American Economic Journal: Applied Economics 2017, 9(4): 216–249
https://doi.org/10.1257/app.20160267

In a Small Moment: Class Size and Moral Hazard in the Italian Mezzogiorno[†]

By Joshua D. Angrist, Erich Battistin, and Daniela Vuri*

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 a 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 illustrate how consequential score manipulation can arise even in assessment systems with few accountability concerns. (*JEL* D82, H75, I21, I26, I28, J24, R23)

School improvement efforts often focus on inputs to education production, the most important of which is staffing ratios. Parents, teachers, and policymakers 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

* Angrist: MIT Department of Economics, 50 Memorial Drive, Cambridge, MA 02142, IZA and NBER (email: angrist@mit.edu); Battistin: School of Economics and Finance, Queen Mary University of London, Mile End Road, London E1 4NS, CEPR, FBK-IRVAPP, and IZA (email: e.battistin@qmul.ac.uk); Vuri: Department of Economics and Finance, University of Rome Tor Vergata, via Columbia 2, 00133 Rome, Italy, CEIS, CESIfo and IZA (email: daniela.vuri@uniroma2.it). 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, Lorenzo Rocco, and Jonah Rockoff for helpful comments and discussion, and to three anonymous referees and 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.

[†] Go to https://doi.org/10.1257/app.20160267 to visit the article page for additional materials and author disclosure statement(s) or to comment in the online discussion forum.

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) shows 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 2015, Diamond and Persson 2016) and in the United Kingdom and Israel, where important nationally administered tests are locally marked.[1] In public school systems with few or no employee performance standards, such as the Italian public school system studied here, lack of accountability or oversight may drive low fidelity to testing protocols.

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 size of 40 in a grade cohort of 40, while with an 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-graders 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 the effect of class size on learning by socioeconomic status 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

[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 (2017) for evidence on UK manipulation. Lavy (2008) documents gender bias in the local grading of Israel's matriculation exams. De Paola, Scoppa, and Pupo (2014) estimates the effects of workplace accountability on productivity in the Italian public sector. Ichino and Tabellini (2014) discusses 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, Apulia, Abruzzo, Molise, and the islands of Sicily and Sardinia. Italy's 20 administrative regions are further divided into over 100 provinces.

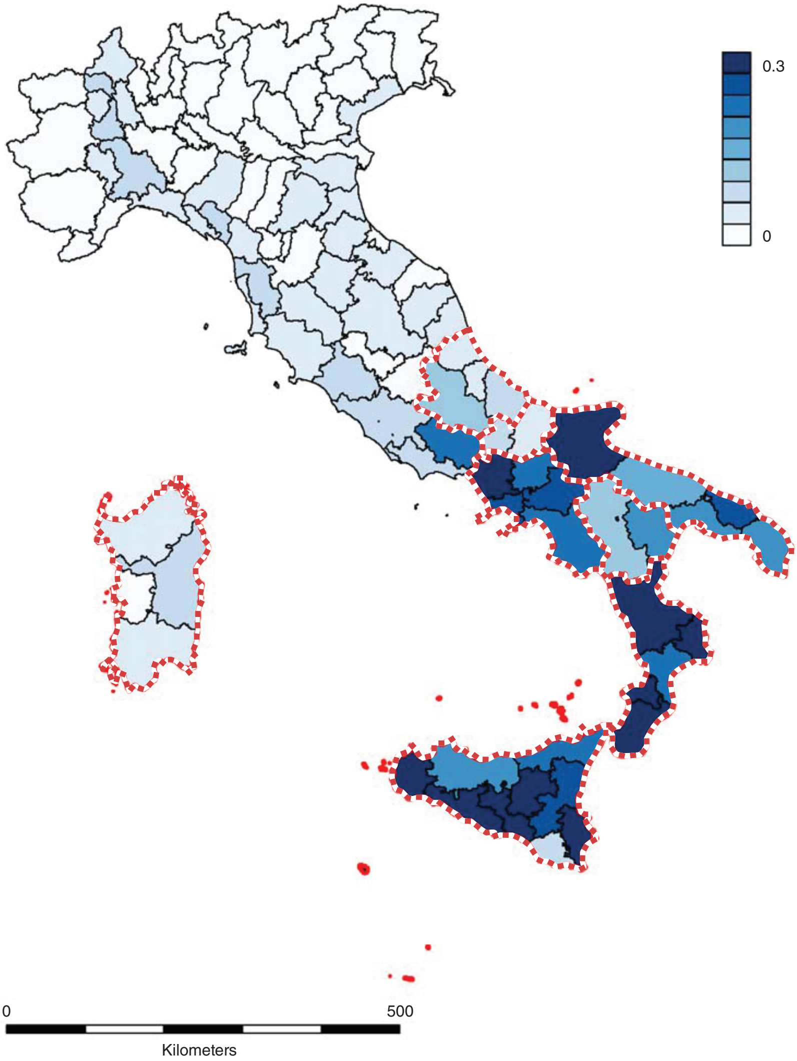

Figure 1. Manipulation Rates by Province

Note: Mezzogiorno regions are bordered with dashed lines.

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 way, roughly 5 percent of

[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

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, below) and reaches 25 percent in some provinces. Further evidence suggesting extensive score manipulation in the South comes from Bertoni, Brunello, and Rocco (2013), which analyzes 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 handwritten 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 student 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 policymakers and to researchers studying the causal effects of school inputs. The Maimonides' Rule research design is not guaranteed to work. Urquiola and Verhoogen (2009) shows 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 inherent in analyses of the causal effects of class size, a problem that arises independently of 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

for each (see Quintano, Castellano, and Longobardi 2009). Figure 1 uses this variable for the 2009–2011 scores of second-graders and fifth-graders.

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, Ladd, and Vigdor (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 II describes our data and documents the Maimonides' Rule first stage. Following a brief graphical analysis, Section III reports Maimonides-style estimates of effects of class size on achievement and score manipulation. Section IV explores the nature of score manipulation by linking score distributions and response patterns with class size and item difficulty. Section V outlines a model of grading behavior, which provides a link between teacher manipulation and class size. Finally, Section VI uses the monitoring experiment and Maimonides' Rule to jointly estimate class size and manipulation effects. This section also reviews possible threats to validity in our research design. Section VII concludes.

I. Background and Context

A. *Italian Schools and Tests*

Primary schooling (*scuola elementare*) in Italy is compulsory from ages 6 to 11. Schools are administrated as single-unit 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 oversubscription, 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 as 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 online 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 enlist 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 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 hours, 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 percent 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 past two years in the towns or at the schools they are assigned to monitor. Monitors are paid for their work and are required to complete transcription by the end of the test day.

B. *Related Work*

Maimonides-style empirical strategies have been used to identify class-size effects in many countries, including the United States (Hoxby 2000), France (Piketty 2004 and Gary-Bobo and Mahjoub 2013), Norway (Bonesrønning 2003; and Leuven, Oosterbeek, and Rønning 2008), and the Netherlands (Dobbelsteen, Levin, and Oosterbeek 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, Wößmann (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. More recent

[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).

estimates for Israel, reported in Angrist et al. (2017), also show no gains from class-size reductions. Results in Sims (2008) suggest class-size reductions obtained through combination classes have 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, Checchi, and Filippin (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, Fort, and Ichino (forthcoming) uses 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, Brunello, and Rocco (2013), which focuses 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 that are not fully accounted for in earlier work.

A second 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, Sapienza, and Zingales 2004, 2011) 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, Leonardi, and Nanetti 1993). Finally, as noted in the introduction, our work connects with research on teacher shirking around the world.

II. Data and First Stage

A. *Data and Descriptive Statistics*

The standardized test score data used in this study come from INVALSI's testing program conducted in Italian elementary schools in the 2009–2010, 2010–2011, and 2011–2012 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 meant 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).

Table 1—Descriptive Statistics

	Grade 2			Grade 5		
	Italy	North/ Center	South	Italy	North/ Center	South
	(1)	(2)	(3)	(4)	(5)	(6)
Panel A. Class characteristics						
Female[a]	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[a]	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[a]	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[a]	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
Panel 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
Panel 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. Data are from the 2009 to 2010, 2010 to 2011, and 2011 to 2012 school years. Standard deviations are reported in parentheses.

