DISCUSSION PAPER SERIES

IZA DP No. 8405

Counting Rotten Apples: Student Achievement and Score Manipulation in Italian Elementary Schools

Erich Battistin
Michele De Nadai
Daniela Vuri

August 2014

Forschungsinstitut
zur Zukunft der Arbeit
Institute for the Study
of Labor

Counting Rotten Apples: Student Achievement and Score Manipulation in Italian Elementary Schools

Erich Battistin
Queen Mary University of London, IRVAPP and IZA

Michele De Nadai
University of Padua

Daniela Vuri
University of Rome Tor Vergata, IZA, CESifo and CEIS

Discussion Paper No. 8405
August 2014

IZA

P.O. Box 7240
53072 Bonn
Germany

Phone: +49-228-3894-0
Fax: +49-228-3894-180
E-mail: iza@iza.org

IZA Discussion Paper No. 8405
August 2014

ABSTRACT

Counting Rotten Apples: Student Achievement and Score Manipulation in Italian Elementary Schools*

We derive bounds for the average of math and language scores of elementary school students in Italy correcting for pervasive score manipulation. Information on the fraction of manipulated data is retrieved from a natural experiment that randomly assigns external monitors to schools. We show how bounds can be tightened imposing restrictions on the measurement properties of the manipulation indicator developed by the government agency charged with test administration and data collection. We additionally assume that manipulation is more likely in those classes at the lower end of the distribution of true scores. Our results show that regional rankings by academic performance are reversed once manipulation is properly taken into account.

JEL Classification: C14, C31, C81, I21, J24

Keywords: corrupt sampling, measurement error, nonparametric bounds, partial identification

Corresponding author:

Daniela Vuri
University of Rome "Tor Vergata"
Via Columbia 2
00133 Rome
Italy
E-mail: 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. Our thanks to Joshua Angrist for helpful discussions and comments. This research is supported by the Fondazione Bruno Kessler. The views expressed here are those of the authors alone.

1 Introduction

International comparisons of standardized tests on cognitive skills from the *Programme for International Student Assessment* (PISA) persistently show Italy at the bottom of achievement league tables (Bratti et al. 2007, Hanushek and Woessmann, 2011). PISA scores also show that, within Italy, Northerners do better than Southerners (see PISA 2012).[1] The same sharp North-South divide is found along many other dimensions. Southern regions are distinguished by persistently higher unemployment, lower per-capita income, higher crime rates, and lower educational attainment. The South also lags in financial development (Guiso et al., 2004), political accountability (Nannicini et al., 2013), and workplace productivity (Ichino and Maggi, 2000). Differential performance across regions is often attributed to cultural differences and differences in residents' view of the role of government (Putnam et al., 1993).

The question of whether regional gaps in Italian scores reflect differences in school quality remains open. In particular, differences in achievement at age 15 documented by PISA may result from deterioration of school quality as students move across grades of compulsory education (which in Italy normally ends at age 16).[2] Italy has only recently embarked on a program of national standardized tests to look into these issues. Surprisingly, in view of the pattern in PISA scores, Italy's own tests (administered by the *Istituto Nazionale per la Valutazione del Sistema dell'Istruzione*, INVALSI in what follows) show Southern students with higher achievement at elementary school. This can be seen from the left hand side panels of Figure 1 and Figure 2 for math and language, respectively, which are obtained pooling data for second and fifth grade students for the years 2009-2011.

The good performance of Southern students at INVALSI tests is however at odds with the evidence from the *Trends in International Mathematics and Science Study* (TIMSS) and the *Progress in International Reading Literacy Study* (PIRLS). The latter surveys, which are representative of the population of Italian fourth grade students, picture the same North-South divide documented by PISA.[3] Moreover, the correlation between student achievement

[1] Italy is divided into 20 administrative regions. The South consists of Basilicata, Campania, Calabria, Puglia, Abruzzo, Molise, and the islands of Sicily and Sardinia.

[2] The policy relevance of this question is strengthened by low mobility of Italian households (Cannari et al. 2000, DiAddario and Patacchini, 2008).

[3] For example, in 2011 fourth grade students in the North score on average 516 in math, 551 in reading and 535 in science with respect to an average of 496 in math, 528 in reading and 510 in science for students living in the South (IEA 2011, CNEL, 2013).

as measured by INVALSI and proxies of school and family inputs unveils patterns which are hard to explain in light of the evidence from the international literature. For example, we show below that social deprivation and lower per-capita income are associated with higher scores, and that public spending - which we proxy with the pupil to teacher ratio at elementary school by region - is inversely related to achievement. Leaving causality aside, these relationships are different from those found at older ages using PISA data.

We aim to reconcile these results, and argue that they can be explained by widespread manipulation on INVALSI tests in the South. The right hand side panels of Figure 1 and Figure 2 reproduce regional estimates of score manipulation for math and language, respectively, as obtained by INVALSI in official publications. The manipulation indicator identifies classes in which scores are likely to have been manipulated through a statistical model that looks for surprisingly high average scores, low within-class variability, and implausible missing data patterns.[4] This measure produces an average of about 6% of compromised scores. In the South the proportion of compromised exams averages about 13% uncovering evidence of a substantial regional gradient. For example, about 16% of classes in Sicily are suspected to have manipulated scores.

Angrist et al. (2014) discuss at large the origin of this phenomenon, focusing on how tests are graded. Differently from international surveys, INVALSI tests are proctored by local administrators and teachers, and institutional features that regulate score transcription seem to facilitate manipulation.[5] This idea is confirmed by reduced manipulation in the presence of external monitors at school (see Bertoni et al. 2013). Score manipulation on the part of teachers is far from unique to Italy. In an early empirical contribution, Jacob and Levitt (2003) documented substantial cheating from teachers in Chicago public schools. More recently, Dee et al. (2011) have shown that scores on New York's Regents exams are manipulated by school staff who grade them in an effort to move marginal students over the performance thresholds. A recent system-wide cheating scandal in Atlanta has raised much

[4]The INVALSI score manipulation variable identifies classes with substantially anomalous score distributions, imputing a probability of manipulation for each (see Quintano et al., 2009 and INVALSI, 2010). We adopt a variant to this definition which is closer in spirit to Jacob and Levitt (2003), as we explain below. Manipulation rates in the figures are computed for 2009-11 scores of second and fifth graders.

[5]PISA tests are graded by a group of test correctors, overseen by a project manager, and the corrections are then cross-checked by other experts. In a similar way, in TIMSS and PIRLS tests scorers are organized into teams, headed by a team leader. The leader's primary responsibility is to monitor scoring reliability by continually checking and rechecking the scores that scorers had assigned.

interest from the media and now threatens to send large numbers of administrators to jail (Severson 2011, Aviv 2014).

