characteristics. As in other RD-type designs, sorting around cutoffs poses a potential threat to this assumption. Urquiola and Verhoogen (2009) and Baker and Paserman (2013) note that discontinuities in student characteristics near Maimonides cutoffs can arise if parents or school authorities try to shift enrollment to schools where expected class size is small. In our setting, however, an evaluation of the sorting hypothesis is complicated by the link between Maimonides' Rule and score manipulation documented in Table 7. The fact that Maimonides' Rule predicts score manipulation, especially in the South, generates the results in Table 8. An important channel for the link between Maimonides' Rule and manipulation is the fact that monitoring rates are lower in small classes. If the behavior driving manipulation also affects data quality, a conjecture supported by the effects of monitoring on data quality seen in Table 4, we might expect Maimonides' Rule to be related to covariates for the same reason that monitoring is related to covariates.

This expectation is borne out by Table 9, which reports estimates of the link between Maimonides' Rule and covariates in a format paralleling that of Table 4. These estimates come from the reduced form specifications used to generate the 2SLS estimates reported in Table 2, after replacing scores with covariates on the left hand side. The pattern of covariate imbalance in Table 9 mirrors that in Table 4: covariates affected by monitoring are also correlated with Maimonides' Rule, while administrative variables that are unrelated to monitoring are largely orthogonal to Maimonides' Rule. Tables 4 and 9 also reflect similar regional differences in the degree of covariate imbalance, with considerably more imbalance in the South. Additional evidence suggesting that the link between covariates is a data quality effect unrelated to sorting appears in Appendix Table A3. This table shows that f_{igkt} is largely unrelated to covariates in schools with monitors, where manipulation is considerably diminished (though not necessarily eliminated, since some classes in monitored institutions remain unmonitored).

rates seem like reasonable descriptions of the classification process. The possible misclassification of manipulators is further investigated by Battistin et al. (2014) with reference to the problem of regional rankings of performance at the INVALSI test.

7 Summary and Directions for Further Work

The causal effects of class size on Italian primary schoolers' test scores are identified by quasi-experimental variation arising from Italy's version of Maimonides' Rule. The resulting estimates show small classes boost test scores in Southern provinces, an area known as the Mezzogiorno, but not elsewhere. Analyses of data on score manipulation and a randomized institution monitoring experiment reveal substantial manipulation in the Italian Mezzogiorno, most likely by teachers. For a variety of institutional and behavioral reasons, teacher score manipulation is inhibited by larger classes as well as by external monitoring. Estimates of a model that jointly captures the causal effects of class size and score manipulation on measured achievement suggest the returns to class size in the Italian Mezzogiorno are explained by the causal effects of class size on score manipulation, with no apparent gains in learning. These findings show how class size effects can be misleading even where internal validity is probably given. Our results also show how score manipulation can arise as a result of shirking in an institutional setting where standardized assessments are largely divorced from accountability.

These findings raise a number of questions, including those of why teacher manipulation is so much more prevalent in the Italian Mezzogiorno, and what can be done to enhance accurate assessment in Italy and elsewhere. Manipulation in the Italian Mezzogiorno arises in part from local exam proctoring and local transcription of answer sheets, a cost-saving measure. New York's venerable Regent's exams were also graded locally until 2013, an arrangement that likewise appears to have facilitated score manipulation. Moreover, as with INVALSI assessments, manipulation of Regent's scores appears to be unrelated to NCLB-style accountability pressure (Dee et al., 2016). By contrast, the UK's Key Stage 2 primary-level assessments are marked by external examiners, a costly effort that our findings suggest may nevertheless be worthwhile.[14] Another reason to favor external anonymous exam grading is the possibility of gender and ethnicity bias (as documented in Lavy, 2008; Lavy and Sand, 2015; Terrier, 2015; and Greaves and Burgess, 2013). It's also worth asking why class size reductions fail to enhance learning in Italy, while evidence from the US, Israel, and a number of other countries suggest class size reductions often increase learning. We hope to address these questions in future work.

[14] See https://home.edexcelgateway.com/pages/job_search_view.aspx?jobId=537 for information on Key Stage 2 marking costs.

Table 1: Descriptive Statistics

	Grade 2 (2009-2011)			Grade 5 (2009-2011)		
	Italy	North/Centre	South	Italy	North/Centre	South
	(1)	(2)	(3)	(4)	(5)	(6)
			A. Class Characteristics			
Female*	0.49	0.49	0.49	0.49	0.49	0.49
	(0,5)	(0.5)	(0.5)	(0.5)	(0.5)	(0.5)
Immigrant*	0.10	0.14	0.03	0.1	0.14	0.03
	(0.30)	(0.35)	(0.17)	(0.3)	(0.34)	(0.18)
Father HS*	0.34	0.34	0.33	0.32	0.33	0.3
	(0.47)	(0.48)	(0.47)	(0.47)	(0.47)	(0.46)
Mother employed*	0.57	0.68	0.39	0.55	0.66	0.38
	(0.49)	(0.47)	(0.49)	(0.5)	(0.47)	(0.49)
Pct correct: math	47.9	46.1	51.1	64.2	63.3	65.6
	(14.6)	(12.9)	(16.7)	(12.9)	(10.9)	(15.5)
Pct correct: language	69.8	69.2	70.8	74.2	74.3	74.1
	(10.9)	(9.2)	(13.3)	(8.9)	(7.5)	(10.8)
Class size	20.1	20.3	19.9	19.7	19.9	19.3
	(3.40)	(3.35)	(3.48)	(3.72)	(3.67)	(3.76)
Score manipulation: math	0.06	0.02	0.14	0.06	0.02	0.13
	(0.24)	(0.13)	(0.35)	(0.25)	(0.15)	(0.34)
Score manipulation: language	0.05	0.02	0.11	0.06	0.02	0.11
	(0.23)	(0.14)	(0.31)	(0.23)	(0.15)	(0.31)
Number of classes	67,453	42,747	24,706	72,536	44,739	27,797
			B. School Characteristics			
Number of classes	1.95	1.87	2.11	1.94	1.85	2.10
	(1.10)	(1.01)	(1.27)	(1.10)	(0.98)	(1.28)
Enrollment	40.5	38.8	43.8	38.9	37.3	41.8
	(25.2)	(23.0)	(28.6)	(25.2)	(22.8)	(28.9)
Number of schools	34,591	22,863	11,728	37,476	24,225	13,251
			C. Institution Characteristics			
Number of schools	2.00	2.32	1.57	2.10	2.42	1.69
	(1.05)	(1.13)	(0.74)	(1.09)	(1.17)	(0.81)
Number of classes	3.89	4.33	3.31	4.07	4.48	3.55
	(1.97)	(1.95)	(1.85)	(1.95)	(1.91)	(1.88)
Enrollment	86.0	95.3	73.7	85.2	94.0	73.9
	(40.6)	(39.5)	(38.7)	(40.5)	(39.1)	(39.3)
External monitor	0.22	0.20	0.23	0.22	0.20	0.23
	(0.41)	(0.40)	(0.42)	(0.41)	(0.4)	(0.42)
Number of institutions	17,333	9,866	7,467	17,830	9,997	7,833

Notes: Means and standard deviations are computed using one observation per class in Panel A, one observation per school in Panel B, and one observation per institution in Panel C

* conditional on non-missing survey response.

Table 2: OLS and IV/2SLS Estimates of the Effect of Class Size on Test Scores

	OLS			IV/2SLS					
	Italy	North/Centre	South	Italy	North/Centre	South	Italy	North/Centre	South
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
				A. Math					
Class size	-0.0078	-0.0224***	0.0091	-0.0519***	-0.0436***	-0.0957***	-0.0609***	-0.0417**	-0.1294**
	(0.0070)	(0.0067)	(0.0146)	(0.0134)	(0.0115)	(0.0362)	(0.0196)	(0.0171)	(0.0507)
Enrollment	✗	✗	✗	✗	✗	✗	✗	✗	✗
Enrollment squared	✗	✗	✗	✗	✗	✗	✗	✗	✗
Interactions							✗	✗	✗
N	140,010	87,498	52,512	140,010	87,498	52,512	140,010	87,498	52,512
				B. Language					
Class size	0.0029	-0.0188***	0.0328***	-0.0395***	-0.0313***	-0.0641**	-0.0409***	-0.0215	-0.0937**
	(0.0055)	(0.0053)	(0.0114)	(0.0106)	(0.0092)	(0.0289)	(0.0155)	(0.0136)	(0.0403)
Enrollment	✗	✗	✗	✗	✗	✗	✗	✗	✗
Enrollment squared	✗	✗	✗	✗	✗	✗	✗	✗	✗
Interactions							✗	✗	✗
N	140,010	87,498	52,512	140,010	87,498	52,512	140,010	87,498	52,512

Notes: Columns 1-3 report OLS estimates of the effect of class size on scores. Columns 4-9 report 2SLS estimates using Maimonides' Rule as an instrument. The unit of observation is the class. Class size coefficients show the effect of 10 students. Models with interactions allow the quadratic running variable control to differ across windows of ±12 students around each cutoff. Robust standard errors, clustered on school and grade, are shown in parentheses. Control variables include: % female students, % immigrants, % fathers at least high school graduate, % employed mothers, % unemployed mothers, % mother NILF, grade and year dummies, and dummies for missing values . All regressions include sampling strata controls (grade enrollment at institution, region dummies and their interactions). * significant at 10%; ** significant at 5%; *** significant at 1%.

Table 3: OLS and IV/2SLS Estimates of the Effect of Class Size on Score Manipulation

	OLS			IV/2SLS					
	Italy	North/Centre	South	Italy	North/Centre	South	Italy	North/Centre	South
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
					A. Math				
Class size	-0.0163***	-0.0074***	-0.0309***	-0.0186***	-0.0042	-0.0542***	-0.0179***	-0.0053	-0.0471**
	(0.0025)	(0.0017)	(0.0058)	(0.0047)	(0.0031)	(0.0143)	(0.0069)	(0.0045)	(0.0202)
Enrollment	✗	✗	✗	✗	✗	✗	✗	✗	✗
Enrollment squared	✗	✗	✗	✗	✗	✗	✗	✗	✗
Interactions							✗	✗	✗
N	139,996	87,491	52,505	139,996	87,491	52,505	139,996	87,491	52,505
					B. Language				
Class size	-0.0166***	-0.0120***	-0.0244***	-0.0202***	-0.0116***	-0.0400***	-0.0161**	-0.0059	-0.0379**
	(0.0023)	(0.0018)	(0.0051)	(0.0043)	(0.0032)	(0.0128)	(0.0063)	(0.0048)	(0.0177)
Enrollment	✗	✗	✗	✗	✗	✗	✗	✗	✗
Enrollment squared	✗	✗	✗	✗	✗	✗	✗	✗	✗
Interactions							✗	✗	✗
N	140,003	87,493	52,510	140,003	87,493	52,510	140,003	87,493	52,510

Notes: Columns 1-3 report OLS estimates of the effect of class size on score manipulation. Columns 4-9 report 2SLS estimates using Maimonides' Rule as an instrument. Class size coefficients show the effect of 10 students. Models with interactions allow the quadratic running variable control to differ across windows of ±12 students around each cutoff. The unit of observation is the class. Robust standard errors, clustered on school and grade, are shown in parentheses. Control variables include: % female students, % immigrants, % fathers at least high school graduate,% employed mothers, % unemployed mothers, % mother NILF, grade and year dummies, and dummies for missing values. All regressions include sampling strata controls (grade enrollment at institution, region dummies and their interactions). * significant at 10%; ** significant at 5%; *** significant at 1%.