[a] Conditional on non-missing survey response

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 ninety-ninth 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. These 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.

B. *Maimonides in Italy*

Our identification strategy for class-size effects exploits minimum and maximum class sizes (these rules are laid out in a regulation known as *Decreto Ministeriale 331/98*). Until the 2008 school year, primary school class sizes were restricted 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 percent 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 percent (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 entering in 2009 and fifth graders in any year were subject to the old rule, while second graders entering 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 nonlinear 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

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

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 $\operatorname{int}(x)$ is the largest integer smaller than or equal to x. Figures 2 and 3 plot average class size and f_{igkt} against enrollment in each grade, separately for pre-reform and post-reform periods. Plotted points show the average actual class size at each level of enrollment. Actual class size follows predicted

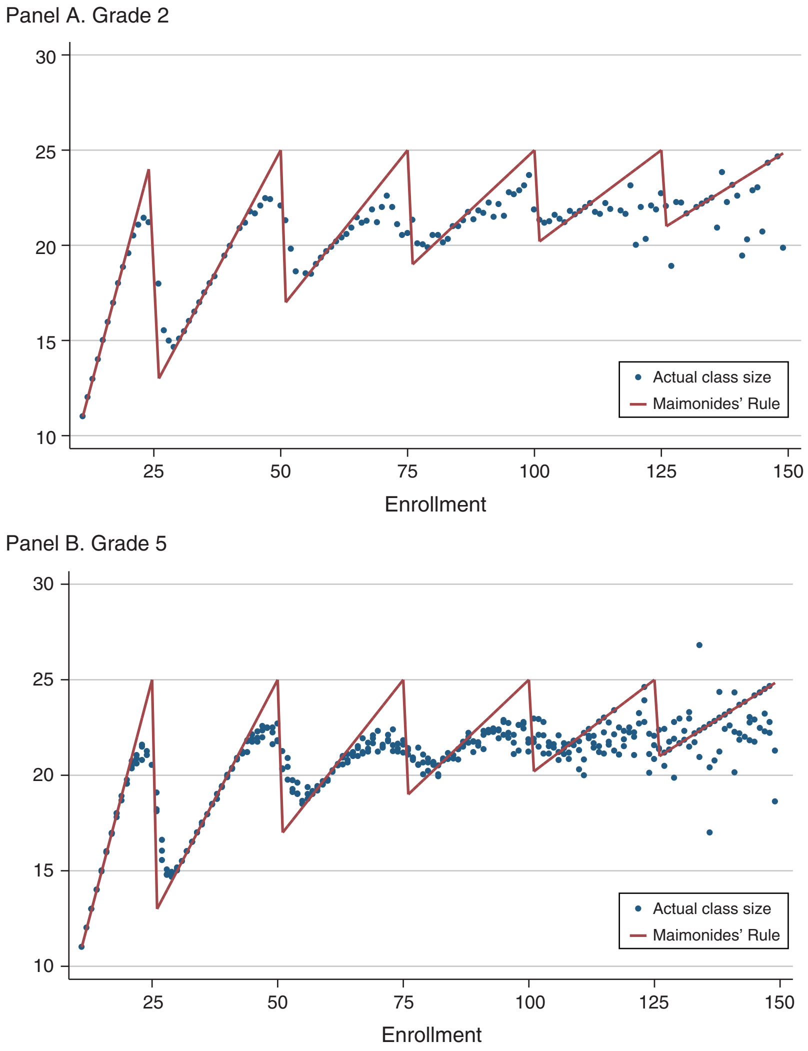

Figure 2. Class Size by Enrollment in Pre-reform Years

Note: The figure shows actual class size and the class size predicted by Maimonides' Rule using data for students enrolled under a class size cap of 25.

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.

C. *Measuring Manipulation*

Our score manipulation flag switches on as a function of implausible score levels, the within-class average and standard deviation of test scores, the number of missing

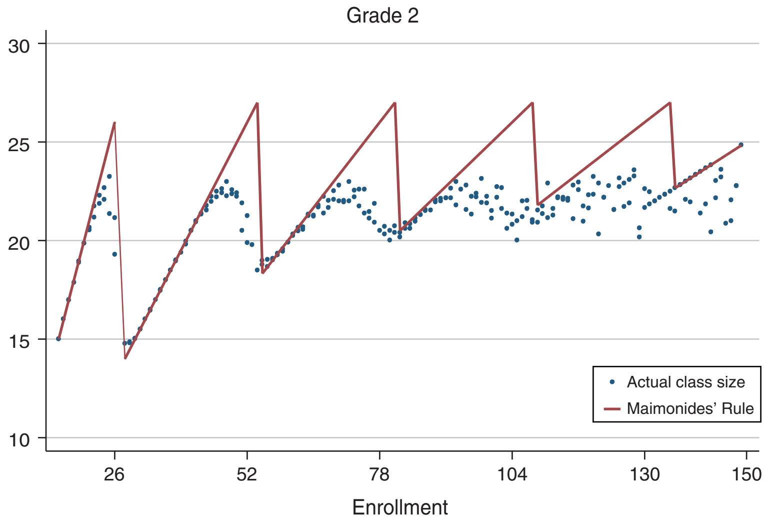

Figure 3. Class Size by Enrollment in Post-reform Years

Note: The figure shows actual class size and the class size predicted by Maimonides' Rule using data for students enrolled under a class size cap of 27.

items, and a Herfindahl index of the share of students with similar response patterns. These indicators are used as inputs to 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, Castellano, and Longobardi (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 online Appendix. A section on threats to validity considers the consequences of possible misclassification of manipulation for our empirical strategy.[5]

III. Class-Size Effects: Achievement and Manipulation

A. *Graphical Analysis*

We begin with nonparametric RD plots that capture class-size effects near enrollment cutoffs. The first in this sequence, Figure 4, documents the relationship

[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 compares 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 their 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.