The contamination of INVALSI data raises the problem of uncovering true score patterns across Italian regions, which is the objective of this paper. The problem of purging official figures from potential manipulation is now widely recognized by researchers, and has received considerable media attention. This prompted the reaction of test administrators, and from 2012 INVALSI down-weights schools with suspiciously large scores in the derivation of aggregate figures by area (see INVALSI 2012, INVALSI 2013). The validity of this correction is not uncontroversial, as it implicitly assumes that manipulators are a random sample from the population (as well as that they can be detected with no error). The distribution of scores retrieved from raw data is a mixture of true and contaminated scores, with mixing weights representing the percentage of manipulators. To retrieve the distribution of true scores one needs to know the true counterfactual score for manipulators, who are arguably a self-selected group. This, together with the fact that manipulators are possibly misclassified by INVALSI, challenge identification.

We derive bounds for the average of math and language true scores of elementary school students, which represents the parameter of interest. Central to the development of our strategy is to employ restrictions that tighten the bounds and strengthen the conclusions about regional rankings by academic performance. We frame the identification problem within the context of the corrupt sampling model of Horowitz and Manski (1995; HM in what follows), and maintain the assumption that manipulation is aimed at boosting scores. The latter represents a monotonicity restriction motivated by both accountability concerns and the conclusions on the anatomy of the manipulation problem presented in Angrist et al. (2014). We exploit a policy followed by INVALSI that randomly assigns external monitors to about 20% of institutions in the country.

This natural experiment ensures that monitored and unmonitored institutions share the same average of true scores, but present a rather different fraction of manipulated scores. We show that restrictions on the parameter of interest naturally arise from the natural experiment. Manipulation is measured at the class level replicating the same indicator used by INVALSI in official publications, as we discuss further below. However, we allow this indicator to be misclassified. Under the assumption that misclassification is independent of

the sampling process that assigns monitors to classes, data on monitored classes classified as manipulators must be informative on the relationship between measured and latent manipulation. It follows that the extent of classification error in the INVALSI indicator is partly revealed by the monitoring experiment.

We further establish a connection with the partially verified model in Dominitz and Sherman (2006) and the literature on misclassification (see for example Mahajan 2006, Lewbel 2007, and Hu 2008). This leads us to investigate the source of identifying information that arises from imposing that the manipulation indicator is a surrogate of latent manipulation in the relationship with scores (see Carroll et al., 2006 , and Chen et al., 2011). We show that, in our application, bounds obtained under this assumption improve on bounds obtained under the corrupt sampling model. We refine bounds by assuming that manipulation is more likely for classes at the lower end of the distribution of true scores. Our strategy allows for monitoring effects on the propensity to manipulate, as well as on the extent to which true scores are boosted. We also show that if manipulation were exogenous, which corresponds to the contaminated sampling model of HM, our assumptions would yield point identification of the parameter of interest. In particular, our setting implies that point identification follows without exclusion restrictions typically invoked in the literature on misclassification.

The resulting bounds are sufficiently tight to revert regional differences in raw scores, under assumptions detailed below. This can be seen by comparing observed scores in the left hand side panels with adjusted scores in the central panels of Figure 1 and Figure 2 for math and language, respectively. For example Sicily - the region with the highest presumed incidence of manipulation - is ranked $3rd$ among the 20 Italian regions using observed math scores, and $20th$ (or $19th$, depending on variants to the procedure implemented) after our correction. The correlation of ranks before and after the correction is -14% and -9% for math and language, respectively. Consistently with PISA, TIMSS and PIRLS our results depict a rather different pattern than does the correction employed by INVALSI. Moreover we show the relationship with family and school inputs is reverted once manipulation is taken into account, concluding that surveys with pervasive score manipulation may lead to bias conclusions on the role of inputs to the education production function.

The rest of the paper is organized as follows. The next section presents the institutional background and data, and describes the monitoring experiment. Section 3 presents bounds

on the percentage of manipulators. Section 4 presents bounds on scores allowing for corrupt sampling as in HM. We then refine these bounds in Section 5 by imposing restrictions on the manipulation indicator and on the behavior of manipulators. Section 6 concludes.

2 Background and Data

Institutional background and sample selection criteria

We use administrative data collected by the INVALSI on testing program in Italian elementary schools in the 2009/10, 2010/11, and 2011/12 school years. Elementary school lasts 5 years starting from 6 years of age and covers grade 1 to 5. Schools are organized into single- or multi-unit institutions; in other words, each institution may comprise more than one school. Standardized testing for evaluation purposes is compulsory in Italy since 2009 for all schools and students. INVALSI assessments considered in what follows cover math and language skills of pupils in second and fifth grade in a national administration lasting two days in the Spring, usually in May.[6] Scores in language and math are measured as number of correct answers, and their number varies by grade and year of test administration. In the empirical analysis we standardize them to have zero mean and unit variance by subject, year of survey and grade. Our statistical unit of analysis is the class since our manipulation variable varies at class level, as explained below. The sample selection replicates that in Angrist et al. (2014): the working sample includes only students attending public schools (more than 90% of the students in Italian primary schools) and consists of about 70,000 classes in each of the two grades covered by three years of data.

Measuring Manipulation

Class-level indicators of compromised scores are defined using within-class information on average and standard deviation of test scores, proportion of items missing, and variability in response patterns (measured by a Herfindahl index). These indicators are used as inputs for a cluster analysis that flags as suspicious classes with abnormally high performance, a small dispersion of scores, a low proportion of missing items, and high concentration in

[6]The testing procedure and its implementation are described in details in the annual reports of INVALSI (see http://www.invalsi.it).

response patterns. This procedure generates a dummy variable indicating classes where score manipulation seems likely, separately for math and language. Our manipulation indicator is similar to that used by Quintano et al. (2009) and employed by INVALSI except that the latter produces a continuous class-level probability of manipulation.[7] The manipulation score indicator might be however affected by misclassification, as implied by the statistical procedure used. We will come back to this point in Section 3.

The Monitoring Experiment

In an effort to increase test reliability, INVALSI randomly selects institutions to be observed by an external monitor. Every year about 7% of classes and 20% of institutions in the country are mandated to external control on the test day. Compliance of institutions is enforced by the Italian law; monitors supervise test administration and are responsible for score sheet transcription in selected classes within schools, which are however not randomly chosen within institutions (as evident from descriptives in Bertoni et al. 2013). The allocation of external monitors to classes follows a two-stage design. First, a sample of institutions stratified by region is selected with probability proportional to grade enrollment; then in sampled institutions one or two classes by grade (depending on total enrollment) are assigned an external monitor. Although within-institution monitoring is supposed to preserve randomness, in practice it appears to be contaminated by negotiation between school principals and INVALSI.