Table 4: Covariate Balance in the Monitoring Experiment

	Italy		North/Centre		South	
	Control Mean	Treatment Difference	Control Mean	Treatment Difference	Control Mean	Treatment Difference
	(1)	(2)	(3)	(4)	(5)	(6)
			A. Administrative Data on Schools			
Class size	19.812	0.0348	20.031	0.0179	19.456	0.0623
	[3.574]	(0.0303)	[3.511]	(0.0374)	[3.646]	(0.0515)
Grade enrollment at school	53.119	-0.4011	49.804	-0.5477	58.483	-0.1410
	[30.663]	(0.3289)	[27.562]	(0.3913)	[34.437]	(0.5909)
% in class sitting the test	0.939	0.0001	0.934	0.0006	0.947	-0.0007
	[0.065]	(0.0005)	[0.066]	(0.0006)	[0.062]	(0.0008)
% in school sitting the test	0.938	-0.0001	0.933	0.0005	0.946	-0.0010
	[0.054]	(0.0005)	[0.055]	(0.0006)	[0.051]	(0.0008)
% in institution sitting the test	0.937	-0.0001	0.932	0.0005	0.945	-0.0010
	[0.045]	(0.0004)	[0.043]	(0.0005)	[0.045]	(0.0007)
			B. Data Provided by School Staff			
Female students	0.482	0.0012	0.483	0.0004	0.479	0.0027*
	[0.121]	(0.0009)	[0.1179]	(0.0011)	[0.126]	(0.0016)
Immigrant students	0.097	0.0010	0.137	0.0004	0.031	0.0020***
	[0.120]	(0.0010)	[0.13]	(0.0014)	[0.056]	(0.0007)
Father HS	0.25	0.0060***	0.258	0.0061***	0.238	0.0056**
	[0.168]	(0.0016)	[0.163]	(0.0019)	[0.176]	(0.0027)
Mother employed	0.441	0.0085***	0.532	0.0067**	0.295	0.0117***
	[0.267]	(0.0024)	[0.258]	(0.0031)	[0.210]	(0.0035)
			C. Non-Response Indicators			
Missing data on father's education	0.223	-0.0217***	0.225	-0.0186***	0.221	-0.0271***
	[0.341]	(0.0034)	[0.340]	(0.0043)	[0.343]	(0.0057)
Missing data on mother's occupation	0.195	-0.0168***	0.196	-0.0083**	0.194	-0.0316***
	[0.328]	(0.0033)	[0.325]	(0.0042)	[0.333]	(0.0054)
Missing data on country of origin	0.033	-0.0115***	0.025	-0.0078***	0.045	-0.0178***
	[0.163]	(0.0013)	[0.143]	(0.0014)	[0.192]	(0.0026)
N	140,010		87,498		52,512	

Notes: Columns 1, 3 and 5 show means and standard deviations for variables listed at left. Other columns report coefficients from regressions of each variable on a treatment dummy (indicating classroom monitoring), grade and year dummies, and sampling strata controls (grade enrollment at institution, region dummies and their interactions). Standard deviations for the control group are in square brackets, robust standard errors are in parentheses. * significant at 10%; ** significant at 5%; *** significant at 1%.

Table 5: Monitoring Effects on Score Manipulation and Test Scores

	Score manipulation			Test scores		
	Italy	North/Centre	South	Italy	North/Centre	South
	(1)	(2)	(3)	(4)	(5)	(6)
			A. Math			
Monitor at institution (M_{igkt})	-0.029***	-0.010***	-0.062***	-0.112***	-0.075***	-0.180***
	(0.002)	(0.001)	(0.004)	(0.006)	(0.005)	(0.012)
Means	0.064	0.020	0.139	0.007	-0.074	0.141
(sd)	(0.246)	(0.139)	(0.346)	(0.637)	(0.502)	(0.796)
N	139,996	87,491	52,505	140,010	87,498	52,512
			B. Language			
Monitor at institution (M_{igkt})	-0.025***	-0.012***	-0.047***	-0.081***	-0.054***	-0.131***
	(0.002)	(0.001)	(0.004)	(0.004)	(0.004)	(0.009)
Means	0.055	0.023	0.110	0.01	-0.005	0.035
(sd)	(0.229)	(0.149)	(0.313)	(0.523)	(0.428)	(0.649)
N	140,003	87,493	52,510	140,010	87,498	52,512

Notes: Columns 1-3 report first stage estimates of the effect of a monitor at institution on score manipulation. Columns 4-6 show the reduced form effect of a monitor at institution on test scores. All models control for a quadratic in grade enrollment, segment dummies and their interactions. The unit of observation is the class. Robust standard errors, clustered on school and grade, are shown in parentheses. Control variables include: % female students, % immigrants, % fathers at least high school graduate, % employed mothers, % unemployed mothers, % mother NILF, grade and year dummies, and dummies for missing values in these variables. All regressions include sampling strata controls (grade enrollment at institution, region dummies and their interactions). * significant at 10%; ** significant at 5%; *** significant at 1%.

Table 6: IV/2SLS Estimates of the Effect of Class Size on Scores by Monitor at Institution

	Italy	North/Centre	South
	(1)	(2)	(3)
		A. Math	
Class size* M_{igkt}	-0.0351	-0.0389*	-0.0347
	(0.0237)	(0.0211)	(0.0605)
Class size* (1- M_{igkt})	-0.0658***	-0.0420**	-0.1433***
	(0.0207)	(0.0180)	(0.0526)
M_{igkt}	-0.1736***	-0.0815**	-0.3947***
	(0.0413)	(0.0376)	(0.0959)
N	140,010	87,498	52,512
		B. Language	
Class size* M_{igkt}	-0.0307	-0.0208	-0.0485
	(0.0188)	(0.0169)	(0.0480)
Class size* (1- M_{igkt})	-0.0419**	-0.0212	-0.0975**
	(0.0164)	(0.0144)	(0.0419)
M_{igkt}	-0.1033***	-0.0545*	-0.2279***
	(0.0328)	(0.0300)	(0.0764)
N	140,010	87,498	52,512

Notes: This table report 2SLS estimates using the interaction of Maimonides' Rule with monitor at institution (M_{igkt}) as instruments. Class size coefficients show the effect of 10 students. The unit of observation is the class. Robust standard errors, clustered on school and grade, are shown in parentheses. Control variables include: % female students, % immigrants, % fathers at least high school graduate,% employed mothers, % unemployed mothers, % mother NILF, grade and year dummies, and dummies for missing values. All regressions include sampling strata controls (grade enrollment at institution, region dummies and their interactions). * significant at 10%; ** significant at 5%; *** significant at 1%

Table 7: Twin First Stages

	A. Score Manipulation					
	Math			Language		
	Italy	North/Centre	South	Italy	North/Centre	South
	(1)	(2)	(3)	(4)	(5)	(6)
Maimonides' Rule (f_{igkt})	-0.0009**	-0.0003	-0.0019**	-0.0008**	-0.0003	-0.0015**
	(0.0004)	(0.0002)	(0.0009)	(0.0003)	(0.0003)	(0.0008)
Monitor at institution (M_{igkt})	-0.029***	-0.010***	-0.062***	-0.025***	-0.012***	-0.047***
	(0.002)	(0.001)	(0.004)	(0.002)	(0.001)	(0.004)
N	139,996	87,491	52,505	140,003	87,493	52,510
	B. Class size					
	Italy	North/Centre	South			
	(1)	(2)	(3)			
Maimonides' Rule (f_{igkt})	0.513***	0.555***	0.433***			
	(0.0006)	(0.0008)	(0.0011)			
Monitor at institution (M_{igkt})	0.013	0.032	-0.009			
	(0.024)	(0.027)	(0.045)			
N	140,010	87,498	52,512			

Notes: Panel A report first stage estimates of the effect of the Maimonides' Rule and a monitor at institution on score manipulation. Panel B report first stage estimates of the effect of the Maimonides' Rule and a monitor at institution on class size. All models control for a quadratic in grade enrollment, segment dummies and their interactions. The unit of observation is the class. Robust standard errors, clustered on school and grade, are shown in parentheses. Control variables include: % female students, % immigrants, % fathers at least high school graduate, % employed mothers, % unemployed mothers, % mother NILF, grade and year dummies, and dummies for missing values in these variables. All regressions include sampling strata controls (grade enrollment at institution, region dummies and their interactions). * significant at 10%; ** significant at 5%; *** significant at 1%.