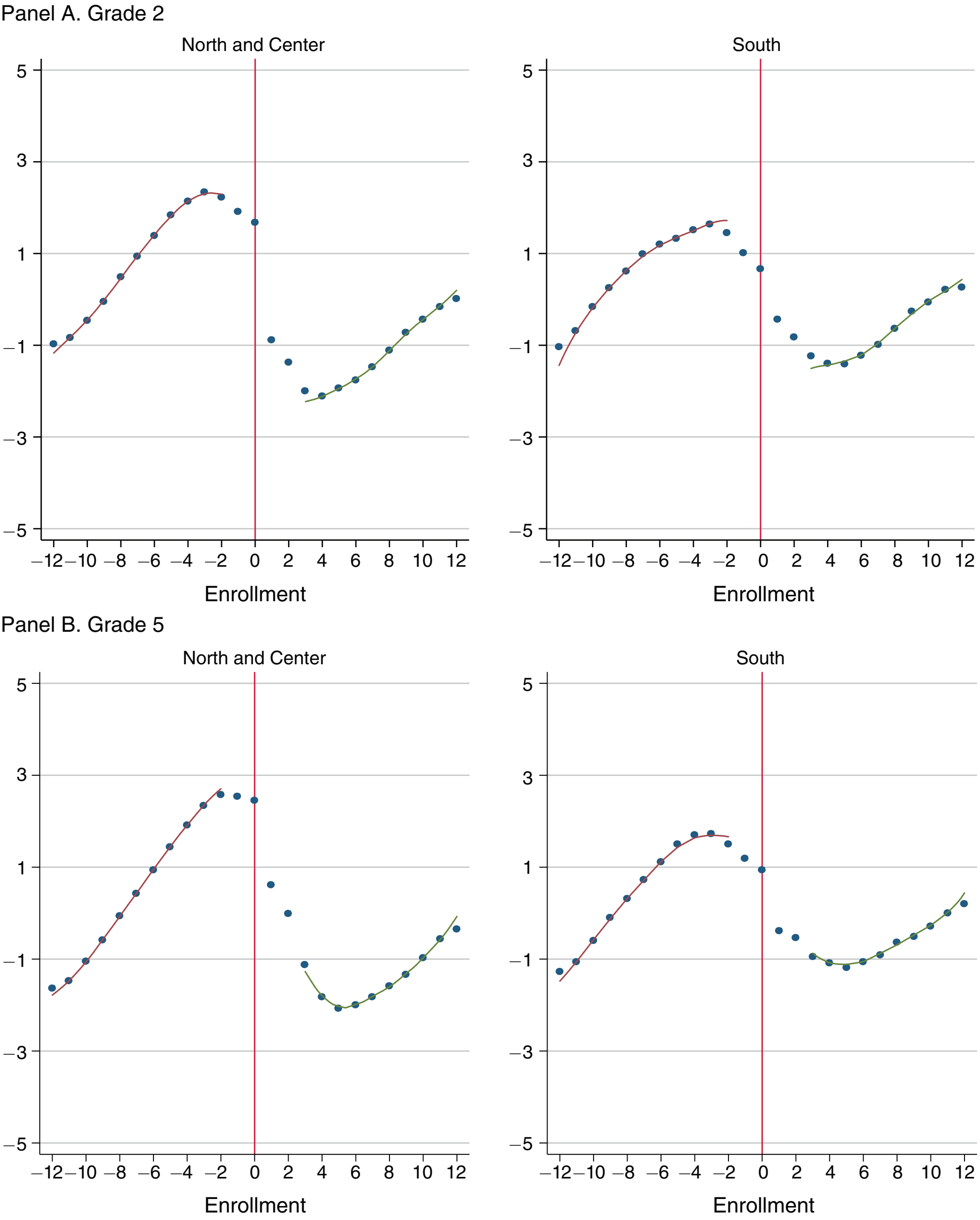

Figure 4. Class Size and Enrollment, Centered at Maimonides' Cutoffs

Notes: Graphs plot residuals from a regression of class size on the controls included in equation (2). The solid line shows a one-sided LLR fit.

between cutoffs (multiples of 25 or 27) and class size. This figure was constructed from a sample of classes at schools with enrollment falling in a [−12,12] window around the first four cutoffs shown in Figures 2 and 3. Enrollment values in each window are centered 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 three-point moving average (smoothing does not cross cutoffs). Figure 4 also plots fitted values generated by local linear regressions (LLR) fit to class-level data. 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 if Maimonides' Rule were tightly enforced. The parametric identification strategy detailed below exploits both the discontinuous variation 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 2009/2010 reform. Near cutoffs, the change in size generated by moving across a cutoff is on the order of two to four students, smaller in the South than elsewhere.

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 scores at schools outside of 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 test 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 mismeasured manipulation might have in generating this pattern. The implications of misclassification for 2SLS estimates of class-size effects are explored in detail in Section VIB below. We note here, however, that classification error is unlikely to change discontinuously at Maimonides' class-size

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

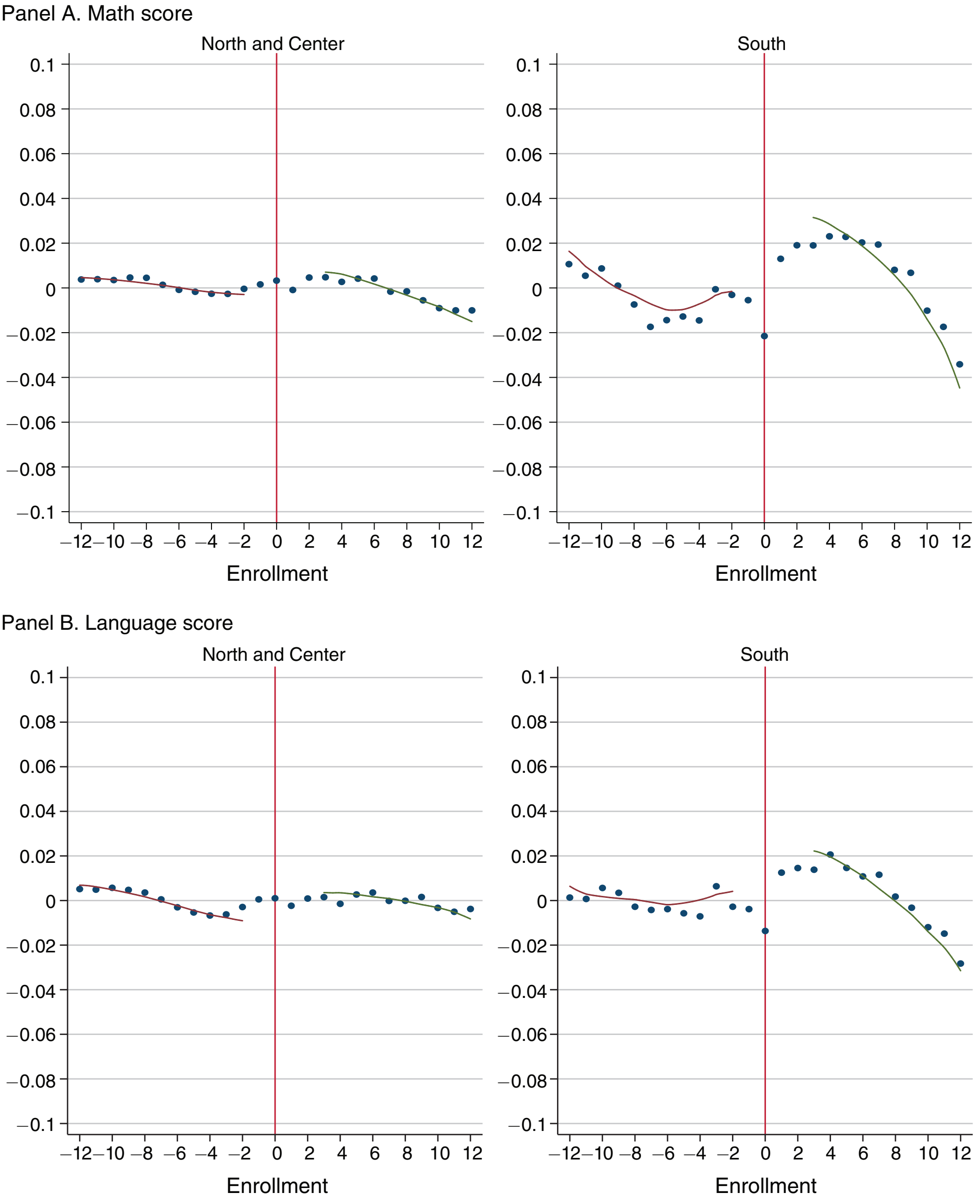

Figure 5. Test Scores and Enrollment, Centered at Maimonides' Cutoffs

Notes: Graphs plot residuals from a regression of test scores on the controls included in equation (2). The solid line shows a one-sided LLR fit.

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).