Monitors are selected by a pool of retired teachers and principals who did not have direct contacts with the schools or worked in town in the two years preceding the test. Monitors have two main duties: supervise test administration and ensure compliance with INVALSI testing standards on the one hand, and perform score sheet transcription on the other hand. Tests without monitors (the majority in the data) are proctored by local school staff, under the rule set by INVALSI that tests are not administered by the class teacher herself, but by a different teacher (of the same school) specialized in a different subject. Proctors and other teachers are expected to copy students' original responses onto machine-readable answer

[7] Our procedure follows Jacob and Levitt (2003) who use patterns of answers within and across tests in a classroom to detect manipulation. When manipulation is proxied by an indicator of concentration of answers within the class, conclusions are informationally equivalent to those presented in what follows. For details on methods and formulas used to classify score manipulation see the appendix in Angrist et al. (2014).

sheets (called *scheda risposta*), which are then sent to INVALSI. The transcription procedure is needed because this task is not totally mechanical. Questions come in the form of multiple choice and open-ended items. Answers to open questions have to be judged by transcribers as correct, wrong or missing, thus making transcription a form of grading. This transcription procedure opens the door to score manipulation, as does the fact that INVALSI tests are typically proctored by teachers and no further checks are done on the similarity between student's original responses and *scheda risposta* sent to INVALSI.[8] Angrist et al. (2014) show that score manipulation in Italian primary schools reflects teacher behavior, and identify shirking rather than accountability concerns as main motivation.

Table 1 shows descriptive statistics for the estimation sample. Scores are lower in classes in monitored institutions and even more so in monitored classes. Classes in the South have higher scores than in the rest of Italy, but not in monitored classes. Finally, manipulation rates are higher in the South and in math, and not surprisingly much lower in classes with external monitor and in monitored institutions.

3 Bounds on the percentage of manipulators

Notation

Let $Y_{ij,1}$ and $Y_{ij,0}$ be test scores for class i in institution j with and without manipulation, respectively. Scores take values in the interval $[k_0, k_1]$. The observed score is $Y_{ij} = Y_{ij,0}(1 - M_{ij}) + Y_{ij,1}M_{ij}$, where M_{ij} is an indicator for manipulation. The latter variable is unobserved, but a proxy W_{ij} of this indicator is available. Monitored institutions are denoted by the dummy Z_j. Finally, D_{ij} indicates whether class i in institution j is monitored. As discussed in the previous section, Z_j is randomly assigned but classes are selected non randomly for monitoring. Data come in the form $(Y_{ij}, D_{ij}, W_{ij}, Z_j)$, and class is our unit of analysis.

[8] No penalization from INVALSI is expected for classes suspected of manipulation. Only from school year 2011-12, INVALSI has actually used the score manipulation indicator to "correct" the average class scores. In classes with manipulation indicator above a certain threshold set by INVALSI average class scores are not returned to the school; in classes with manipulation indicator in a given range, again set up by INVALSI, class average score are returned appropriately adjusted according to the extent of manipulation detected. However, this procedure was unknown at the time the test took place and therefore our data are not affected by changes in teachers behavior due to the threat of scores' adjustment.

Effects of manipulation

Monitors reduce manipulation markedly, as can be seen in Table 2. The first row reports the coefficient on the dummy Z_j from the following regression:

$$W_{ij} = \rho_0 + \alpha Z_j + \rho_1 X_{ij} + \varepsilon_{ij}, \tag{1}$$

where X_{ij} contains a full set of grade and year effects as well as the stratification variables used in the monitoring experiment (region, grade enrollment at institution and their interactions). Here and in what follows standard errors are clustered on institution, which we reckon to be a conservative strategy in this context. OLS results are presented for Italy first and are then disaggregated by area, pooling Northern and Central regions and keeping Southern regions separated. The monitoring effects are shown in columns 1-3 for math and columns 4-6 for language, which suggest monitoring reduces manipulation rates by about 2% to 2.5% for Italy. Effects are considerably more pronounced in the South.

The second row in Table 2 reports the coefficient on D_{ij} from the following regression:

$$W_{ij} = \tau_0 + \beta D_{ij} + \tau_1 X_{ij} + \zeta_{ij},$$

which we estimate using 2SLS instrumenting D_{ij} with Z_j to correct for the endogenous choice of monitored classes within institutions. The coefficient estimated identifies the effect of class monitoring on manipulation. The effect is important and estimated at 7% to 8% for Italy. As before, the size of the effect doubles in the South.

Table 3 replicates the same analysis presented in Table 2 using scores on the left hand side of the equations considered. Institutional monitoring reduces math scores by 0.18σ, while the estimated monitoring effect on language scores is about -0.16σ. Here too effects are much larger in the South, ranging from -0.29σ for math to -0.26σ for language, estimates that appear in columns 3 and 6 of the table, respectively. Not surprisingly, the presence of an external monitor in class produces effects on tests scores which are more pronounced. Class monitoring reduces test score by about 0.6σ for math and 0.5σ for language in Italy, with effects twice as large in the South.

Misclassified manipulation

Assuming that classification error is unrelated to monitoring, we can use the monitoring experiment to bound the extent of true manipulation. This idea is formally stated through an exclusion restriction, implying that the correlation between monitors and W_{ij} only depends on manipulation M_{ij}. Recall that the monitoring sampling design implies $P(D_{ij} = 0|Z_j = 0) = 1$. Our key assumption is:

Assumption 1. *For $d = 0, 1$:*

$$\begin{aligned} P(W_{ij} = 1|M_{ij} = 0, D_{ij} = d, Z_j = 1) &= P(W_{ij} = 1|M_{ij} = 0, Z_j = 0), \\ &= P(W_{ij} = 1|M_{ij} = 0) \equiv 1 - \pi_0, \\ P(W_{ij} = 1|M_{ij} = 1, D_{ij} = d, Z_j = 1) &= P(W_{ij} = 1|M_{ij} = 1, Z_j = 0), \\ &= P(W_{ij} = 1|M_{ij} = 1) \equiv \pi_1. \end{aligned}$$

The terms π_1 and π_0 are probabilities of correct detection of manipulated and honest scores, respectively. We also assume that monitoring eliminates manipulation.

Assumption 2. $P(M_{ij} = 1|D_{ij} = 1) = 0$.

Under Assumption 1 and Assumption 2, π_0 is identified from:

$$P(W_{ij} = 1|D_{ij} = 1) = P(W_{ij} = 1|M_{ij} = 0, D_{ij} = 1, Z_j = 1) = 1 - \pi_0.$$

Define $p_z \equiv P(M_{ij} = 1|Z_j = z)$ as the true manipulation rate of interest, for $z = 0, 1$. Since:

$$P(W_{ij} = 1|Z_j = z) = (1 - \pi_0) + (\pi_0 + \pi_1 - 1)p_z,$$

it follows that p_z is linked to π_1 by:

$$p_z = \frac{P(W_{ij} = 1|Z_j = z) - P(W_{ij} = 1|D_{ij} = 1)}{\pi_1 - P(W_{ij} = 1|D_{ij} = 1)}. \tag{2}$$

This expression shows how knowledge of misclassification rates allows us to construct true manipulation rates. We also impose that W_{ij} is better than a coin toss.

Assumption 3. $\pi_1 \geq 0.5$.