Table 8: IV/2SLS Estimates of the Effect of Class Size and Score Manipulation on Test Scores

	IV/2SLS			IV/2SLS (overidentified)			IV/2SLS (overidentified-interacted)		
	Italy	North/Centre	South	Italy	North/Centre	South	Italy	North/Centre	South
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
				A. Math					
Class size	0.0075	-0.0029	0.0062	0.0024	-0.0113	0.0133	0.0116	0.0136	0.0473
	(0.0213)	(0.0298)	(0.0441)	(0.0190)	(0.0251)	(0.0378)	(0.0316)	(0.0482)	(0.0675)
Score manipulation	3.82***	7.33***	2.88***	3.82***	7.02***	2.87***	4.10***	9.21**	3.33***
	(0.19)	(0.79)	(0.16)	(0.19)	(0.73)	(0.16)	(0.96)	(4.41)	(0.86)
Class size * Score manipulation							-0.1464	-1.2700	-0.2273
							(0.4814)	(2.1598)	(0.4304)
Overid test [P-value]				[0.914]	[0.600]	[0.541]	[0.914]	[0.475]	[0.476]
N	139,996	87,491	52,505	139,996	87,491	52,505	139,996	87,491	52,505
				B. Language					
Class size	0.0121	0.0049	0.0127	0.0218	0.0109	0.0491	0.0325	0.0098	0.1337*
	(0.0173)	(0.0196)	(0.0385)	(0.0153)	(0.0174)	(0.0329)	(0.0308)	(0.0320)	(0.0800)
Score manipulation	3.29***	4.50***	2.80***	3.21***	4.34***	2.74***	3.59***	4.31*	4.18***
	(0.18)	(0.45)	(0.18)	(0.18)	(0.42)	(0.18)	(1.03)	(2.25)	(1.30)
Class size * Score manipulation							-0.2130	-0.0029	-0.7058
							(0.4980)	(1.0898)	(0.6214)
Overid test [P-value]				[0.129]	[0.796]	[0.036]	[0.216]	[0.844]	[0.109]
N	140,003	87,493	52,510	140,003	87,493	52,510	140,003	87,493	52,510

Notes: Columns 1-3 show 2SLS estimates using Maimonides' Rule and monitor at institution as instruments. Columns 4-6 show overidentified 2SLS estimates which also use dummies for grade enrollment being in a 10 percent window below and above each cutoff (2 students) as instrument. Columns 7-9 add the interaction between class size and score manipulation and use the interaction of Maimonide's Rule with monitor at institution and the interactions of dummies for grade enrollment being in a 10 percent window below and above each cutoff with monitor at institution as instruments. Class size coefficients show the effect of 10 students. All models control for a quadratic in grade enrollment, segment dummies and their interactions. The unit of observation is the class. Robust standard errors, clustered on school and grade, are shown in parentheses. Control variables include: % female students, % immigrants, % fathers at least high school graduate,% employed mothers, % unemployed mothers, % mother NILF, grade and year dummies, and dummies for missing values in these variables. All regressions include sampling strata controls (grade enrollment at institution, region dummies and their interactions). * significant at 10%; ** significant at 5%; *** significant at 1%.

Table 9: Maimonides' Rule and Covariate Balance

	Italy		North/Centre		South	
	Control Mean	Treatment Difference	Control Mean	Treatment Difference	Control Mean	Treatment Difference
	(1)	(2)	(3)	(4)	(5)	(6)
			A. Administrative Data on Schools			
% in class sitting the test	0.9392	0.0000	0.9345	0.0001	0.9471	0.0000
	[0.0643]	(0.0001)	[0.0657]	(0.0001)	[0.061]	(0.0001)
% in school sitting the test	0.9386	0.0001	0.9339	0.0001	0.9464	0.0001
	[0.0534]	(0.0001)	[0.0548]	(0.0001)	[0.05]	(0.0001)
% in institution sitting the test	0.9374	-0.0001	0.9327	-0.0001	0.9451	-0.0000
	[0.0436]	(0.0001)	[0.0426]	(0.0001)	[0.0441]	(0.0001)
			B. Data Provided by School Staff			
Female	0.482	0.0000	0.4836	0.0002	0.4792	-0.0002
	[0.1205]	(0.0002)	[0.1176]	(0.0002)	[0.1251]	(0.0003)
Immigrant	0.0981	-0.0007***	0.1375	-0.0007***	0.0324	-0.0004***
	[0.1198]	(0.0002)	[0.1298]	(0.0003)	[0.0572]	(0.0001)
Father HS	0.2546	0.0006**	0.2613	0.0002	0.2434	0.0013***
	[0.1678]	(0.0003)	[0.1626]	(0.0003)	[0.1755]	(0.0005)
Mother employed	0.4503	0.0012***	0.5356	0.0010*	0.3082	0.0016***
	[0.2658]	(0.0004)	[0.2574]	(0.0005)	[0.2138]	(0.0006)
			C. Non-Response Indicators			
Missing data on father's education	0.2187	0.0003	0.2216	0.0015**	0.2139	-0.0018*
	[0.3361]	(0.0006)	[0.3358]	(0.0007)	[0.3367]	(0.0010)
Missing data on mother's occupation	0.1925	0.0002	0.1963	0.0014**	0.1861	-0.0019*
	[0.3239]	(0.0006)	[0.3231]	(0.0007)	[0.3251]	(0.0010)
Missing data on country of origin	0.0296	-0.0001	0.0232	-0.0001	0.0401	-0.0000
	[0.1544]	(0.0002)	[0.1361]	(0.0003)	[0.1804]	(0.0005)
N	140,010		87,498		52,512	

Notes: Columns 1, 3 and 5 show means and standard deviations for variables listed at left. Other columns report coefficients from regressions of each variable on predicted class size (Maimonides' Rule), a quadratic in grade enrollment, segment dummies and their interactions, grade and year dummies, and sampling strata controls (grade enrollment at institution, region dummies and their interactions). Standard deviations for the control group are in square brackets, robust standard errors are in parentheses. * significant at 10%; ** significant at 5%; *** significant at 1%.

Figure 1: Manipulation Rates by Province

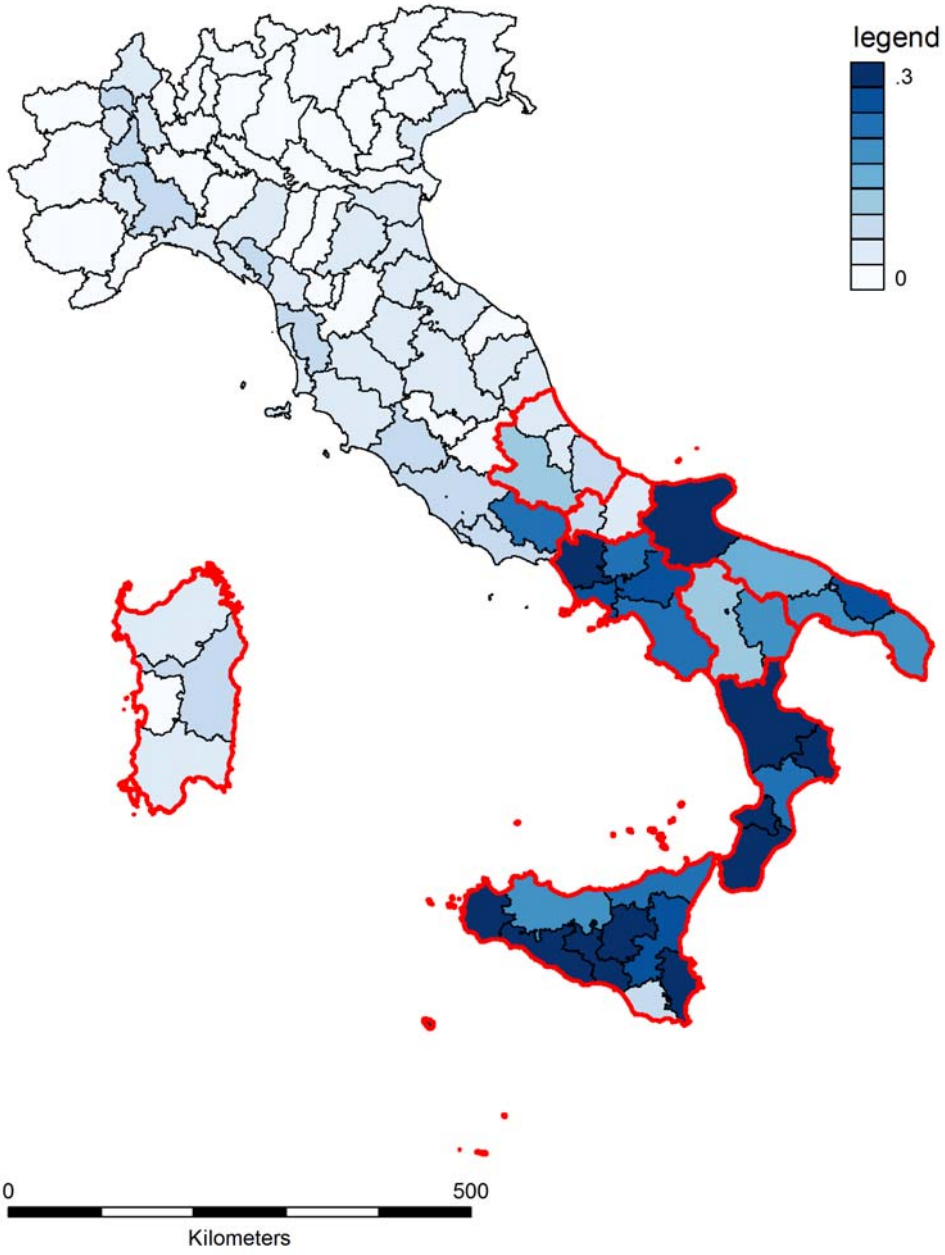
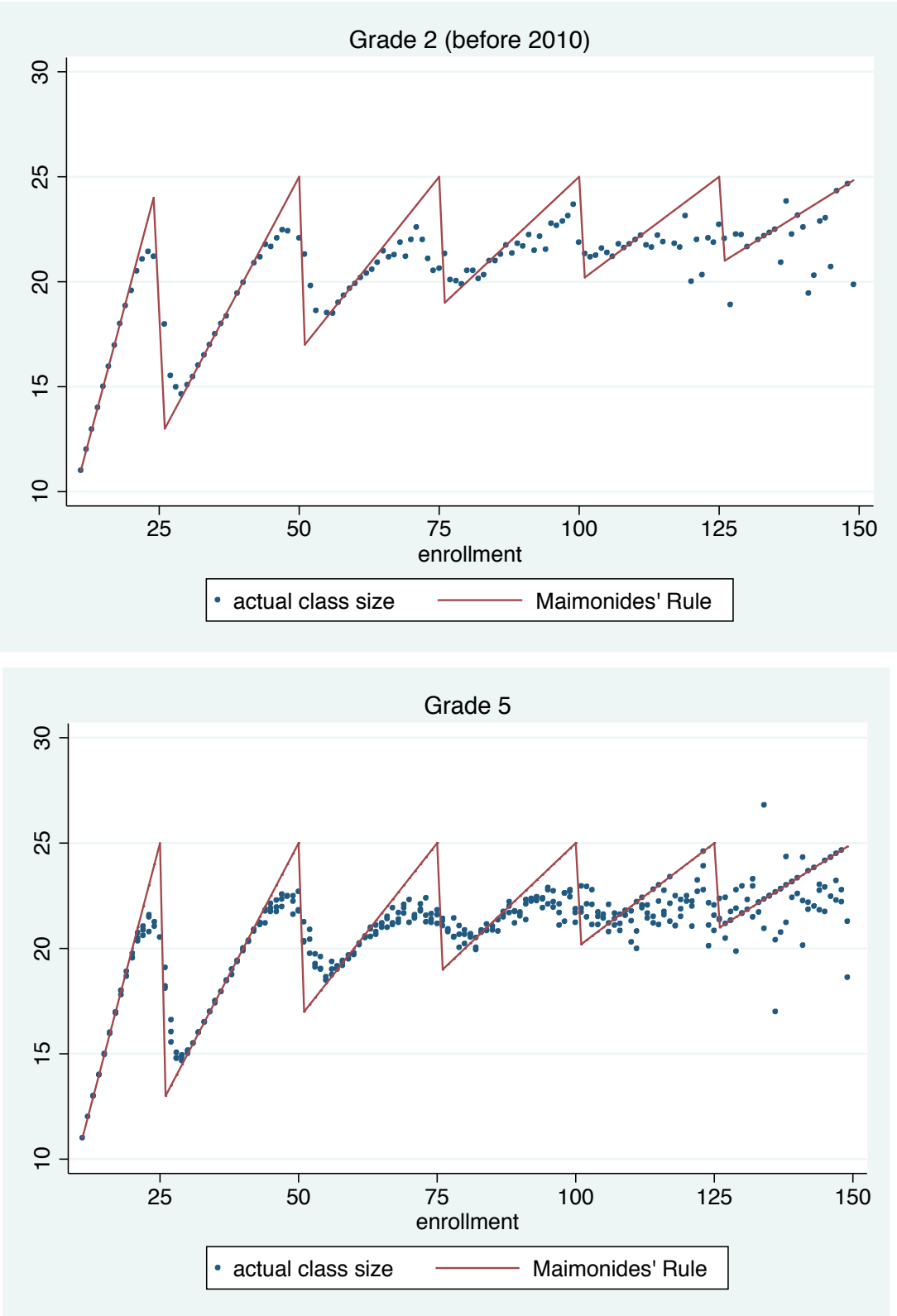