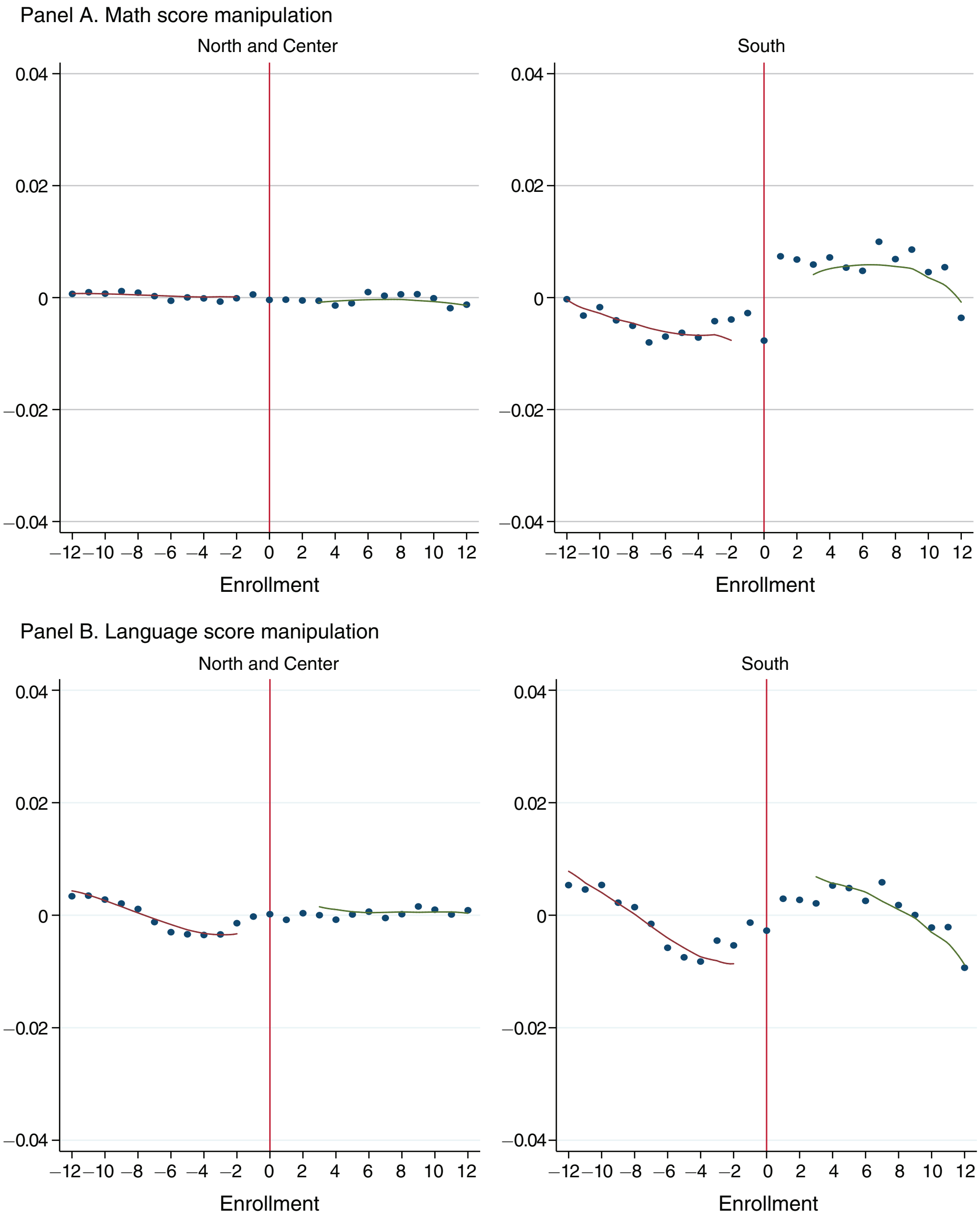

Figure 6. Score Manipulation and Enrollment, Centered at Maimonides' Cutoffs

Notes: Graphs plot residuals from a regression of score manipulation on the controls included in equation (2). The solid line shows a one-sided LLR fit.

B. *Empirical Framework for Class-Size Effects*

Figures 4 and 5 suggest that variation in class size near Maimonides' cutoffs can be used to identify class-size effects in a nonparametric fuzzy regression discontinuity (RD) framework. In what follows, however, we opt for parametric

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

	OLS			IV/2SLS					
	Italy	North/ Center	South	Italy	North/ Center	South	Italy	North/ Center	South
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Panel A. Math									
Class size	−0.008	−0.022	0.009	−0.052	−0.044	−0.096	−0.061	−0.042	−0.129
	(0.007)	(0.007)	(0.015)	(0.013)	(0.012)	(0.036)	(0.020)	(0.017)	(0.051)
Enrollment	X	X	X	X	X	X	X	X	X
Enrollment squared	X	X	X	X	X	X	X	X	X
Interactions							X	X	X
Observations	140,010	87,498	52,512	140,010	87,498	52,512	140,010	87,498	52,512
Panel B. Language									
Class size	0.003	−0.019	0.033	−0.040	−0.031	−0.064	−0.041	−0.022	−0.094
	(0.006)	(0.005)	(0.011)	(0.011)	(0.009)	(0.029)	(0.016)	(0.014)	(0.040)
Enrollment	X	X	X	X	X	X	X	X	X
Enrollment squared	X	X	X	X	X	X	X	X	X
Interactions							X	X	X
Observations	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 ten 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. All regressions include sampling strata controls (grade enrollment at institution, region dummies, and their interactions). Other control variables are listed in footnote 7.

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 test 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, specifications 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]

[7] Control variables include percent of female students in the class; percent of immigrant students; dummies for missing values in these variables; percent of students with high school dropout, high school graduate, or college graduate fathers (missing information is the omitted category); percent of students with employed, unemployed, or inactive mothers (missing information is the omitted category). Stratification controls consist of total enrollment in grade, region dummies, and the interaction between enrollment and region.

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, OLS estimates 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 associated with somewhat higher language scores in the South, while Southern class sizes appear to be unrelated to achievement in math.

The 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 appear in online 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. The 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.

C. *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 effects of class size on manipulation rates, with effects on the order of 5 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 in the non-interacted specification). OLS estimates of effects of class

[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.

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

	OLS			IV/2SLS					
	Italy (1)	North/ Center (2)	South (3)	Italy (4)	North/ Center (5)	South (6)	Italy (7)	North/ Center (8)	South (9)
Panel A. Math									
Class size	−0.016	−0.008	−0.031	−0.019	−0.004	−0.054	−0.018	−0.005	−0.047
	(0.003)	(0.002)	(0.006)	(0.005)	(0.003)	(0.014)	(0.007)	(0.005)	(0.020)
Enrollment	X	X	X	X	X	X	X	X	X
Enrollment squared	X	X	X	X	X	X	X	X	X
Interactions							X	X	X
Observations	139,996	87,491	52,505	139,996	87,491	52,505	139,996	87,491	52,505
Panel B. Language									
Class size	−0.017	−0.012	−0.024	−0.020	−0.012	−0.040	−0.016	−0.006	−0.038
	(0.002)	(0.002)	(0.005)	(0.004)	(0.003)	(0.013)	(0.006)	(0.005)	(0.018)
Enrollment	X	X	X	X	X	X	X	X	X
Enrollment squared	X	X	X	X	X	X	X	X	X
Interactions							X	X	X
Observations	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 ten 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. All regressions include sampling strata controls (grade enrollment at institution, region dummies and their interactions). Other control variables are listed in footnote 7.

size on score manipulation, though smaller in magnitude, reflect the same negative effects as 2SLS.

IV. 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.[9]

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

[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, Brunello, and Rocco (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.

TABLE 4—COVARIATE BALANCE IN THE MONITORING EXPERIMENT

	Italy		North/Center		South	
	Control mean	Treatment difference	Control mean	Treatment difference	Control mean	Treatment difference
	(1)	(2)	(3)	(4)	(5)	(6)
Panel A. Administrative variables						
Class size	19.8	0.035	20.0	0.018	19.4	0.062
	[3.57]	(0.030)	[3.51]	(0.037)	[3.65]	(0.052)
Grade enrollment at school	53.1	−0.401	49.8	−0.548	58.5	−0.141
	[30.7]	(0.329)	[27.6]	(0.391)	[34.44]	(0.591)
Percent 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)
Percent 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)
Percent 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)
Panel B. Data provided by school staff						
Female students	0.482	0.0012	0.483	0.0004	0.479	0.0027
	[0.121]	(0.0009)	[0.118]	(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	0.250	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	0.012
	[0.267]	(0.0024)	[0.258]	(0.0031)	[0.210]	(0.003)
Panel C. Nonresponse indicators						
Missing data on father's education	0.223	−0.022	0.225	−0.019	0.221	−0.027
	[0.341]	(0.003)	[0.340]	(0.0043)	[0.343]	(0.008)
Missing data on mother's occupation	0.195	−0.017	0.196	−0.0083	0.194	−0.032
	[0.328]	(0.003)	[0.325]	(0.0042)	[0.333]	(0.005)
Missing data on country of origin	0.033	−0.012	0.025	−0.0078	0.045	−0.018
	[0.163]	(0.001)	[0.143]	(0.0014)	[0.192]	(0.003)
Observations	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.

reported in panel A of Table 4. 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 statistically significant treatment-control differential in missing data rates documented at the bottom of the table. Among other salutary effects, randomly assigned monitors reduce item nonresponse 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 VI.