Knowledge of π_0 together with Assumption 3 produce useful bounds on misclassification rates, which can then be used to bound true manipulation rates. In particular the upper

bound on p_z, which we denote by $\bar{p}_z$, is obtained when $\pi_1 = 0.5$. Assumptions 1-3 therefore allow to determine the maximum extent of manipulated data in monitored and unmonitored institution, which is a condition needed to develop bounds from the corrupt sampling model in Section 4.

Counting rotten apples

Bounds on p_z obtained under Assumptions 1-3 are presented in Table 6, for institutions with and without monitors. Results are presented for values of π_1 consistent with bounds developed in Section 4. We shall see that the monitoring experiment conveys information on the largest possible value of π_1 which is consistent with our data. It follows that the range of possible values is $0.5 \leq \pi_1 \leq 0.7$ for math scores in Northern and Central regions, and $\pi_1 \geq 0.5$ is all remaining cases.

The first row of Panel B and Panel C in Table 6 reports the fraction of presumed manipulators as computed from raw data for monitored and unmonitored institutions, respectively. The second row reports bounds on the true percentage of manipulators for the same subgroups. Here and in what follows bounds are estimated by replacing population probabilities or conditional expectations with their empirical analogues. As the focus is identification, we ignore the problem of sampling variability and of drawing inference for partially identified models (see, for example, Horowitz and Manski 2000, Imbens and Manski 2004, and Molinari 2008).

When monitored institutions are considered the manipulation rate for math in the South ranges between 5.0% and 11.0%, and is estimated at 9.0% in raw data. In unmonitored institutions, it ranges between 12.0% and 26.0%, and is estimated at 16.0% in raw data. By constructing:

$$p_1 - p_0 = \frac{P(W_{ij} = 1|Z_j = 1) - P(W_{ij} = 1|Z_j = 0)}{\pi_1 - P(W_{ij} = 1|D_{ij} = 1)},$$

we can bound the effect of external monitoring on manipulation. For math in the South this is found to range between -7.0% and -14.7%, which should be compared to the point estimate of -5.1% shown in column (3) of Table 2. The usual attenuation bias result induced by classification errors applies to regression estimates based on equation (1), hence characterizing the extent of the bias of 2SLS estimates of Y on W using Z as instrument (see Kane et al. 1999). The same analysis for math in Northern and Central regions yields bounds that are

markedly different, pointing to manipulation in unmonitored institutions between 2.0% and 3.0% (2.0% if computed from raw data). It is also clear from the table that manipulation can be ignored when monitored institutions are considered, being at most 1.0%. The effect $p_1 - p_0$ for math in Northern and Central regions varies between -1.2% and -1.7%.

The same pattern applies to manipulation of language score. The effect $p_1 - p_0$ ranges between -5.2% and -10.7% in the South, and between -1.0% and -2.1% in Northern and Central regions.

4 Bounds on scores

Naive bounds

The quantity of interest is the average of $Y_{ij,0}$ across classes in the population. Conditional on the strata used for random assignment of monitors, we have:[9]

$$E(Y_{ij,0}|Z_j = 0) = E(Y_{ij,0}|Z_j = 1). \tag{3}$$

We start by assuming that manipulation is aimed at boosting scores, a behavioral restriction motivated by background evidence from INVALSI data. Differences in observed scores between Northern and Southern regions, the existence of substantial manipulation in the South and the evidence from international surveys documented in the Introduction suggest that manipulated scores are higher than true scores. Moreover, the assumption is consistent with the evidence in Angrist et al. (2014), who show that manipulation is primarily the result of curbstoning by teachers during transcription.

Assumption 4. $Y_1 \geq Y_0$.

This is the Monotone Treatment Response assumption in Manski and Pepper (2000), and amounts to stating that each class' reported score is weakly increasing with manipulation regardless of the extent of monitoring at institution. Since

$$E(Y_{ij,0}|Z_j = z) = E(Y_{ij,0}|M_{ij} = 0, Z_j = z)[1 - p_z] + E(Y_{ij,0}|M_{ij} = 1, Z_j = z)p_z, \tag{4}$$

[9] In our empirical exercise population figures are obtained using sampling weights constructed from strata controls. To ease notation, reference to the sampling scheme and the use of sampling weights will be left implicit in what follows.

Assumption 4 implies that the average of observed scores is above the average of true scores. Let the following sets be defined for $z = 0, 1$:

$$\mathcal{I}_{1z} = \{x : k_0 \leq x \leq E(Y_{ij}|Z_j = z)\}.$$

Naive bounds on $E(Y_{ij,0})$ under Assumption 4 are defined as follows:

$$\mathcal{I}_1 = \mathcal{I}_{10} \cap \mathcal{I}_{11}.$$

These are presented in the second row of each panel in Table 4 setting k_0 and k_1 to the minimum and maximum values of Y_{ij} in the data, respectively (in the first row observed scores are reported to ease the comparison). Much uncertainty about the true ranking of scores across areas is revealed, for both math and language. This holds for both monitored (Panel A) and unmonitored institutions (Panel B) and therefore for the intersection of their bounds, i.e. considering all institutions (Panel C). The range of variation for scores in monitored institutions is within that observed without monitors, with one exception. It follows that the intersection of bounds constructed from the two groups coincides in most cases with bounds obtained for monitored institutions. Despite the coarse information employed, these bounds eliminate regional differences by subject observed in raw data. For example, observed math scores in Table 1 are -0.12σ and 0.21σ for North/Centre and South, respectively, pointing to figures well above the average in the latter area. We now learn that test scores in math in the South are between $[-4.57\sigma, -0.05\sigma]$, the upper bound being considerably below the value of scores computed from raw data. Still results don't reveal any ranking of areas with respect to their performance by subject. Bounds are however informative about the distance between true scores and the national average computed from raw scores, as all signs in the bottom panel of Table 4 are negative.

Bounds from the corrupt sampling model

Results in HM allow to define sharp bounds on $E(Y_{ij,0}|Z_j = z)$ when (2) is limited from above by $\bar{p}_z$. As discussed in Section 3, such upper limit on $\bar{p}_z$ is achieved when $\pi_1 = 0.5$. Observed data $Y_{ij} = Y_{ij,0}(1 - M_{ij}) + Y_{ij,1}M_{ij}$ are drawn from a mixture of the distribution of interest, $F(y_0|Z_j = z)$, and another distribution which follows from manipulation, $F(y_1|Z_j = z)$. Mixing weights p_z define the percentage of corrupted data. Under the plausible scenario

that manipulators have scores $Y_{ij,0}$ selectively different from those of non-manipulators, this setting defines the corrupt sampling model of HM (for empirical applications of this idea see, for example, Kreider and Pepper 2007, 2008, and 2011). Corollary 4.1 in HM yields the following bounds for $z = 0, 1$:

$$E(Y_{ij}|Y_{ij} \leq Q_{1-\bar{p}_z}, Z_j = z) \leq E(Y_{ij,0}|M_{ij} = 0, Z_j = z) \leq E(Y_{ij}|Y_{ij} \geq Q_{\bar{p}_z}, Z_j = z),$$