Figure 2: Class Size by Enrollment in Pre-reform Years

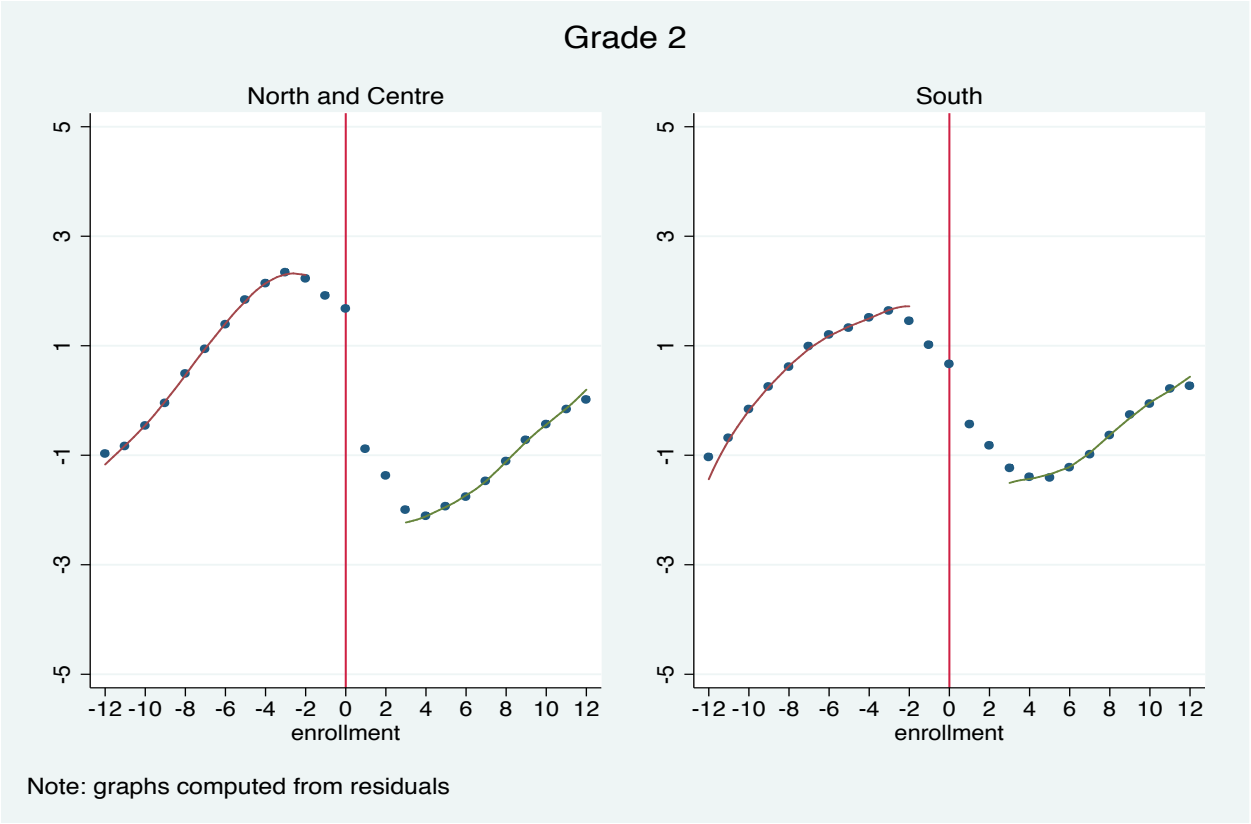


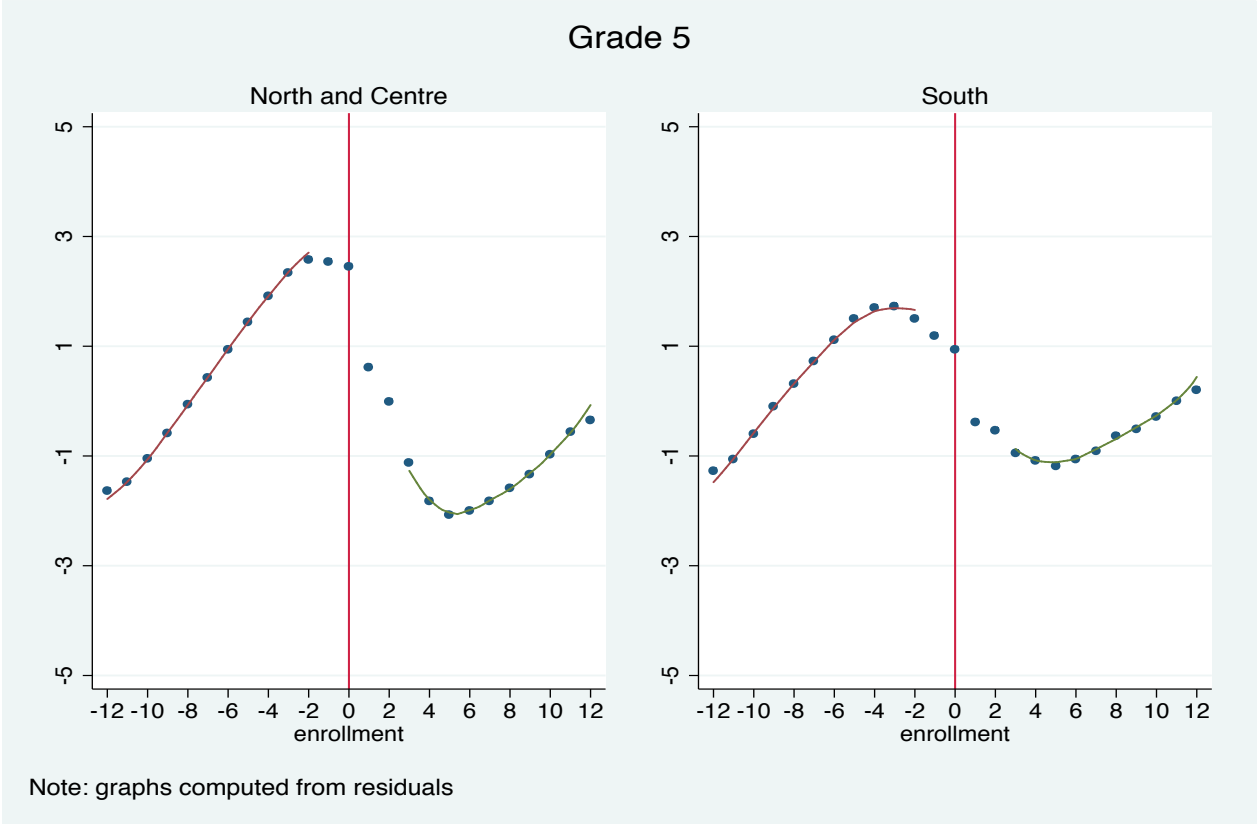
Notes: The figure shows actual class size and as predicted by Maimonides' Rule in pre-reform years

Figure 3: Class Size by Enrollment in Post-reform Years

Notes: The figure shows actual class size and as predicted by Maimonides' Rule in post-reform years

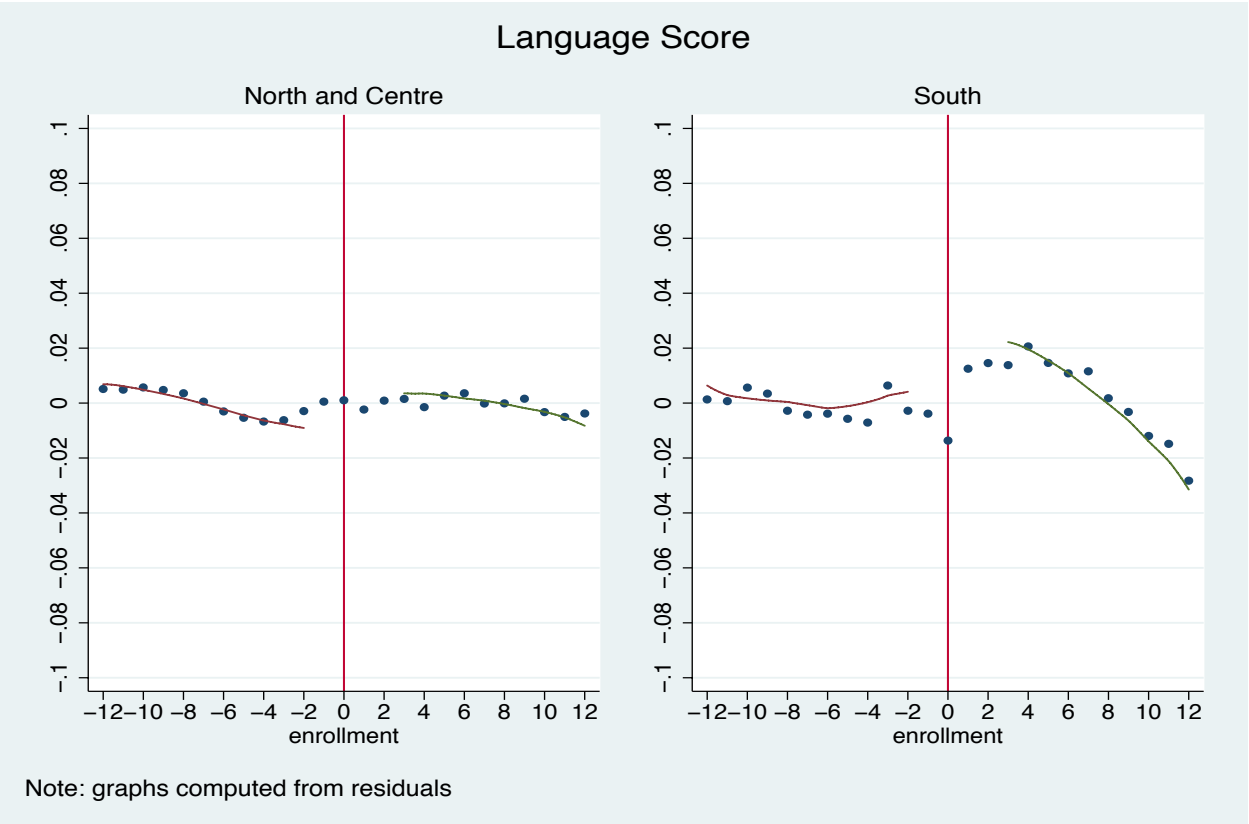
Figure 4: Class Size and Enrollment, centered at Maimonides Cutoffs