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—Monitoring Effects on Score Manipulation and Test Scores

	Score manipulation			Test scores		
	Italy	North/Center	South	Italy	North/Center	South
	(1)	(2)	(3)	(4)	(5)	(6)
Panel 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)
Dependent variable mean	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)
Observations	139,996	87,491	52,505	140,010	87,498	52,512
Panel 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)
Dependent variable mean	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)
Observations	140,003	87,493	52,510	140,010	87,498	52,512

Notes: Columns 1–3 report estimates of the effect of monitoring on score manipulation. Columns 4–6 show the 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. All regressions include sampling strata controls (grade enrollment at institution, region dummies, and their interactions). Other controls are listed in footnote 7.

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 score manipulation indicators 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 6 of Table 5. 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 can be interpreted as 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 misclassification are discussed in Section VI.

Manipulation Is Curbstoning.—The fact that monitoring reduces score manipulation and that manipulation *decreases* with class size suggests that teachers are

[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.

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 show that score manipulation decreases with class size and therefore weigh against student cheating. Finally, because individual test scores are never disclosed even to those tested, it's 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 estimates of the difficulty gradient, that is, average reported scores as a function of item difficulty. We identify this gradient for manipulators and others by assuming reported scores reflect two underlying potential score distributions for each item, j, one revealed in the presence of manipulation, denoted $y_{igkt}^{j}(1)$, and one revealed otherwise, denoted $y_{igkt}^{j}(0)$. Observed scores in class i on item j, denoted by y_{igkt}^{j}, are determined by

$$y_{igkt}^{j} = (1 - m_{igkt}) y_{igkt}^{j}(0) + m_{igkt} y_{igkt}^{j}(1),$$

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

The means of the underlying potential scores determining y_{igkt}^{j} are identified here by adapting methods developed by Abadie (2002) (an application of this approach appears in Angrist, Pathak, and Walters 2013). Specifically, we compute 2SLS estimates of the parameters β_1^j and β_0^j in models of the form

$$\begin{aligned} y_{igkt}^{j} m_{igkt} &= \rho_1(t,g) + \beta_1^j m_{igkt} + \epsilon_{igkt}, \\ y_{igkt}^{j}(1 - m_{igkt}) &= \rho_0(t,g) + \beta_0^j (1 - m_{igkt}) + \epsilon_{igkt}, \end{aligned}$$

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_1^j 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_0^j 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).[11]

[11] 2SLS estimates of potential manipulation rates are not constrained to fall between zero and one. Estimates are for items grouped by difficulty using ten equally spaced bands. Figure 7 below also shows linear fits by subject, grading effort, and potential score weighted by the inverse standard error for each cell. Marker size is proportional

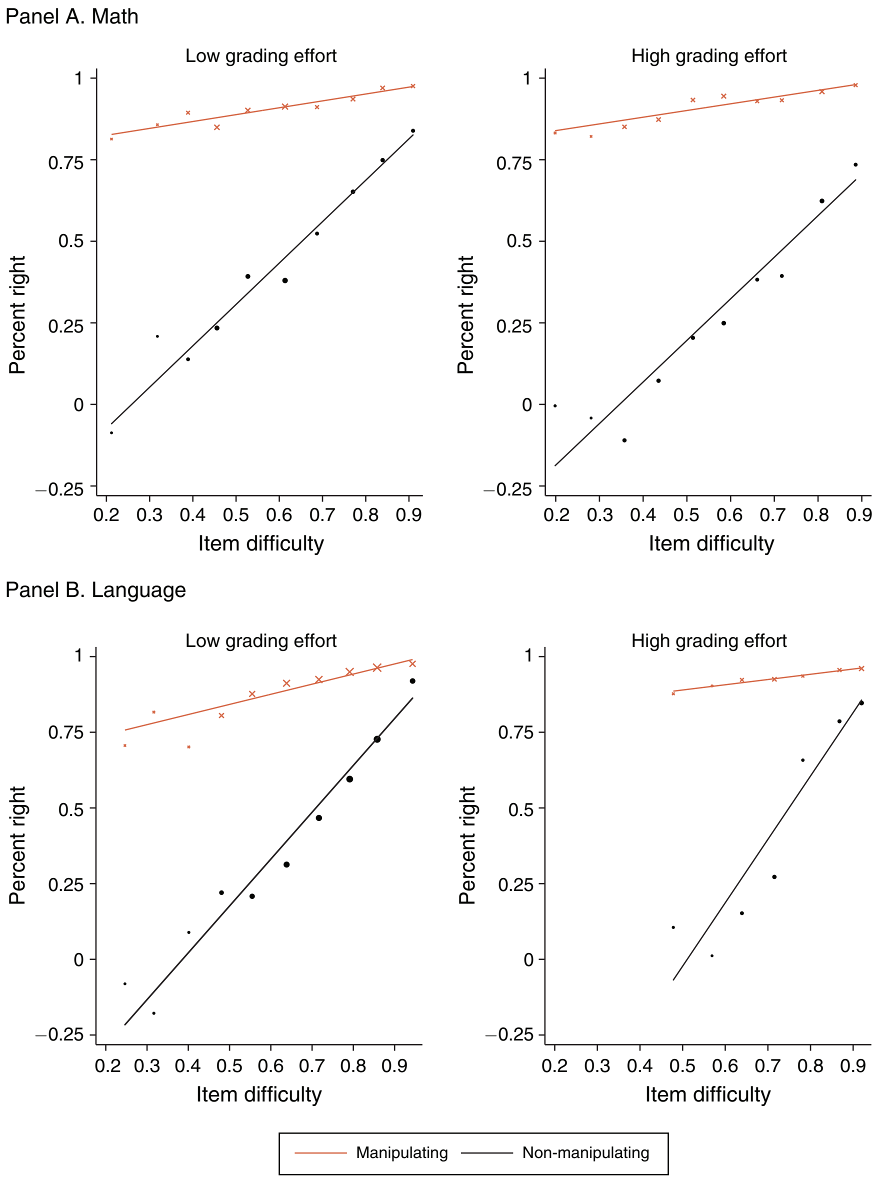

Figure 7. Score Gradient by Item Difficulty

Notes: This figure shows the average potential score by item under manipulation and the average potential score by item without manipulation plotted against the percent correct on the item in monitored institutions in Veneto. The sample is restricted to the South. Additional details appear in footnote 11.

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,

to the number of items used to compute estimates in each band. Three outlier items were dropped from the language/high grading-effort panel.

which also shows a least squares fit to this 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 Figure 7.

Figure 7 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 online 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-grading-effort 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 explanation 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.

V. 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, Table 6 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 likewise reveal a strong negative effect of class size on achievement, but much more so in the absence of monitoring (and, again, in the South). They also suggest 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 test for which scores vary across items and as a result of manipulation, but not otherwise. Without manipulation, the

Table 6—IV/2SLS Estimates of the Effect of Class Size on Scores by Monitoring Status

	Math			Language		
	Italy (1)	North/ Center (2)	South (3)	Italy (4)	North/ Center (5)	South (6)
Class size × M_{igkt}	−0.035	−0.039	−0.035	−0.031	−0.021	−0.048
	(0.024)	(0.021)	(0.061)	(0.019)	(0.017)	(0.048)
Class size × $(1 - M_{igkt})$	−0.066	−0.042	−0.143	−0.042	−0.021	−0.098
	(0.021)	(0.018)	(0.053)	(0.016)	(0.014)	(0.042)
M_{igkt}	−0.174	−0.082	−0.395	−0.103	−0.055	−0.228
	(0.041)	(0.038)	(0.096)	(0.033)	(0.030)	(0.076)
Observations	140,010	87,498	52,512	140,010	87,498	52,512

Notes: This table reports 2SLS estimates using the interaction of Maimonides' Rule with monitor at institution (M_{igkt}) as instruments. Class-size coefficients show the effect of ten students. The unit of observation is the class. Robust standard errors, clustered on school and grade, are shown in parentheses. All regressions include sampling strata controls (grade enrollment at institution, region dummies, and their interactions). Other controls are listed in footnote 7.