where Q_τ is the τ-th quantile of $F(y|Z_j = z)$. Using (4) the following bounds on $E(Y_{ij,0}|Z_j = z)$ are defined:

$$\begin{aligned} LB_2(z) &\equiv (1 - \bar{p}_z)E(Y_{ij}|Y_{ij} \leq Q_{1-\bar{p}_z}, Z_j = z) + \bar{p}_z k_0, \\ UB_2(z) &\equiv (1 - \bar{p}_z)E(Y_{ij}|Y_{ij} \geq Q_{\bar{p}_z}, Z_j = z) + \bar{p}_z k_1, \end{aligned}$$

upon noting that in the absence of additional information we have:

$$k_0 \leq E(Y_{ij,0}|M_{ij} = 1, Z_j = z) \leq k_1. \tag{5}$$

Define the following sets for $z = 0, 1$:

$$\mathcal{I}_{2z} = \{x : LB_2(z) \leq x \leq UB_2(z)\},$$

Sets are defined for $\pi_1 = 0.5$, the lowest admissible value for π_1 implied by Assumption 3. As $\mathcal{I}_{2z}$ shrinks for increasing values of π_1, the identifying information conveyed by HM bounds depends on the lowest value of the parameter space for π_1. The quantity of interest $E(Y_{ij,0})$ lies in the intersection of the sets $\mathcal{I}_{2z}$ across all values $Z_j = z$ (Manski 1990). It follows that HM-like bounds on $E(Y_{ij,0})$ under Assumptions 1-4 are obtained as follows:

$$\mathcal{I}_2 = \mathcal{I}_{20} \cap \mathcal{I}_{21} \cap \mathcal{I}_1,$$

and are presented in the third row of each panel in Table 4.

The gain with respect to using only Assumption 4 is clear-cut. The result is driven by more informative lower bounds as - by construction - Assumption 4 yields more informative upper bounds than does the corrupt sampling model. For math in the South, for example, bounds narrow from $[-4.57\sigma, -0.05\sigma]$ to $[-0.84\sigma, -0.05\sigma]$, a 83% reduction in width for monitored institutions, and from $[-3.82\sigma, 0.29\sigma]$ to $[-1.24\sigma, 0.29\sigma]$, a 63% reduction for unmonitored institutions. Similar figures are found for language: 86% and 74% reduction in institutions

with and without monitors, respectively. In the North/Centre bounds shrink even more, a 98% reduction in monitored institutions and a 95% reduction in unmonitored institutions for both math and language. These results follow from the lower percentage of manipulators in the North, as documented in Table 6. The identifying power of HM-like lower bounds is apparent when we consider the ranking of scores across areas. Despite not being able to exclude that scores are identical across regions, lower bounds for math and language in the North are now relatively close to upper bounds in the South.

Bounds from partially verified data

HM-like bounds can be tightened using W_{ij} to partition observed scores into two sets, under the idea that values in one set are more likely to be determinations from the distribution of interest. Dominitz and Sherman (2006, DS in what follows) refer to this setting as corrupt sampling model with verification. In our setting, the verified set consists of scores with $W_{ij} = 0$. For $w = 0, 1$ denote by:

$$\gamma_{wz} = P(M_{ij} = 1 | W_{ij} = w, Z_j = z),$$

the percentage of manipulators in the $W_{ij} = 0$ and $W_{ij} = 1$ groups, respectively. Under Assumption 1 we have:

$$\gamma_{1z} = \frac{\pi_1 p_z}{P(W_{ij} = 1 | Z_j = z)}, \qquad \gamma_{0z} = \frac{(1 - \pi_1) p_z}{1 - P(W_{ij} = 1 | Z_j = z)},$$

which depend on the unknown π_1. Under Assumptions 1-3, it can be shown that the maximum values of γ_{1z} and γ_{0z}, which we denote by $\overline{\gamma}_{1z}$ and $\overline{\gamma}_{0z}$ respectively, are obtained when $\pi_1 = 0.5$. The derivation of bounds mirrors the steps followed above. First, write for $z = 0, 1$:

$$\begin{aligned} E(Y_{ij,0} | M_{ij} = 0, Z_j = z) &= E(Y_{ij,0} | M_{ij} = 0, W_{ij} = 0, Z_j = z)\pi_0 \\ &+ E(Y_{ij,0} | M_{ij} = 0, W_{ij} = 1, Z_j = z)[1 - \pi_0], \end{aligned} \tag{6}$$

where - differently from DS - mixture weights are known due to Assumptions 1-2. Second, bound mixture components in (6) following the argument of HM conditional on W_{ij}. Finally,

use (4) and (6) to define the following bounds on $E(Y_{ij,0}|Z_j = z)$:

$$\begin{aligned} LB_3(z) &\equiv (1-\bar{p}_z)\pi_0 E(Y_{ij}|Y_{ij} \leq Q^0_{1-\bar{\gamma}_{0z}}, W_{ij}=0, Z_j=z) \\ &+ (1-\bar{p}_z)(1-\pi_0)E(Y_{ij}|Y_{ij} \leq Q^1_{1-\bar{\gamma}_{1z}}, W_{ij}=1, Z_j=z) + \bar{p}_z k_0, \\ UB_3(z) &\equiv (1-\bar{p}_z)\pi_0 E(Y_{ij}|Y_{ij} \geq Q^0_{\bar{\gamma}_{0z}}, W_{ij}=0, Z_j=z) \\ &+ (1-\bar{p}_z)(1-\pi_0)E(Y_{ij}|Y_{ij} \geq Q^1_{\bar{\gamma}_{1z}}, W_{ij}=1, Z_j=z) + \bar{p}_z k_1, \end{aligned}$$

where Q^w_τ is the τ-th quantile of $F(y|W_{ij}=w, Z_j=z)$ for $w=0,1$. Define the following sets for $z=0,1$:

$$\mathcal{I}_{3z} = \{x : LB_3(z) \leq x \leq UB_3(z)\},$$

which, as before, are defined for $\pi_1 = 0.5$ and shrink as π_1 increases. DS-like bounds on $E(Y_{ij,0})$ under Assumptions 1-4 are obtained from intersections as follows:

$$\mathcal{I}_3 = \mathcal{I}_{30} \cap \mathcal{I}_{31} \cap \mathcal{I}_2,$$

and are presented in the last row of each panel in Table 4.

Given the small fraction of manipulators, score verification for Northern and Central regions is expected to have less impact on bounds than it may have for the South. A similar idea applies to differences between monitored and unmonitored institutions. Bounds are narrowed by about 4% (1%) in monitored (unmonitored) institutions in the South, for both math and language, as shown in the last row of each panel in Table 4. We find that the gain in width is overall contained, and confined to the South.