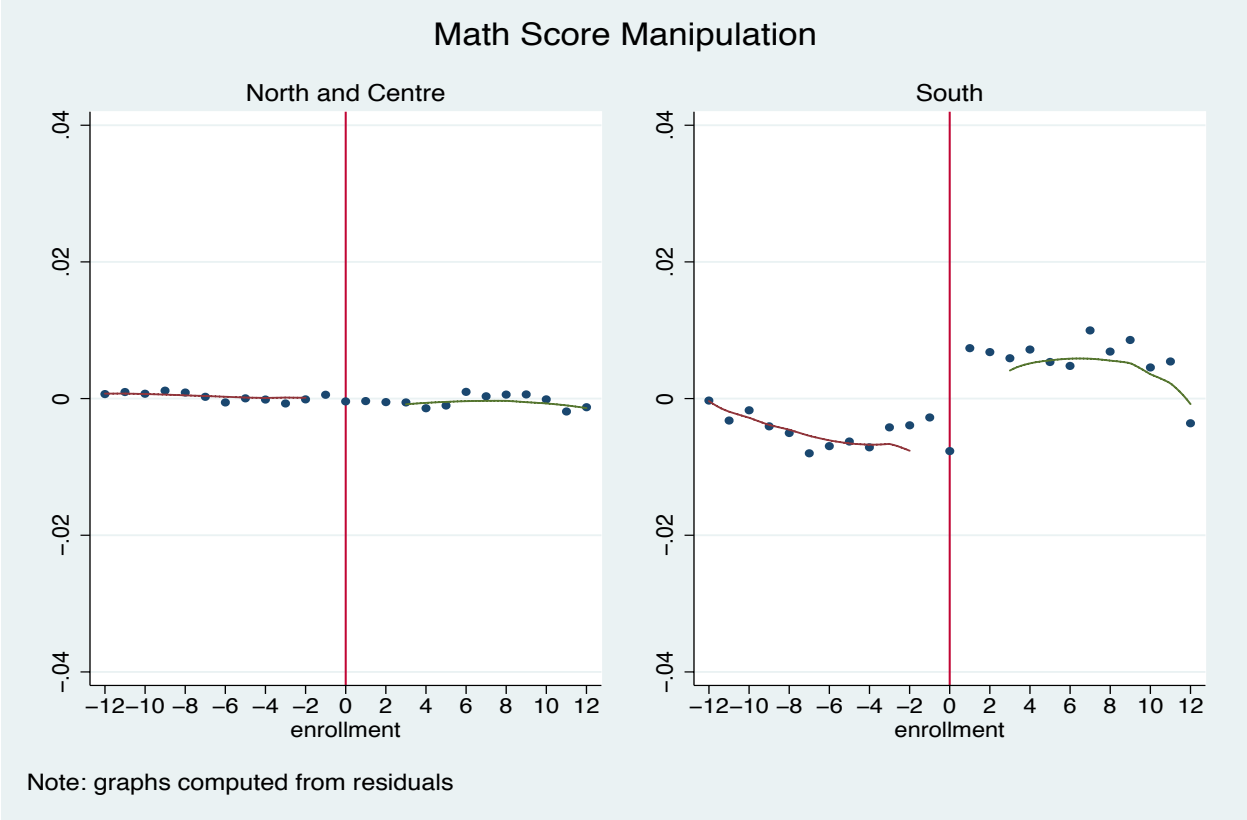
Notes: The solid line shows a one-sided LLR fit.

Figure 5: Test Scores and Enrollment, centered at Maimonides Cutoffs



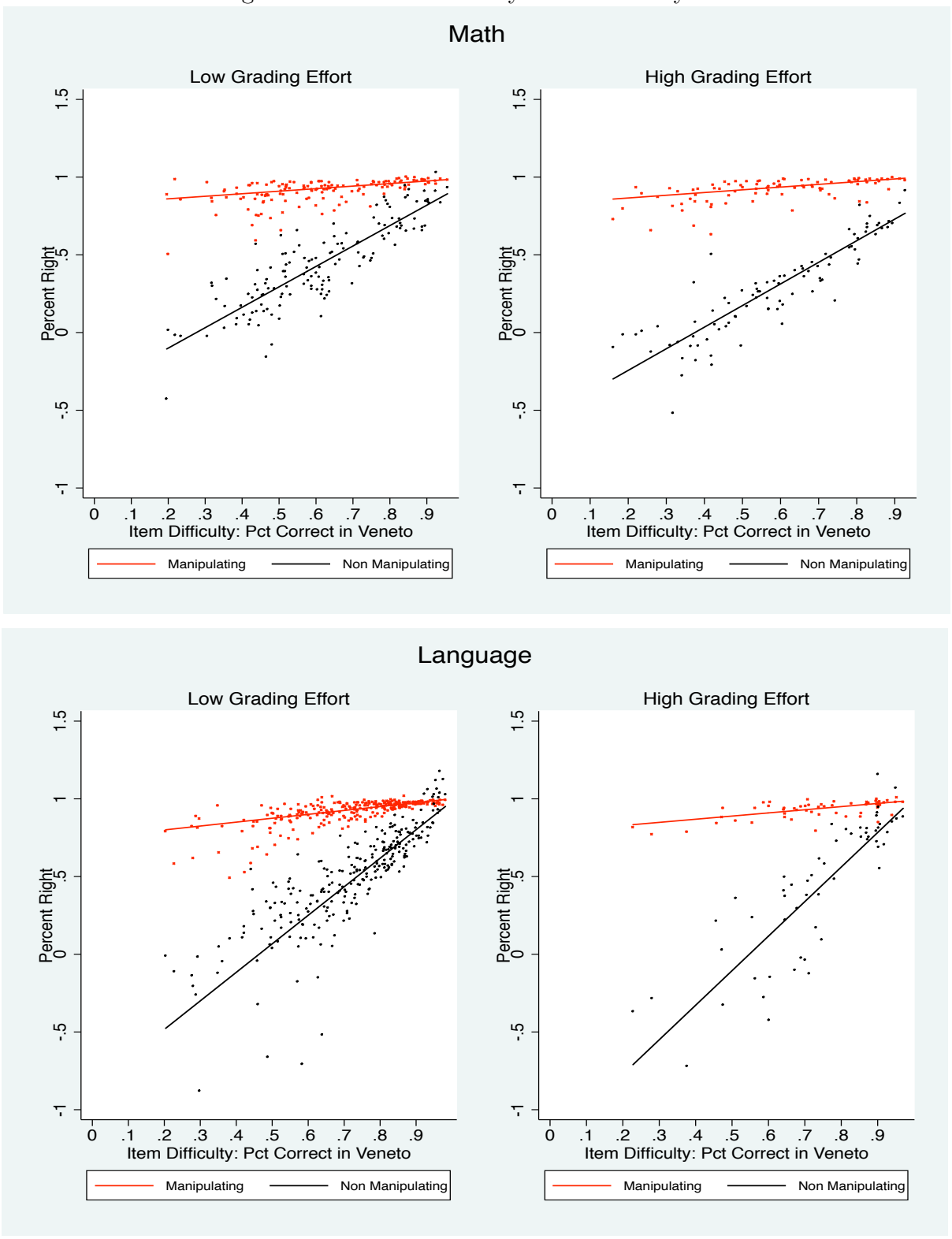
Notes: The solid line shows a one-sided LLR fit.

Figure 6: Score Manipulation and Enrollment, centered at Maimonides Cutoffs



Notes: The solid line shows a one-sided LLR fit.

Figure 7: Score Gradient by Item Difficulty



Notes: The figures plot the average potential score on item j under manipulation for complying classes and the average potential score on item j without manipulation for the same classes against the percent correct answers in monitored institutions in Veneto.The sample is restricted to the South.

References

Abadie, A. (2002): "Bootstrap Tests for Distributional Treatment Effects in Instrumental Variables Models," *Journal of the American Statistical Association*, 97, 284–292.

Angrist, J. D., and V. Lavy (1999): "Using Maimonides' Rule to Estimate the Effect of Class Size on Scholastic Achievement," *Quarterly Journal of Economics*, 114(2), 533–575.

Angrist, J. D., P. Pathak, and C. R. Walters (2013): "Explaining Charter School Effectiveness," *American Economic Journal: Applied Economics*, 5(4), 1–27.

Baker, O., and D. Paserman (2013): "Grade Enrollment Sorting under an Incentives-Based Class Size Reduction Program," Unpublished mimeo.

Ballatore, R., M. Fort, and A. Ichino (2014): "The Tower of Babel in the Classroom: Immigrants and Natives in Italian Schools," IZA Discussion Papers 8732, Institute for the Study of Labor.

Banerjee, A., and E. Duflo (2006): "Addressing Absence," *Journal of Economic Perspectives*, 20(1), 117–132.

Battistin, E., M. De Nadai, and D. Vuri (2014): "Counting Rotten Apples: Student Achievement and Score Manipulation in Italian Elementary Schools," IZA Discussion Papers 8405, Institute for the Study of Labor.

Battistin, E., and L. Neri (2016): "Discretion in Grading and Economic Geography," Queen Mary University of London, Unpublished mimeo.

Bertoni, M., G. Brunello, and L. Rocco (2013): "When the Cat is Near, the Mice Won't Play: The Effect of External Examiners in Italian Schools," *Journal of Public Economics*, 104, 65–77.

Bezdek, J. (1981): *Pattern Recognition with Fuzzy Objective Function Algorithms*. Plenum Press, New York.

Blundell, R., and T. McCurdy (1999): "Labor Supply: A Review of Alternative Approaches," *Handbook of Labor Economics*, 3, 1559–1695.