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 $p_j = n_j/s$ 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 average score gain from manipulation is $\tilde{p}_j \equiv \tau_j p_j$, implying $\partial \tilde{p}_j / \partial n_j = \tau_j / s$. This reflects the facts that the value of a single manipulated exam 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 and in monitored institutions. Suppose teachers maximize 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 teachers' problem is to choose n_h (and hence p_h) to solve

$$\max \left(1 - \gamma(s) \sum_h n_h\right) \alpha \sum_h \tau_h p_h - \beta \sum_h (s - n_h),$$

where $\alpha \sum_h \tau_h p_h$ is the utility of overall exam performance; $\gamma(s) \sum_h n_h$ is discovery risk; $\beta \sum_h (s - n_h)$ is the disutility of honest grading, and utility falls to zero when manipulation is discovered. Parameters α and β reflect discovery-weighted score gains and the relative weight teachers place on grading effort. 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 risk-attenuated utility gain that decreases with class size, specifically a gain of $\alpha\frac{\tau_j}{s}$, with the additional (constant) utility of reduced grading effort, β. These gains are offset by increased disclosure risk in amount $\gamma(s)$ per item. Dividing the maximand by s, the first-order conditions (FOCs) for optimal manipulation rates can be written

$$\frac{\tau_j}{s} + \frac{\beta}{\alpha} - \gamma(s)\sum_h\left(\tau_j + \tau_h\right)p_h \;=\; 0,$$

for each item indexed by j, where $j \;=\; 1, \ldots, J$.

When $\gamma(s) \;\approx\; 0$, as we might expect in many small classes where teachers transcribe score sheets unassisted by peers, this model predicts manipulation of all items for entire classes because the FOC is constant and positive. This behavior produces the pattern seen in Figure 7, which shows near-perfect exams on all items in classes identified as having manipulated scores. When $\gamma(s)$ is large enough, as we might expect in some large classes, the FOC becomes negative and the model predicts no manipulation whatsoever. For $\gamma(s)$ between these two extremes, a comparative statics result derived in the online Appendix yields

$$\frac{dp_j}{ds} \;=\; -\frac{1 + \gamma'(s)\,s^2\sum_h\left(1 + \frac{\tau_h}{\tau_j}\right)p_h}{2\gamma(s)\,s^2} \;<\; 0, \tag{3}$$

where $\gamma'(s) \;>\; 0$, and hence

$$\frac{dy_j}{ds} \;=\; \tau_j\frac{dp_j}{ds} \;=\; -\frac{\tau_j + \gamma'(s)\,s^2\sum_h\left(\tau_j + \tau_h\right)p_h}{2\gamma(s)\,s^2} \;<\; 0,$$

implying a score gradient decreasing with class size. This can be written as

$$\frac{dy_j}{ds} \;=\; -\frac{\tau_j}{2\gamma(s)\,s^2} - \frac{d\log\gamma(s)}{ds}\frac{\sum_h\left(\tau_j + \tau_h\right)p_h}{2}.$$

The first term here reflects diminishing score gains from manipulation as class size increases $\left(\frac{\alpha^{-1}d\left(\alpha\frac{\tau_j}{s}\right)}{ds}\right)$, while the increased risk of discovery, that is, $\gamma(s)$ in the denominator, is constant. The second term reflects increased risk of discovery by peers (or monitors) in larger classes, that is, $\gamma'(s)$.

The online Appendix gives a more general version of these results, assuming log-linear preferences (as in, for example, Blundell and McCurdy 1999). The Appendix also shows that allowing for convex disutility of effort moderates the negative effect of increasing class size on manipulation. For example, when the cost of honest grading for item j is

$$C\left(s - n_j\right) \;=\; \beta_1\left(s - n_j\right) + \beta_2\left(s - n_j\right)^2,$$

(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

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

The bottom line is still unclear, however; what matters is the contrast with the disclosure risk parameterized by $\gamma(s)$.

VI. Score Manipulation Explains Class-Size Effects

A. *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 as

$$(4) \qquad y_{igkt} = \rho_0(t,g) + \beta_1 s_{igkt} + \beta_2 m_{igkt} + \rho_1 r_{gkt} + \rho_2 r_{gkt}^2 + \eta_{igkt},$$

where $\rho_0(t,g)$ is again a shorthand for year and grade effects. We interpret equation (4) 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 (4) represents a model for potential outcomes indexed against two jointly manipulable treatments.

We estimate equation (4) 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 the dummy indicating classes at institutions with randomly assigned monitors, M_{igkt}. The first-stage equations associated with these two instruments can be written as

$$(5) \qquad s_{igkt} = \lambda_{10}(t,g) + \mu_{11} f_{igkt} + \mu_{12} M_{igkt} + \lambda_{11} r_{gkt} + \lambda_{12} r_{gkt}^2 + \xi_{ik},$$

$$(6) \qquad m_{igkt} = \lambda_{20}(t,g) + \mu_{21} f_{igkt} + \mu_{22} M_{igkt} + \lambda_{21} r_{gkt} + \lambda_{22} r_{gkt}^2 + \upsilon_{ik},$$

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 is unrelated to class size. This is consistent with random assignment of monitors to institutions.

The 2SLS estimates of β_2 in equation (4), reported in Table 8, show large effects of manipulation on test scores. At the same time, this table also shows small and mostly insignificant estimates of β_1, the coefficient on class size in the multivariate

Table 7—Twin First Stages

	Math			Language		
	Italy (1)	North/ Center (2)	South (3)	Italy (4)	North/ Center (5)	South (6)
Panel A. Score manipulation						
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)
Observations	139,996	87,491	52,505	140,003	87,493	52,510
Panel B. Class size						
Maimonides' Rule (f_{igkt})	0.513	0.555	0.433			
	(0.001)	(0.001)	(0.001)			
Monitor at institution (M_{igkt})	0.013	0.032	−0.009			
	(0.024)	(0.027)	(0.045)			
Observations	140,010	87,498	52,512			

Notes: Panel A reports first-stage estimates of the effect of the Maimonides' Rule and a monitor at institution on score manipulation. Panel B reports 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. All regressions include sampling strata controls (grade enrollment at institution, region dummies, and their interactions). Other control variables are listed in footnote 7.

model. In an effort to boost the precision of these estimates, we estimate overidentified models that add four dummies for values of the running variable that fall within 10 percent of each cutoff, a specification motivated by the nonparametric first stage captured in Figure 4.[12] The most precise of the estimated zeros reported in Table 8, generated by the overidentified 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 are the estimates for language in column 4. It's also worth noting that the overidentification p-values reported at the bottom of Table 4 are mostly far from conventional significance levels.

Table 8 also reports 2SLS estimates computed by adding an interaction term, $s_{igkt}m_{igkt}$, to equation (4), 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 subsample, while an additive model like equation (4) 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 positive class-size effects for the Italian Mezzogiorno, results that contrast with the much larger and statistically significant negative class-size effects for the same area reported in Table 2. In column 9 of the latter table, for example, a 10-student

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

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/ Center	South	Italy	North/ Center	South	Italy	North/ Center	South
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Panel A. Math									
Class size	0.0075	−0.0029	0.0062	0.0024	−0.011	0.013	0.012	0.014	0.047
	(0.0213)	(0.0298)	(0.0441)	(0.0190)	(0.025)	(0.038)	(0.032)	(0.048)	(0.068)
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							−0.146	−1.270	−0.227
× Score manipulation							(0.481)	(2.160)	(0.430)
Overid test p-value				0.914	0.600	0.541	0.914	0.475	0.476
Observations	139,996	87,491	52,505	139,996	87,491	52,505	139,996	87,491	52,505
Panel B. Language									
Class size	0.012	0.0049	0.013	0.022	0.011	0.049	0.033	0.0098	0.134
	(0.017)	(0.0196)	(0.039)	(0.015)	(0.017)	(0.033)	(0.031)	(0.0320)	(0.080)
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							−0.213	−0.0029	−0.706
× Score manipulation							(0.498)	(1.0898)	(0.621)
Overid test p-value				0.129	0.796	0.036	0.216	0.844	0.109
Observations	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 Maimonides' 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 ten 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. Models include sampling strata controls (grade enrollment at institution, region dummies, and their interactions). Other control variables are listed in footnote 7.

reduction in class size is estimated to boost achievement by 0.10σ or more. The corresponding multivariate estimates in column 6 of Table 8 are of the opposite sign, showing that larger classes increase achievement, though not by very much. The overidentified 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 reasonably 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.