Implications for the classification error

The identity in (3) imposes restrictions on the support of π_1. In particular, values of the parameter space for π_1 which correspond to an empty set $\mathcal{I}_3$ are not plausible. Since $\mathcal{I}_{30}$ and $\mathcal{I}_{31}$ shrink as π_1 increases, this condition might result in an upper bound for the parameter space of π_1. Knowledge of this upper bound does not affect the HM-like and DS-like bounds. It is however interesting to study the identifying power of this result when combined with assumptions that define bounds alternative to those from the corrupt sampling model.

The implied ranges for π_1 in our data are $[0.5, 0.7]$ for math scores in North/Centre, and $[0.5, 1]$ in all remaining cases. These ranges are consistent with having more misclassification in the North/Centre than in the South. For example the percentage of true math manipula-

tors amongst those deemed to manipulate in unmonitored institutions, γ_{10}, ranges between 66.9% and 67.2% in the North/Centre and 76.3% and 80.0% in the South. Such differential patterns help interpret the manipulation coefficients in Table 7 of Angrist et al. (2014).

5 Refining bounds on scores

Restrictions on the manipulation indicator

Under the assumption that W_{ij} does not contain more information on scores than the latent indicator M_{ij}, it is possible to define alternative bounds on $E(Y_{ij,0}|M_{ij}=0,Z_j=z)$.

Assumption 5. *For $z=0,1$ and $w=0,1$:*

$$\begin{aligned} E(Y_{ij,0}|M_{ij}=0,W_{ij}=w,Z_j=z) &= E(Y_{ij,0}|M_{ij}=0,Z_j=z), \\ E(Y_{ij,1}|M_{ij}=1,W_{ij}=w,Z_j=z) &= E(Y_{ij,1}|M_{ij}=1,Z_j=z). \end{aligned}$$

This assumption is often referred to as non-differential misclassification and qualifies W_{ij} as a surrogate of M_{ij} (see Carroll et al., 2006, and Chen et al., 2011). The conditioning on $Z_j=z$ plays an important role in light of the endogenous determination of M_{ij}. Manipulators can be selectively different with and without monitors at institution, and this may result in different distributions of $Y_{ij,0}$ for correct reporters $M_{ij}=0$. Moreover reported scores for manipulators $Y_{ij,1}$ may differ with and without external monitoring. Monitors at institution not only reduce the percentage of classes with manipulated scores, but may also change the composition of manipulators as well as the reporting of scores.[10]

To fix ideas, assume that π_1 is known. Under Assumption 5, for $z=0,1$ we have:

$$\begin{aligned} E(Y_{ij}|W_{ij}=0,Z_j=z) &= E(Y_{ij,0}|M_{ij}=0,Z_j=z)(1-\gamma_{0z})+E(Y_{ij,1}|M_{ij}=1,Z_j=z)\gamma_{0z}, \\ E(Y_{ij}|W_{ij}=1,Z_j=z) &= E(Y_{ij,0}|M_{ij}=0,Z_j=z)(1-\gamma_{1z})+E(Y_{ij,1}|M_{ij}=1,Z_j=z)\gamma_{1z}, \end{aligned}$$

[10]The setting considered for our analysis has interesting connections with the literature on identification with imperfect instruments. To fix ideas, assume W_{ij}=M_{ij} and that the causal effect of manipulation on scores $E(Y_{ij,1}-Y_{ij,0}|M_{ij}=1)$ is the parameter of interest. Under a valid exclusion restriction Z_j can be used as instrument for M_{ij} in the relationship between Y_{ij} and M_{ij}. This restriction is violated if the extent of manipulation in unmonitored classes depends on the presence of monitors at institution, for example because $Y_{ij,1}$ is decreasing in $Z_j=z$. Assumptions can be made to sign the role of unobservables that cause such violation, as in Nevo and Rosen (2012). These assumptions yield partial identification of the parameter of interest.

which defines a system of two equations in two unknowns. It is easy to show that the system has a solution if $\gamma_{1z} \neq \gamma_{0z}$, a condition implied by Assumption 3.[11] It follows that:

$$\begin{aligned} E(Y_{ij,0}|M_{ij}=0,Z_j=z) &= (1-\Gamma_{0z})E(Y_{ij}|W_{ij}=0,Z_j=z)+\Gamma_{0z}E(Y_{ij}|W_{ij}=1,Z_j=z), \\ E(Y_{ij,1}|M_{ij}=1,Z_j=z) &= (1-\Gamma_{1z})E(Y_{ij}|W_{ij}=0,Z_j=z)+\Gamma_{1z}E(Y_{ij}|W_{ij}=1,Z_j=z), \end{aligned}$$

where:

$$\Gamma_{0z} \equiv -\frac{\gamma_{0z}}{\gamma_{1z}-\gamma_{0z}}, \qquad \Gamma_{1z} \equiv \frac{1-\gamma_{0z}}{\gamma_{1z}-\gamma_{0z}}.$$

Bounds on $E(Y_{ij,0})$ for $z=0,1$ under Assumptions 1-5 can be constructed from (4) by varying π_1 over its support:[12]

$$\begin{aligned} LB_4(z) &\equiv \inf_{\pi_1}\{(1-\Gamma_{0z})E(Y_{ij}|W_{ij}=0,Z_j=z)[1-p_z] \\ &+ \Gamma_{0z}E(Y_{ij}|W_{ij}=1,Z_j=z)[1-p_z]+k_0p_z\}, \\ UB_4(z) &\equiv \sup_{\pi_1}\{(1-\Gamma_{0z})E(Y_{ij}|W_{ij}=0,Z_j=z)[1-p_z] \\ &+ \Gamma_{0z}E(Y_{ij}|W_{ij}=1,Z_j=z)[1-p_z]+k_1p_z\}, \end{aligned}$$

and then taking the intersection of resulting bounds for $z=0,1$:

$$\mathcal{I}_4 = \mathcal{I}_{40} \cap \mathcal{I}_{41} \cap \mathcal{I}_3,$$

where:

$$\mathcal{I}_{4z} = \{x : LB_4(z) \leq x \leq UB_4(z)\}.$$

Differently from bounds constructed under the corrupt sampling model, $\mathcal{I}_4$ and $\mathcal{I}_3$ are not necessarily nested. In our empirical exercise, however, we find that Assumption 5 reduces the upper bound for the parameter of interest.

Results are presented in the second row of each panel of Table 5 (in the first row observed scores are reported to ease the comparison). The identifying power of Assumption 5 is investigated separately from that conveyed by the contaminated sampling model, ignoring

[11] Since there is $P(W_{ij}=0|Z_j=z)>0$ in our data, the condition $\gamma_{1z} \neq \gamma_{0z}$ is met if $\pi_1 \neq P(W_{ij}=1|Z_j=z)$. This requirement is implied by Assumption 3, as in our data the maximum value of $P(W_{ij}=1|Z_j=z)$ across z's is 0.16. This can be equivalently stated by saying that classes with $W_{ij}=1$ are more likely to have $M_{ij}=1$ than classes with $W_{ij}=0$, which is A1 in DS.