Böhlmark, A., and M. Lindahl (2012): "Independent Schools and Long-Run Educational Outcomes - Evidence from Sweden's Large Scale Voucher Reform," *Economica*, 82(327).

Bonesronning, H. (2003): "Class Size Effects on Student Achievement in Norway: Patterns and Explanations," *Southern Economic Journal*, 69(4), 952–965.

Bratti, M., D. Checchi, and A. Filippin (2007): "Territorial Differences in Italian Students' Mathematical Competences: Evidence from PISA," *Giornale degli Economisti e Annali di Economia*, 66(3), 299–335.

Brunello, G., and D. Checchi (2005): "School Quality and Family Background in Italy," *Economics of Education Review*, 24, 563–577.

Chaudhury, N., J. Hammer, M. Kremer, K. Muralidharan, and F. H. Rogers (2006): "Missing in Action: Teacher and Health Worker Absence in Developing Countries," *Journal of Economic Perspectives*, 20(1), 91–116.

Chetty, R., J. Friedman, N. Hilger, E. Saez, D. Schanzenbach, and D. Yagan (2011): "How Does Your Kindergarten Classroom Affect Your Earnings? Evidence from Project STAR," *Quarterly Journal of Economics*, 126(4), 1593–1660.

Clotfelter, C. T., H. F. Ladd, and J. L. Vigdor (2009): "Are Teacher Absences Worth Worrying About in the United States?," *Education Finance and Policy*, 4(2), 115–149.

Costantini, M., and C. Lupi (2006): "Divergence and Long-run Equilibria in Italian Regional Unemployment," *Applied Economics Letters*, 13(14), 899–904.

De Paola, M., V. Scoppa, and V. Pupo (2014): "Absenteeism in the Italian Public Sector: The Effects of Changes in Sick Leave Policy," *Journal of Labor Economics*, 32(2), 337–360.

Dee, T. S., B. A. Jacob, J. McCrary, and J. Rockoff (2016): "The Causes and Consequences of Test Score Manipulation: Evidence from the New York Regents Examinations," NBER Working Paper, 22165.

Diamond, R., and P. Persson (2016): "The Long-term Consequences of Teacher Discretion in Grading of High-Stakes Tests," NBER Working Paper, 22207.

Dobbelsteen, S., J. Levin, and H. Oosterbeek (2002): "The Causal Effect of Class Size on Scholastic Achievement: Distinguishing the Pure Class Size Effect from the Effect of Changes in Class Composition," *Oxford Bulletin of Economics and Statistics*, 64(1), 17–38.

Falzetti, P. (2013): "L'esperienza di Restituzione dei Dati al Netto del Cheating," presentation at the Workshop "Metodi di Identificazione, Analisi e Trattamento del Cheating", 8 February, available at: http://www.invalsi.it/invalsi/ri/sis/documenti/022013/falzetti.pdff.

Gary-Bobo, R. J., and M.-B. Mahjoub (2013): "Estimation of Class-Size Effects, Using "Maimonides' Rule" and Other Instruments: the Case of French Junior High Schools," *Annals of Economics and Statistics*, 111/112, 193–255.

Greaves, E., and S. Burgess (2013): "Test Scores, Subjective Assessment and Stereotyping of Ethnic Minorities," *Journal of Labor Economics*, 31, 535–576.

Guiso, L., P. Sapienza, and L. Zingales (2004): "The Role of Social Capital in Financial Development," *American Economic Review*, 94(3), 526–556.

——— (2010): "Civic Capital as the Missing Link," in *Handbook of Social Economics*, ed. by A. B. Jess Benhabib, and M. Jackson. North Holland.

Hanushek, E. A. (1995): "Interpreting Recent Research on Schooling in Developing Countries," *The World Bank Research Observer*, 10(2), 227–246.

Hoxby, C. (2000): "The Effects of Class Size on Student Achievement: New Evidence from Population Variation," *Quarterly Journal of Economics*, 115 (4), 1239–1285.

Ichino, A., and P. Ichino (1997): "Culture, Discrimination and Individual Productivity: Regional Evidence from Personnel Data in a Large Italian Firm," CEPR Discussion Papers 1709.

Ichino, A., and G. Maggi (2000): "Work Environment and Individual Background: Explaining Regional Shirking Differentials in a Large Italian Firm," *Quarterly Journal of Economics,*, 115(3), 933–959.

Ichino, A., and G. Tabellini (2014): "Freeing the Italian School System," *Labour Economics*, 30, 113–128.

Imbens, G., and K. Kalyanaraman (2012): "Optimal Bandwidth Choice for the Regression Discontinuity Estimator," *Review of Economic Studies*, 79(3), 933–959.

INVALSI (2010): "Sistema Nazionale di Valutazione - A.S. 2009/2010, Rilevazione degli apprendimenti," *Technical Report*.

Jacob, B., and S. Levitt (2003): "Rotten Apples: An Investigation of the Prevalence and Predictors of Teacher Cheating," *Quarterly Journal of Economics*, 118(3), 843–77.

Kane, T. J., C. E. Rouse, and D. Staiger (1999): "Estimating Returns to Schooling When Schooling is Misreported," NBER Working Paper 7235.

Krueger, A. (1999): "Experimental Estimates of Education Production Functions," *Quarterly Journal of Economics*, 114, 497–532.

Lavy, V. (2008): "Do Gender Stereotypes Reduce Girls' or Boys' Human Capital Outcomes? Evidence from a Natural Experiment," *Journal of Public Economics*, 92(10-11), 2083–2105.

Lavy, V., and E. Sand (2015): "On The Origins of the Gender Human Capital Gap: Short and Long Term Effect of Teachers' Stereotypes," NBER Working paper, No. 20909.

Leuven, E., H. Oosterbeek, and M. Ronning (2008): "Quasi-Experimental Estimates of the Effect of Class Size Achievement in Norway," *The Scandinavian Journal of Economics*, 110(4), 663–693.

Lewbel, A. (2007): "Estimation of Average Treatment Effects with Misclassification," *Econometrica*, 2(3), 537–551.

Mahajan, A. (2006): "Identification and Estimation of Regression Models with Misclassification," *Econometrica*, 74(3), 631–665.

Neal, D. (2013): "The Consequences of Using One Assessment System to Pursue Two Objectives," *The Journal of Economic Education*, 44(4), 339–352.

PIKETTY, T. (2004): "Should We Reduce Class Size or School Segregation? Theory and Evidence from France," presentation at the Roy Seminars, Association pour le Dévelopement de la Recherche en Économie et en Statistique (ADRES), 22 November, available at: http://www.adres.polytechnique.fr/SEMINAIRE/221104b.pdf.

PUTNAM, R., R. LEONARDI, AND R. NANETTI (1993): *Making Democracy Work*. Princeton University Press, Princeton.

QUINTANO, C., R. CASTELLANO, AND S. LONGOBARDI (2009): "A Fuzzy Clustering Approach to Improve the Accuracy of Italian Student Data. An Experimental Procedure to Correct the Impact of the Outliers on Assessment Test Scores," *Statistica & Applicazioni*, Vol.VII(2), 149–171.

SEVERSON, K. (2011): "Systematic Cheating Is Found in Atlanta's School System," *New York Times, July 11*, Accessed at: http://www.nytimes.com/2011/07/06/education/06atlanta.html.

SIMS, D. (2008): "A Strategic Response to Class Size Reduction: Combination Classes and Student Achievement in California," *Journal of Policy Analysis and Management*, 27(3), 457–478.

TERRIER, C. (2015): "Boys Lag Behind: How Teachers' Gender Biases Affect Student Achievement," CEP Working Paper No 1341, London School of Economics.

URQUIOLA, M., AND E. VERHOOGEN (2009): "Class Size Caps, Sorting, and the Regression Discontinuity Design," *American Economic Review*, 99(1), 179–215.

WOESSMANN, L. (2005): "Educational Production in Europe," *Economic Policy*, 43, 445–493.

Appendix (for online publication)

Score Manipulation Imputation

Our imputation is closely related to that used by INVALSI and described in Quintano et al. (2009). INVALSI assigns a manipulation probability to each class in three steps.

The first step computes the following four summary statistics.

(1) Within-class average score

$$\bar{p}_i = \frac{\sum_{j=1}^{N_i} p_{ji}}{N_i}, \tag{10}$$

where p_{ji} denotes the score of student j in class i; N_i denotes the number of test-takers in class i.

(2) Within-class standard deviation of scores

$$\sigma_i = \sqrt{\frac{\sum_{j=1}^{N_i} (p_{ji} - \bar{p}_i)^2}{N_i}}. \tag{11}$$

(3) Within-class average percent missing

$$MC_i = \frac{\sum_{j=1}^{N_i} M_{ji}}{N_i}, \tag{12}$$

where M_{ji} is the fraction of test items skipped by student j in class i.

(4) Within-class index of answer homogeneity

$$\bar{E}_i = \frac{\sum_{q=1}^{Q} E_{qi}}{Q}, \tag{13}$$

where $q = 1, .., Q$ indexes test items and E_{qi} is a Gini measure of homogeneity that equals value zero if all students in class i provide the same answer to item q. This can be interpreted as the Herfindahl index of the share of students with similar response patterns in the class.

In the second step, the first two principal components are extracted from the 4×4 correlation matrix determined by these indicators, yielding a percentage of explained variance which is - across years, subjects and grades - well above 90%. Denote these principal com-

ponents by ψ_{1i} and ψ_{2i}. The third step consists of a cluster analysis that creates G groups from the distribution of (ψ_{1i}, ψ_{2i}). INVALSI sets $G = 8$, yielding a matrix whose elements are, for each class, eight group membership probabilities. This procedure is known as "fuzzy clustering" (see Bezdek, 1981), since data elements (classes, in our setting) can be assigned to one or more groups. With "hard clustering", data elements belong to exactly one cluster.

INVALSI identifies likely manipulators as those in the group with values of (ψ_{1i}, ψ_{2i}) that are most extreme (see Figure 8 in Quintano et al. 2009). In practice, the suspicious group is characterized by (i) abnormally large values of $\bar{p}_i$, and (ii) small values of σ_i, MC_i and $\bar{E}_i$, relative to the population average of these indicators. This group is flagged as the "outlier" or manipulating cluster. The INVALSI manipulation indicator gives, for each class, the membership probability for this cluster. Our hard clustering computations codes a dummy for manipulating classes. This dummy indicates classes whose values of (ψ_{1i}, ψ_{2i}) belong to the manipulating cluster identified by INVALSI.