B. *Threats to Validity*

We briefly consider three possible threats to validity relevant for a 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

$$y_i = \rho_0 + \beta_1 s_i + \beta_2 m_i^* + \zeta_i, \tag{7}$$

where instruments are assumed to be uncorrelated with the error, ζ_i, as in equation (4). 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

$$m_i = (1 - \pi_0) + (\pi_0 + \pi_1 - 1)m_i^* + \omega_i, \tag{8}$$

where the residual, ω_i, is defined by

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

and the probability of manipulation given z_i and m_i^* satisfies

$$\Pr[m_i = d | z_i, m_i^* = d] = \Pr[m_i = d | m_i^* = d] \equiv \pi_d, \tag{9}$$

for $d = 0, 1$. Note that $E[z_i \omega_i] = 0$ by definition of ω_i. Using (8) to substitute for m_i^*, equation (7) can be rewritten as

$$y_i = \left[\rho_0 - \frac{\beta_2(1-\pi_0)}{\pi_0 + \pi_1 - 1}\right] + \beta_1 s_i + \left[\frac{\beta_2}{\pi_0 + \pi_1 - 1}\right] m_i + \left[\zeta_i - \beta_2 \frac{\omega_i}{\pi_0 + \pi_1 - 1}\right]. \tag{10}$$

We assume that the π_d 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 (10) is finite and has the same sign as β_2.

The 2SLS estimate of the coefficient on reported score manipulation is biased upward, since $\pi_0 + \pi_1 - 1$ lies between 0 and 1 given our assumptions. Because equation (10) has a residual uncorrelated with the instruments and the coefficient on class size is unchanged in this model, misclassification of the sort described by (9) leaves estimates of the class-size coefficient, β_1, unchanged. Similar results for the consequences of classification error appear in Kane, Rouse, and Staiger (1999); Mahajan (2006); and Lewbel (2007), among others, though other work focuses on the consequences of misclassification for the coefficient on a variable subject to error rather than implications for other regressors in the model.[13]

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. A likely channel for the link between Maimonides' Rule and manipulation is increased teacher shirking (perhaps due to reduced peer monitoring) 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: some of the variables reported by school staff and nonresponse indicators are 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 more imbalance in the South.

Additional evidence suggesting that the link between covariates is a data quality effect unrelated to sorting appears in online Appendix Table A3. This table shows that f_{igkt} is largely unrelated to covariates in schools with monitors, where

[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 percent 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_d around 0.8 for Italy $\left(\text{since } \frac{2.25}{2 \times 0.8 - 1} = 3.75\right)$, though the implied conditional classification rates are closer to 0.65 for math scores outside the South. These rates seem like reasonable descriptions of the classification process. The possible misclassification of manipulators is further investigated by Battistin, De Nadai, and Vuri (forthcoming) with reference to the problem of regional rankings of performance on INVALSI tests.

Table 9—Maimonides' Rule and Covariate Balance

	Italy		North/Center		South	
	Control mean	Treatment difference	Control mean	Treatment difference	Control mean	Treatment difference
	(1)	(2)	(3)	(4)	(5)	(6)
Panel A. Administrative data on schools						
Percent in class sitting the test	0.939	0.0000	0.935	0.0001	0.947	0.0000
	[0.064]	(0.0001)	[0.066]	(0.0001)	[0.061]	(0.0001)
Percent in school sitting the test	0.939	0.0001	0.934	0.0001	0.946	0.0001
	[0.053]	(0.0001)	[0.055]	(0.0001)	[0.050]	(0.0001)
Percent in institution sitting the test	0.937	−0.0001	0.933	−0.0001	0.945	−0.0000
	[0.044]	(0.0001)	[0.043]	(0.0001)	[0.044]	(0.0001)
Panel B. Data provided by school staff						
Female	0.482	0.0000	0.484	0.0002	0.479	−0.0002
	[0.121]	(0.0002)	[0.118]	(0.0002)	[0.125]	(0.0003)
Immigrant	0.098	−0.0007	0.138	−0.0007	0.032	−0.0004
	[0.120]	(0.0002)	[0.130]	(0.0003)	[0.057]	(0.0001)
Father HS	0.255	0.0006	0.261	0.0002	0.243	0.0013
	[0.168]	(0.0003)	[0.163]	(0.0003)	[0.176]	(0.0005)
Mother employed	0.450	0.0012	0.536	0.0010	0.308	0.0016
	[0.266]	(0.0004)	[0.257]	(0.0005)	[0.214]	(0.0006)
Panel C. Nonresponse indicators						
Missing data on father's education	0.219	0.0003	0.222	0.0015	0.214	−0.0018
	[0.336]	(0.0006)	[0.336]	(0.0007)	[0.337]	(0.0010)
Missing data on mother's occupation	0.193	0.0002	0.196	0.0014	0.186	−0.0019
	[0.324]	(0.0006)	[0.323]	(0.0007)	[0.325]	(0.0010)
Missing data on country of origin	0.030	−0.0001	0.023	−0.0001	0.040	−0.0000
	[0.1544]	(0.0002)	[0.136]	(0.0003)	[0.180]	(0.0005)
Observations	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 control means are in square brackets; robust standard errors are in parentheses.

manipulation is considerably diminished (though not necessarily eliminated, since some classes in monitored institutions remain unmonitored).

VII. 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 suggest that 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 monitoring experiment reveal substantial test score 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 not an issue. 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 United Kingdom's Key Stage 2 primary-level assessments are marked by external examiners, a costly effort that our findings suggest may be worthwhile. 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 2016, and Burgess and Greaves 2013). It's also worth asking why class-size reductions fail to enhance learning in Italy, while evidence from the United States, Israel, and a number of other countries suggest class-size reductions often increase learning. We hope to address these questions in future work.

REFERENCES

Abadie, Alberto. 2002. "Bootstrap Tests for Distributional Treatment Effects in Instrumental Variables Models." *Journal of the American Statistical Association* 97 (457): 284–92.

Angrist, Joshua D., Erich Battistin, and Daniela Vuri. 2017. "In a Small Moment: Class Size and Moral Hazard in the Italian Mezzogiorno: Dataset." *American Economic Journal: Applied Economics*. https://doi.org/10.1257/app.20160267.

Angrist, Joshua D., and Victor Lavy. 1999. "Using Maimonides' Rule to Estimate the Effect of Class Size on Scholastic Achievement." *Quarterly Journal of Economics* 114 (2): 533–75.

Angrist, Joshua D., Victor Lavy, Jetson Leder-Luis, and Adi Shany. 2017. "Maimonides Rule Redux." National Bureau of Economic Research (NBER) Working Paper 23486.

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

Baker, Olesya, and M. Daniele Paserman. 2013. "Grade Enrollment Sorting under an Incentives-Based Class Size Reduction Program." Unpublished.

Ballatore, Rosario Maria, Margherita Fort, and Andrea Ichino. Forthcoming. "The Tower of Babel in the Classroom: Immigrants and Natives in Italian Schools." *Journal of Labor Economics*.

Banerjee, Abhijit, and Esther Duflo. 2006. "Addressing Absence." *Journal of Economic Perspectives* 20 (1): 117–32.

Battistin, Erich, Michele De Nadai, and Daniela Vuri. Forthcoming. "Counting Rotten Apples: Student Achievement and Score Manipulation in Italian Elementary Schools." *Journal of Econometrics*.