[12] It may be shown that Γ_0 increases in π_1 and that $\Gamma_0<0$. It follows that if $E[Y_{ij}|W_{ij}=1,Z_j=z] > E[Y_{ij}|W_{ij}=0,Z_j=z]$, a condition which always holds in our empirical application, $E[Y_{ij,0}|M_{ij}=0,Z_j=0]$ also increases in π_1. This, since p_z decreases in π_1, implies that $LB_4(z)$ is obtained when $\pi_1=0.5$.

intersections with bounds presented in Section 4. Assumption 4 is instead maintained in the derivation of bounds throughout the table. Panel C of Table 5 almost depicts a reversal in the regional ranking of language scores with respect to raw data. For math scores, bounds in the South largely overlap to bounds in the North/Centre and it is not possible to establish a clear regional ranking.

Behavioral restrictions

Restrictions on the origin of manipulation can be used to tighten bounds derived in the previous section. Suppose, for example, that scores are manipulated more often in classes with low average achievement. This restricts the support of $E(Y_{ij,0}|M_{ij} = 1, Z_j = z)$ previously stated in (5). We state this assumption in the form of multiplicative mean independence, allowing two conditional means to differ by a factor of proportionality. This is in the spirit of Kreider and Pepper (2011), although random assignment of monitors to institutions adds to the informational content of this assumption.

We start from the following re-parametrization for $z = 0, 1$:

$$E(Y_{ij,0}|M_{ij} = 1, Z_j = z) = \delta_z E(Y_{ij,0}|M_{ij} = 0, Z_j = z),$$

that relates true scores for manipulators to true scores of honest reporters. Exogeneity of manipulation is equivalent to $\delta_z = 1$, and defines the contaminated sampling model of HM. Equation (4) implies:

$$E(Y_{ij,0}|Z_j = z) = E(Y_{ij,0}|M_{ij} = 0, Z_j = z)[1 - p_z(1 - \delta_z)], \tag{7}$$

and since Z_j is assigned at random:

$$E(Y_{ij,0}|M_{ij} = 0, Z_j = 1)[1 - p_1(1 - \delta_1)] = E(Y_{ij,0}|M_{ij} = 0, Z_j = 0)[1 - p_0(1 - \delta_0)]. \tag{8}$$

If manipulation is exogenous, under Assumption 5 the last equation can be solved for the only unknown π_1. It follows that Assumptions 1-5 yield point identification in the contaminated sampling model of HM. When manipulation is not exogenous, the identification breakdown is evident from equation (8). The following assumption restricts the composition of manipulators, as corrupted data are assumed to come from the lower end of the distribution $Y_{ij,0}$. Calculations to derive bounds in the previous sections, not presented here,

show that the upper bound on $E(Y_{ij,0}|M_{ij}=0,Z_j=z)$ is always negative. The condition $E(Y_{ij,0}|M_{ij}=0,Z_j=z) \geq E(Y_{ij,0}|M_{ij}=1,Z_j=z)$ in this setting is therefore equivalent to $\delta_z \geq 1$.

Assumption 6. *$\delta_0 \geq 1$ and $\delta_1 \geq 1$.*

Bounds under Assumption 6 are obtained by varying the three unknowns δ_0, δ_1 and π_1 in their corresponding parameter space and by imposing equation (4).[13] For $z=0,1$ define:

$$\begin{aligned} LB_5(z) &\equiv \inf_{\pi_1,\delta_z}\{E(Y_{ij,0}|M_{ij}=0,Z_j=z)[1-p_z(1-\delta_z)]\}, \\ UB_5(z) &\equiv \sup_{\pi_1,\delta_z}\{E(Y_{ij,0}|M_{ij}=0,Z_j=z)[1-p_z(1-\delta_z)]\}, \end{aligned}$$

and:

$$\mathcal{I}_{5z} = \{x : LB_5(z) \leq x \leq UB_5(z)\}.$$

Bounds on $E(Y_{ij,0})$ under Assumptions 1-6 are defined as follows:

$$\mathcal{I}_5 = \mathcal{I}_{50} \cap \mathcal{I}_{51} \cap \mathcal{I}_4.$$

Results are presented in the third row of each panel of Table 5. By construction Assumption 6 affects the width of bounds only by changing their upper limit, and unveils geographic differences in scores that revert the picture obtained from raw data. Language scores for the North/Centre are now in the $[-0.23\sigma, -0.15\sigma]$ interval, which is above the interval $[-1.02\sigma, -0.38\sigma]$ found for the South. The interval for math scores in the South, $[-1.03\sigma, -0.28\sigma]$, overlaps only partially with the interval $[-0.29\sigma, -0.26\sigma]$ for the Northern and Central regions.

6 Implications and Directions for Further Work

Table 6 summarizes the main conclusions from our analysis. Panel A presents bounds on $E(Y_{ij,0})$ that result from taking the intersection of bounds in Table 4 and Table 5. These are our best estimates of bounds on the parameter of interest. For both math and language

[13] The derivation is obtained in two steps. Using results from the previous section, it is easy to show that the quantity in (7) is minimized for the largest value of δ_z when $\pi_1 = 0,5$. We use $E(Y_{ij,0}|Z_j=z) \geq k_0$ to limit the set of possible values that can be taken by δ_z. A similar argument shows that the quantity is maximized when $\delta_z = 1$ and π_1 is set to its maximum value. Equation (8) implies that the intersection of bounds constructed for $E(Y_{ij,0}|Z_j=0)$ and $E(Y_{ij,0}|Z_j=1)$ must not be empty.

bounds are sharp enough to provide a ranking of areas in terms of performance at national tests. Bounds are tighter in Northern and Central regions; most importantly, there is almost no overlap of bounds across areas. This allows us to conclude that true scores in the South are consistently lower than in the rest of Italy. The comparison with average scores computed from raw data, also reported in Table 6, reveals that regional inequality is reverted after the adjustment.

The issue is further explored in the central panels of Figure 1 and Figure 2. Maps are derived with a descriptive purpose and offer a graphical inspection of the distribution of test scores across areas after our correction. When monitored institutions are considered, some regions in Northern Italy present little variability of the indicator W_{ij}. For these regions the map reports observed average scores in monitored institutions.[14] For all remaining regions we compute area-specific bounds replicating the analysis presented in the previous sections, and report the mid point of the interval. The central panels for adjusted scores should be compared with the left panels obtained for observed scores. The ranking of regions in terms of academic performance of students is reversed once manipulation is taken into account, and matches that obtained from TIMSS and PIRLS (see INVALSI, 2011). The correlation of ranks across the 20 Italian regions before and after the adjustment is -14% and -9% for math and language scores, respectively.[15]

The second row in Panel A of Table 6 together with Figure A2 give insight on the geographical distribution of scores after the adjustment used by INVALSI in official publications. As explained above, their procedure builds upon Quintano et al. (2009) to derive a continuous class-level probability of presumed manipulation. The latter is then used to define weights for all classes according to the following steps (see Falzetti, 2013). First manipulation is bench-marked against Veneto, the region with lowest presumed manipulation. All classes in

[14]For example Veneto, the region with the lowest level of measured manipulation, has 8 classes in monitored institutions with $W_{ij} = 1$ for math out of $2,505$ classes (0.32%) in our data. Out of $1,878$ classes in monitored institutions in Sicily, 199 have $W_{ij} = 1$ for math scores (10.6%). We use observed math scores from monitored institutions for the following seven regions: Piemonte, Lombardia, Trentino Alto-Adige, Veneto, Friuli Venezia-Giulia, Toscana, and Lazio. For language scores we use raw data for the following three regions: Lombardia, Trentino Alto-Adige and Friuli Venezia-Giulia.