Manipulation and Class Size

Class size is denoted by s and, in the absence of manipulation, the score on item j is $L_j \in [0, 1]$. Manipulated scores are equal to 1. The manipulated class average score is therefore $y_j = (1 - L_j)p_j + L_j$, where $p_j = \frac{n_j}{s}$ is the fraction of score sheets manipulated for item j. The score gain from manipulation is $\widetilde{p}_j \equiv \tau_j p_j$, with $\tau_j = 1 - L_j \geq 0$. Large τ_j denotes difficult items, so the returns to manipulation vary with item difficulty. Probability of exposure cumulates across items, $\gamma \sum_j n_j$, where n_j is number of score sheets manipulated for item j. Assuming additively separable across items utility, teachers have the following objective function (assuming utility is zero if caught)

$$\underbrace{\left(1 - \gamma \sum_j n_j\right)}_{disclosure\ risk} \underbrace{U\left(\sum_j \widetilde{p}_j\right)}_{utility\ of\ score\ gain} - \underbrace{\beta \sum_j (s - n_j)}_{honest\ grading\ effort} .$$

Divide by s to write the problem as

$$\max_{\widetilde{\mathbf{p}}} \left(\frac{1}{s} - \gamma \sum_j \frac{\widetilde{p}_j}{\tau_j}\right) U\left(\sum_j \widetilde{p}_j\right) - \beta \sum_j \left(1 - \frac{1}{\tau_j}\widetilde{p}_j\right).$$

The first order condition for optimal $\widetilde{p}_j$ yields

$$\left(\frac{1}{s}-\gamma\sum_j\frac{\widetilde{p}_j}{\tau_j}\right)h\left(\sum_j\widetilde{p}_j\right)+\frac{\beta}{\tau_j}g\left(\sum_j\widetilde{p}_j\right)-\frac{\gamma}{\tau_j}=0, \tag{14}$$

where $g(p)=\frac{1}{U(p)}>0$ and $h(p)=\frac{U'(p)}{U(p)}>0$. Using $\frac{g'(p)}{h(p)}=-g(p)$, comparative statics implies

$$\frac{d\widetilde{p}_j}{ds}=\frac{\tau_j}{s}\left[\tau_j\left(1-\gamma\sum_j n_j\right)\frac{h'\left(\sum_j\widetilde{p}_j\right)}{h\left(\sum_j\widetilde{p}_j\right)}-\beta sg\left(\sum_j\widetilde{p}_j\right)-\gamma s\right]^{-1}.$$

This is negative if

$$\frac{h'\left(\sum_j\widetilde{p}_j\right)}{h\left(\sum_j\widetilde{p}_j\right)}\leq\frac{s}{\tau_j}\left(\beta g\left(\sum_j\widetilde{p}_j\right)+\gamma\right)\left(1-\gamma\sum_j n\right)^{-1},$$

which is more likely to hold in large classes. A sufficient condition is $h'(p)<0$, a diminishing condition for marginal log-utility. With commonly used log-linear preferences we have

$$\frac{h'\left(\sum_j\widetilde{p}_j\right)}{h\left(\sum_j\widetilde{p}_j\right)}=-\frac{1}{\sum_j\widetilde{p}_j},$$

and

$$\frac{d\widetilde{p}_j}{ds}=-\frac{\tau_j}{s}\left[\tau_j\left(1-\gamma\sum_j n_j\right)\frac{1}{\sum_j\widetilde{p}_j}+\beta sg\left(\sum_j\widetilde{p}_j\right)+\gamma s\right]^{-1},$$

which is clearly negative. If $\tau_j=0$, then $\frac{d\widetilde{p}_j}{ds}=0$. If τ_j grows, the class size gradient also depends on γ. This can be used to obtain the score gradient

$$\frac{dy_j}{ds}=\frac{d\left[(1-L_j)p_j+L_j\right]}{ds}=\frac{d\widetilde{p}_j}{ds}.$$

We can also write

$$\frac{dp_j}{ds}=\frac{1}{\tau_j}\frac{d\widetilde{p}_j}{ds}=-\frac{1}{s}\left[\tau_j\left(1-\gamma\sum_j n_j\right)\frac{1}{\sum_j\widetilde{p}_j}+\beta sg\left(\sum_j\widetilde{p}_j\right)+\gamma s\right]^{-1},$$

which shows that, keeping $\sum_j\widetilde{p}_j$ constant, the effect of class size on manipulation rates flattens as item difficulty increases.

Finally, we consider the quantity $\frac{dp_j}{ds}$ for teachers motivated solely by grading effort. When teachers care little about measured achievement, the utility of overall exam performance is constant. In this case, the first order condition in equation (14) becomes

$$\frac{\beta}{\tau_j}g\left(\sum_j\widetilde{p}_j\right)-\frac{\gamma}{\tau_j}=0,$$

and we have that

$$\frac{dp_j}{ds} = -\frac{1}{s^2}\left[\beta g\left(\sum_j \widetilde{p}_j\right) + \gamma\right]^{-1}.$$

This suggests that in a curbstoning scenario, item-level manipulation rates should decrease similarly as class size increases.

Table A1: Reduced Form Estimates of the Effect of Maimonides' Rule on Class Size, Test Scores, and Score Manipulation

	Math			Language		
	Italy	North/Centre	South	Italy	North/Centre	South
	(1)	(2)	(3)	(4)	(5)	(6)
		A. Class size				
Maimonides' Rule	0.513***	0.555***	0.433***			
	(0.006)	(0.008)	(0.011)			
Means	19.88	20.07	19.58			
(sd)	(3.58)	(3.52)	(3.64)			
N	140,010	87,498	52,512			
			B. Test Scores			
Maimonides' Rule	-0.0031***	-0.0023**	-0.0056**	-0.0021***	-0.0012	-0.0041**
	(0.0010)	(0.0009)	(0.0022)	(0.0008)	(0.0008)	(0.0017)
Means	0.007	-0.074	0.141	0.01	-0.005	0.035
(sd)	(0.637)	(0.502)	(0.796)	(0.523)	(0.428)	(0.649)
N	140,010	87,498	52,512	140,010	87,498	52,512
			C. Score Manipulation			
Maimonides' Rule	-0.0009***	-0.0003	-0.0020**	-0.0008**	-0.0003	-0.0016**
	(0.0004)	(0.0002)	(0.0009)	-0.0003	-0.0003	-0.0008
Means	0.065	0.02	0.139	0.055	0.023	0.110
(sd)	(0.246)	(0.139)	(0.346)	(0.229)	(0.149)	(0.313)
N	139,996	87,491	52,505	140,003	87,493	52,510

Notes: This table shows the reduced form effect of the Maimonides' Rule on class size (Panel A), test scores (Panel B), score manipulation (Panel C). All models control for a quadratic in grade enrollment, segment dummies and their interactions. The unit of observation is the class. Robust standard errors, clustered on school and grade, are shown in parentheses. Control variables include: % female students, % immigrants, % fathers at least high school graduate, % employed mothers, % unemployed mothers, % mother NILF grade and year dummies, and dummies for missing values in these variables. All regressions include sampling strata controls (grade enrollment at institution, region dummies and their interactions). * significant at 10%; ** significant at 5%; *** significant at 1%.

Table A2: First Stage Estimates for Over-Identified Models

	Class size			Score manipulation math			Score manipulation language		
	Italy	North/Centre	South	Italy	North/Centre	South	Italy	North/Centre	South
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Maimonides' Rule (f_{igkt})	0.704***	0.753***	0.617***	-0.0009**	-0.0003	-0.0021*	-0.0014***	-0.0008**	-0.0024**
	(0.0059)	(0.0069)	(0.0107)	(0.0005)	(0.0003)	(0.0011)	(0.0004)	(0.0003)	(0.0010)
Monitor at institution (M_{igkt})	0.010	0.029	-0.013	-0.029***	-0.010***	-0.062***	-0.025***	-0.012***	-0.047***
	(0.023)	(0.026)	(0.044)	(0.002)	(0.001)	(0.004)	(0.002)	(0.001)	(0.004)
2 students below cutoff	-1.427***	-1.154***	-1.865***	0.002	-0.002	0.008	0.010**	0.005	0.018
	(0.083)	(0.101)	(0.138)	(0.005)	(0.003)	(0.012)	(0.005)	(0.004)	(0.011)
1 student below cutoff	-2.258***	-2.053***	-2.580***	0.001	0.001	0.000	0.007	0.009**	0.002
	(0.093)	(0.116)	(0.150)	(0.005)	(0.004)	(0.012)	(0.005)	(0.004)	(0.011)
1 student above cutoff	2.411***	3.026***	1.519***	0.000	0.003	-0.004	-0.001	-0.001	-0.001
	(0.097)	(0.132)	(0.138)	(0.006)	(0.005)	(0.013)	(0.005)	(0.004)	(0.012)
2 students above cutoff	1.247***	1.546***	0.826***	0.001	-0.004	0.007	-0.007	-0.005	-0.012
	(0.083)	(0.114)	(0.120)	(0.006)	(0.004)	(0.013)	(0.005)	(0.004)	(0.009)
N	140,010	87,498	52,512	139,996	87,491	52,505	140,003	87,493	52,510

Notes: Columns 1-3 report first stage estimates of the effect of the Maimonides' Rule, a monitor at institution and dummies for grade enrollment being in a 10 percent window below and above each cutoff on class size. Columns 4-9 show first stage estimates of the effect of the Maimonides' Rule, a monitor at institution and dummies for grade enrollment being in a 10 percent window (2 students) above and below each cutoff on score manipulation. All models control for a quadratic in grade enrollment, segment dummies and their interactions. The unit of observation is the class. Robust standard errors, clustered on school and grade, are shown in parentheses. Control variables include: % female students, % immigrants, % fathers at least high school graduate, % employed mothers, % unemployed mothers, % mother NILF, grade and year dummies, and dummies for missing values in these variables. All regressions include sampling strata controls (grade enrollment at institution, region dummies and their interactions). * significant at 10%; ** significant at 5%; *** significant at 1%.