Battistin, Erich, and Lorenzo Neri. 2017. "School Accountability, Score Manipulation, and Economic Geography." Unpublished.

Bertoni, Marco, Giorgio Brunello, and Lorenzo Rocco. 2013. "When the cat is near, the mice won't play: The effect of external examiners in Italian schools." *Journal of Public Economics* 104: 65–77.

Blundell, Richard, and Thomas McCurdy. 1999. "Labor Supply: A Review of Alternative Approaches." In *Handbook of Labor Economics*, Vol. 3A, edited by Orley C. Ashenfelter and David Card, 1559–1695. Amsterdam: North-Holland.

Böhlmark, Anders, and Mikael Lindahl. 2015. "Independent Schools and Long-Run Educational Outcomes: Evidence from Sweden's Large-scale Voucher Reform." *Economica* 82 (327): 508–51.

Bonesrønning, Hans. 2003. "Class Size Effects on Student Achievement in Norway: Patterns and Explanations." *Southern Economic Journal* 69 (4): 952–65.

Bratti, Massimiliano, Daniele Checchi, and Antonio Filippin. 2007. "Geographical Differences in Italian Students' Mathematical Competencies: Evidence from PISA 2003." *Giornale degli Economisti e Annali di Economia* 66 (3): 299–335.

Brunello, Giorgio, and Daniele Checchi. 2005. "School quality and family background in Italy." *Economics of Education Review* 24 (5): 563–77.

Burgess, Simon, and Ellen Greaves. 2013. "Test Scores, Subjective Assessment, and Stereotyping of Ethnic Minorities." *Journal of Labor Economics* 31 (3): 535–76.

Chaudhury, Nazmul, Jeffrey Hammer, Michael Kremer, Karthik Muralidharan, and F. Halsey Rogers. 2006. "Missing in Action: Teacher and Health Worker Absence in Developing Countries." *Journal of Economic Perspectives* 20 (1): 91–116.

Chetty, Raj, John N. Friedman, Nathaniel Hilger, Emmanuel Saez, Diane Whitmore Schanzenbach, and Danny Yagan. 2011. "How Does Your Kindergarten Classroom Affect Your Earnings? Evidence from Project STAR." *Quarterly Journal of Economics* 126 (4): 1593–1660.

Clotfelter, Charles T., Helen F. Ladd, and Jacob L. Vigdor. 2009. "Are Teacher Absences Worth Worrying About in the United States?" *Education Finance and Policy* 4 (2): 115–49.

Costantini, Mauro, and Claudio Lupi. 2006. "Divergence and long-run equilibria in Italian regional unemployment." *Applied Economics Letters* 13 (14): 899–904.

De Paola, Maria, Vincenzo Scoppa, and Valeria Pupo. 2014. "Absenteeism in the Italian Public Sector: The Effects of Changes in Sick Leave Policy." *Journal of Labor Economics* 32 (2): 337–60.

Dee, Thomas S., Will Dobbie, Brian A. Jacob, and Jonah Rockoff. 2016. "The Causes and Consequences of Test Score Manipulation: Evidence from the New York Regents Examinations." National Bureau of Economic Research (NBER) Working Paper 22165.

Diamond, Rebecca, and Petra Persson. 2016. "The Long-term Consequences of Teacher Discretion in Grading of High-Stakes Tests." National Bureau of Economic Research (NBER) Working Paper 22207.

Dobbelsteen, Simone, Jesse Levin, and Hessel 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, Patrizia. 2013. "L'esperienza di Restituzione dei Dati al Netto del Cheating." Presentation at the Metodi di Identificazione, Analisi e Trattamento del Cheating Workshop, Rome, February 8.

Gary-Bobo, Robert J., and Mohamed-Badrane 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.

Guiso, Luigi, Paola Sapienza, and Luigi Zingales. 2004. "The Role of Social Capital in Financial Development." *American Economic Review* 94 (3): 526–56.

Guiso, Luigi, Paola Sapienza, and Luigi Zingales. 2011. "Civic Capital as the Missing Link." In *Handbook of Social Economics*, edited by Jess Benhabib, Alberto Bisin, and Matthew O. Jackson, 417–80. Amsterdam: North-Holland.

Hanushek, Eric A. 1995. "Interpreting Recent Research on Schooling in Developing Countries." *World Bank Research Observer* 10 (2): 227–46.

Hoxby, Caroline M. 2000. "The Effects of Class Size on Student Achievement: New Evidence from Population Variation." *Quarterly Journal of Economics* 115 (4): 1239–85.

Ichino, Andrea, and Pietro Ichino. 1997. "Culture, Discrimination and Individual Productivity: Regional Evidence from Personnel Data in a Large Italian Firm." Centre for Economic Policy Research (CEPR) Discussion Paper 1709.

Ichino, Andrea, and Giovanni Maggi. 2000. "Work Environment and Individual Background: Explaining Regional Shirking Differentials in a Large Italian Firm." *Quarterly Journal of Economics* 115 (3): 1057–90.

Ichino, Andrea, and Guido Tabellini. 2014. "Freeing the Italian school system." *Labour Economics* 30: 113–28.

Imbens, Guido, and Karthik Kalyanaraman. 2012. "Optimal Bandwidth Choice for the Regression Discontinuity Estimator." *Review of Economic Studies* 79 (3): 933–59.

Instituto Nazionale per la Valutazione del Sistema Educativo di Istruzione e di Formazione (INVALSI). 2010. *Servizio Nazionale di Valutazione: A.S. 2009/2010: Rilevazione degli apprendimenti—SNV: Prime Analisi*. Rome: INVALSI.

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

Kane, Thomas J., Cecilia Elena Rouse, and Douglas Staiger. 1999. "Estimating Returns to Schooling When Schooling Is Misreported." National Bureau of Economic Research (NBER) Working Paper 7235.

Krueger, Alan B. 1999. "Experimental Estimates of Education Production Functions." *Quarterly Journal of Economics* 114 (2): 497–532.

Lavy, Victor. 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, Victor, and Edith Sand. 2015. "On The Origins of Gender Human Capital Gaps: Short and Long Term Consequences of Teachers' Stereotypical Biases." National Bureau of Economic Research (NBER) Working Paper 20909.

Leuven, Edwin, Hessel Oosterbeek, and Marte Rønning. 2008. "Quasi-experimental Estimates of the Effect of Class Size on Achievement in Norway." *Scandinavian Journal of Economics* 110 (4): 663–93.

Lewbel, Arthur. 2007. "Estimation of Average Treatment Effects with Misclassification." *Econometrica* 75 (2): 537–51.

Mahajan, Aprajit. 2006. "Identification and Estimation of Regression Models with Misclassification." *Econometrica* 74 (3): 631–65.

Neal, Derek. 2013. "The Consequences of Using One Assessment System to Pursue Two Objectives." *Journal of Economic Education* 44 (4): 339–52.

Piketty, Thomas. 2004. "Should We Reduce Class Size or School Segregation? Theory and Evidence from France." http://piketty.pse.ens.fr/fichiers/public/PikettySlides.pdf.

Putnam, Robert D., Robert Leonardi, and Raffaelle Y. Nanetti. 1993. *Making Democracy Work.* Princeton: Princeton University Press.

Quintano, Claudio, Rosalia Castellano, and Sergio 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 and Applicazioni* 2: 149–71.

Severson, Kim. 2011. "Systematic Cheating Is Found in Atlanta's School System." *New York Times*, July 5, A13.

Sims, David. 2008. "A strategic response to class size reduction: Combination classes and student achievement in California." *Journal of Policy Analysis and Management* 27 (3): 457–78.

Terrier, Camille. 2016. "Boys Lag Behind: How Teachers' Gender Biases Affect Student Achievement." Massachusetts Institute of Technology (MIT) School Effectiveness and Inequality Initiative (SEII) Working Paper 2016.07.

Urquiola, Miguel, and Eric Verhoogen. 2009. "Class Size Caps, Sorting, and the Regression-Discontinuity Design." *American Economic Review* 99 (1): 179–215.

Wößmann, Ludger. 2005. "Educational Production in Europe." *Economic Policy* 20 (43): 446–504.