Table A3: Covariates and Maimonides' Rule with and without External Monitors

	Institutions with Monitor			Institutions without Monitor		
	Italy	North/Centre	South	Italy	North/Centre	South
	(1)	(2)	(3)	(4)	(5)	(6)
			A. Administrative Data on Schools			
% in class sitting the test	0.0001	0.0002	0.0000	0.0000	0.0000	0.0000
	(0.0002)	(0.0002)	(0.0003)	(0.0001)	(0.0001)	(0.0002)
% in school sitting the test	0.0003	0.0003	0.0002	0.0001	0.0001	0.0001
	(0.0002)	(0.0002)	(0.0003)	(0.0001)	(0.0001)	(0.0002)
% in institution sitting the test	-0.0000	-0.0000	0.0001	-0.0001*	-0.0002*	-0.0000
	(0.0001)	(0.0002)	(0.0003)	(0.0001)	(0.0001)	(0.0001)
			B. Data Provided by School Staff			
Female	-0.0003	-0.0006	0.0001	0.0001	0.0005*	-0.0003
	(0.0003)	(0.0004)	(0.0006)	(0.0002)	(0.0002)	(0.0003)
Immigrant	-0.0005	-0.0002	-0.0007**	-0.0007***	-0.0009***	-0.0003*
	(0.0003)	(0.0005)	(0.0003)	(0.0002)	(0.0003)	(0.0002)
Father HS	-0.0005	-0.0002	-0.0014	0.0010***	0.0003	0.0020***
	(0.0005)	(0.0006)	(0.0010)	(0.0003)	(0.0004)	(0.0005)
Mother employed	0.0001	0.0003	-0.0004	0.0015***	0.0012**	0.0022***
	(0.0008)	(0.0010)	(0.0012)	(0.0004)	(0.0006)	(0.0006)
			C. Non-Response Indicators			
Missing data on father's education	0.0014	0.0012	0.0019	0.0000	0.0016**	-0.0026**
	(0.0011)	(0.0013)	(0.0020)	(0.0007)	(0.0008)	(0.0012)
Missing data on mother's occupation	0.0018*	0.0017	0.0020	-0.0002	0.0012	-0.0028**
	(0.0011)	(0.0013)	(0.0019)	(0.0007)	(0.0008)	(0.0011)
Missing data on country of origin	0.0006	0.0003	0.0011	-0.0002	-0.0002	-0.0003
	(0.0004)	(0.0004)	(0.0008)	(0.0003)	(0.0003)	(0.0006)
N	34,325	22,174	12,151	105,685	65,324	40,361

Notes: This table reports coefficients from regressions of the variables listed at left on Maimonides' Rule, controlling for a quadratic in grade enrollment, enrollment segment dummies and their interactions, grade and year dummies, and sampling strata controls (grade enrollment at institution, region dummies and their interactions). Columns 1-3 show results for the sample with monitors; columns 4-6 show results for the sample without monitors. Robust standard errors, clustered on school and grade, are shown in parentheses. * significant at 10%; ** significant at 5%; *** significant at 1%.

Figure A1: Answer sheet for V grade in 2010/11

Servizio Nazionale di Valutazione a.s. 2010/11

CLASSE:

Scheda Risposte Studente n°

Risultati delle prove

Codice istituto: Codice Scuola:

Codice plesso: Livello:

Codice Classe: **NON CAMPIONE**

Codice studente: Numero progressivo studente:

PROVA ITALIANO (1)

A1	☐A	☐B	☐C	☐D	☐NV	**C1_a1**	☐0	☐1	☐NV		
A2	☐A	☐B	☐C	☐D	☐NV	**C1_a2**	☐0	☐1	☐NV		
A3	☐A	☐B	☐C	☐D	☐NV	**C1_b1**	☐0	☐1	☐NV		
A4	☐A	☐B	☐C	☐D	☐NV	**C1_b2**	☐0	☐1	☐NV		
A5	☐A	☐B	☐C	☐D	☐NV	**C1_b3**	☐0	☐1	☐NV		
A6	☐A	☐B	☐C	☐D	☐NV	**C2**	☐A	☐B	☐C	☐D	☐NV
A7	☐A	☐B	☐C	☐D	☐NV	**C3_a**	☐Nome		☐Non_Nome		☐NV
A8	☐A	☐B	☐C	☐D	☐NV	**C3_b**	☐Nome		☐Non_Nome		☐NV
A9	☐A	☐B	☐C	☐D	☐NV	**C3_c**	☐Nome		☐Non_Nome		☐NV
A10	☐A	☐B	☐C	☐D	☐NV	**C3_d**	☐Nome		☐Non_Nome		☐NV
A11	☐A	☐B	☐C	☐D	☐NV	**C3_e**	☐Nome		☐Non_Nome		☐NV
A12	☐A	☐B	☐C	☐D	☐NV	**C3_f**	☐Nome		☐Non_Nome		☐NV
A13	☐A	☐B	☐C	☐D	☐NV	**C3_g**	☐Nome		☐Non_Nome		☐NV
A14	☐A	☐B	☐C	☐D	☐NV	**C3_h**	☐Nome		☐Non_Nome		☐NV
A15	☐A	☐B	☐C	☐D	☐NV	**C3_i**	☐Nome		☐Non_Nome		☐NV
A16	☐A	☐B	☐C	☐D	☐NV	**C3_l**	☐Nome		☐Non_Nome		☐NV
A17	☐A	☐B	☐C	☐D	☐NV	**C3_m**	☐Nome		☐Non_Nome		☐NV
B1	☐A	☐B	☐C	☐D	☐NV	**C3_n**	☐Nome		☐Non_Nome		☐NV
B2	☐A	☐B	☐C	☐D	☐NV	**C3_o**	☐Nome		☐Non_Nome		☐NV
B3	☐A	☐B	☐C	☐D	☐NV	**C3_p**	☐Nome		☐Non_Nome		☐NV
B4	☐A	☐B	☐C	☐D	☐NV	**C3_q**	☐Nome		☐Non_Nome		☐NV
B5	☐A	☐B	☐C	☐D	☐NV	**C3_r**	☐Nome		☐Non_Nome		☐NV
B6	☐A	☐B	☐C	☐D	☐NV	**C4**	☐0	☐1	☐NV		
B7	☐A	☐B	☐C	☐D	☐NV	**C5**	☐0	☐1	☐NV		
B8	☐A	☐B	☐C	☐D	☐NV	**C6**	☐A	☐B	☐C	☐D	☐NV
B9	☐A	☐B	☐C	☐D	☐NV	**C7**	☐0	☐1	☐NV		
B10	☐A	☐B	☐C	☐D	☐NV	**C8**	☐A	☐B	☐C	☐D	☐NV
B11	☐A	☐B	☐C	☐D	☐NV	**C9**	☐A	☐B	☐C	☐D	☐NV
B12	☐A	☐B	☐C	☐D	☐NV	**C10**	☐0	☐1	☐NV		
B13	☐A	☐B	☐C	☐D	☐NV						
B14	☐A	☐B	☐C	☐D	☐NV						
B15	☐A	☐B	☐C	☐D	☐NV						

PROVA MATEMATICA (1)

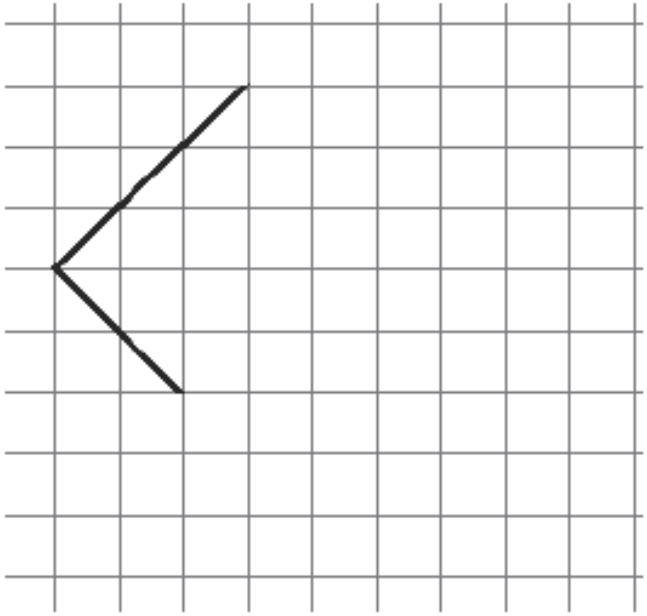
D1_a	☐V	☐F	☐NV		
D1_b	☐V	☐F	☐NV		
D1_c	☐V	☐F	☐NV		
D1_d	☐V	☐F	☐NV		
D2	☐A	☐B	☐C	☐D	☐NV
D3	☐0	☐1	☐NV		
D4_a	☐A	☐B	☐C	☐D	☐NV
D4_b	☐0	☐1	☐NV		
D5	☐A	☐B	☐C	☐D	☐NV
D6	☐A	☐B	☐C	☐D	☐NV
D7	☐A	☐B	☐C	☐D	☐NV
D8	☐A	☐B	☐C	☐D	☐NV
D9	☐0	☐1	☐NV		
D10	☐A	☐B	☐C	☐D	☐NV
D11	☐A	☐B	☐C	☐D	☐NV
D12	☐0	☐1	☐NV		
D13	☐A	☐B	☐C	☐D	☐NV
D14	☐A	☐B	☐C	☐D	☐NV
D15	☐A	☐B	☐C	☐D	☐NV
D16_a	☐0	☐1	☐NV		
D16_b	☐0	☐1	☐NV		
D17_a	☐V	☐F	☐NV		
D17_b	☐V	☐F	☐NV		
D17_c	☐V	☐F	☐NV		
D17_d	☐V	☐F	☐NV		
D18	☐A	☐B	☐C	☐D	☐NV
D19	☐0	☐1	☐NV		
D20	☐A	☐B	☐C	☐D	☐NV
D21_a	☐0	☐1	☐NV		
D21_b	☐0	☐1	☐NV		
D22	☐A	☐B	☐C	☐D	☐NV
D23_a	☐0	☐1	☐NV		
D23_b	☐0	☐1	☐NV		
D24_a	☐0	☐1	☐NV		
D24_b	☐0	☐1	☐NV		
D24_c	☐0	☐1	☐NV		
D25	☐0	☐1	☐NV		
D26	☐A	☐B	☐C	☐D	☐NV
D27	☐A	☐B	☐C	☐D	☐NV
D28_a	☐km	☐m	☐cm	☐mm	☐NV
D28_b	☐km	☐m	☐cm	☐mm	☐NV
D28_c	☐km	☐m	☐cm	☐mm	☐NV
D29_a	☐V	☐F	☐NV		
D29_b	☐V	☐F	☐NV		
D29_c	☐V	☐F	☐NV		
D29_d	☐V	☐F	☐NV		
D30	☐A	☐B	☐C	☐D	☐NV

(1) Barrare NV per risposta non valida (2 risposte o risposta incomprensibile) e non barrare nulla in caso di risposta omessa

(**ATTENZIONE Non spillare, non modificare per nessun motivo i dati precompilati della scheda)**

Figure A2: Example of open-ended question in math test - V grade 2010/11

D23. Osserva la seguente figura.



a. Completa la figura in modo da ottenere un quadrato.

b. Spiega come hai fatto per disegnare il quadrato.

.....

.....

.....

Figure A3: Example of open-ended question in language test - V grade 2010/11

C4. Nella frase che segue inserisci le parole mancanti scegliendole da questa lista: *così, dove, perché, però, se, siccome.*

……. non conoscevo la strada, ho chiesto a una signora ……. dovevo andare; …….. non mi sono perso.