[15]Figure A1 in the Appendix is obtained following the same procedure, but reporting the lower bound for Northern regions and the upper bound for Southern regions. This represents the worse case scenario for detecting differences across areas, but still conveys a message similar to that in the central panels of Figure 1 and Figure 2. Using Figure A1 in the Appendix, the correlation of ranks before and after the adjustment is 34% and -3% for math and language scores, respectively.

Italy with value of the manipulation probability below the median value in Veneto are given weight one. It is worth noting that this implicitly admits the existence of classification errors in the manipulation probability. All remaining classes are weighted one minus the probability of presumed manipulation. The correction employed by INVALSI yields a correlation of ranks with raw data of 99% and 66% for math and language scores, respectively.[16]

Our correction heavily affects the ranking of regions because the effects of manipulation on scores are large, as shown in the last row of Panel B and Panel C of Table 6. Our setting allows to identify bounds on $E(Y_{ij,1} - Y_{ij,0}|M_{ij} = 1, Z_j = z)$ since $E(Y_{ij,1}|M_{ij} = 1, Z_j = z)$ and $E(Y_{ij,0}|M_{ij} = 1, Z_j = z)$ can be retrieved under Assumption 5 and Assumption 6. As explained, bounds are defined without exclusion restrictions for Z_j, and allow for mismeasurement of the indicator W_{ij}. Here too bounds are marginally tighter in Northern and Central regions, and in monitored institutions. Bounds for monitored institutions are however not disjoint from those for unmonitored institutions, pointing to effects of at least 3σ. The same conclusion applies if the intersection between monitored and unmonitored institutions is considered, assuming constant effects of manipulation with respect to Z_j. This finding is consistent with the idea that when manipulators rig scores, the result is independent of the presence of an external monitor at institution.

Why is the fact that score manipulation distorts regional rankings in Italy of general interest? Micro-data on student achievement are employed in empirical research to learn about the most effective determinants in the education production function. Figure 3 presents the association between observed and adjusted scores and selected proxies of family and school inputs. Only math scores are considered, as the figure for language scores conveys a similar message. We consider two indicators of family background: per-capita income (top panel) and an index of deprivation (central panel) distributed by the National Statistical Office. The bottom panel reports score profiles by pupil-to-teacher ratio, which we interpret as a proxy for public spending on education at primary school. The association between achievement and inputs is reverted by manipulation.

Our findings raise a number of questions, including why teacher manipulation is so much more prevalent in the South, and what can be done to enhance accurate assessment in Italy

[16] A variant to this procedure is also considered by INVALSI, and assigns weight zero to all classes with a probability value above 50%. Figure A3 in the Appendix shows results from this variant. In this case the correlations with observed scores are still very high (92% and 43% for math and language, respectively).

and elsewhere. It's also worth asking what are the determinants of low performance of students in the South, in light of the ongoing education policies in those areas (Objective 1 regions) eligible to receive EU Regional Development Funds and EU Social Funds (see, for example Battistin and Meroni, 2013) and the positive trend in PISA scores of some regions. We hope to answer these questions in future work.

References

Angrist, J. D., E. Battistin, and D. Vuri (2014): "In a Small Moment: Class Size and Moral Hazard in the Mezzogiorno," NBER Working Paper, 20173.

Aviv, R. (2014): "Wrong Answer: In an era of high-stakes testing, a struggling school made a shocking choice.," *The New Yorker, Annals of Education, July 21*, Accessed at: http://www.newyorker.com/magazine/2014/07/21/wrong-answer.

Battistin, E., and E. C. Meroni (2013): "Should We Increase Instruction Time in Low Achieving Schools? Evidence from Southern Italy," IZA Discussion Papers 7437, Institute for the Study of Labor.

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

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

Cannari, L., F. Nucci, and P. Sestito (2000): "Geographic labour mobility and the cost of housing: evidence from Italy," *Applied Economics*, 32(14), 1899–1906.

Carroll, R., D. Ruppert, L. Stefanski, and C. Crainiceanu (2006): *Measurement Error in Nonlinear Models, A Modern Perspective, Second Edition*. Chapman & Hall.

Chen, X., X. Hong, and D. Nekipelov (2011): "Nonlinear Models of Measurement Errors," *Journal of Economic Literature*, 49, 901–937.

CNEL (2013): "Relazione annuale 2013 al Parlamento e al Governo sui livelli e la qualita' dei servizi erogati dalle pubbliche amministrazioni centrali e locali alle imprese e ai cittadini," *Technical Report*.

Dee, T. S., B. A. Jacob, J. McCrary, and J. Rockoff (2011): "Rules and Discretion in the Evaluation of Students and Schools: The Case of the New York Re-

gents Examinations," Columbia Business School Research Paper. Available at SSRN: http://ssrn.com/abstract=1915387.

DiAddario, S., and E. Patacchini (2008): "Wages and the City. Evidence from Italy," *Labor Economics*, 15, 1040–1061.

Dominitz, J., and R. P. Sherman (2006): "Identification and estimation of bounds on school performance measures: a nonparametric analysis of a mixture model with verification," *Journal of Applied Econometrics*, 21(8), 1295–1326.

Falzetti, P. (2013): "L'esperienza di restituzione dei dati al netto del cheating," presentation at the Workshop "Metodi di identificazione, analisi e trattamento del cheating", 8 February, available at: http://www.invalsi.it/invalsi/ri/sis/documenti/022013/falzetti.pdff.

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

Hanushek, E., and L. Woessmann (2011): "The Economics of International Differences in Educational Achievement," in *Handbook of Economics of Education*, ed. by E. A. Hanushek, S. Machin, and L. Woessmann. Elsevier.

Horowitz, J. L., and C. F. Manski (1995): "Identification and Robustness with Contaminated and Corrupted Data," *Econometrica*, 63(2), 281–302.

——— (2000): "Nonparametric Analysis of Randomized Experiments With Missing Covariate and Outcome Data," *Journal of the American Statistical Association*, 95, 77–84.

Hu, Y. (2008): "Identification and estimation of nonlinear models with misclassification error using instrumental variables: a general solution," *Journal of Econometrics*, 144 (1), 27–61.

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

IEA (2011): "IEA's Progress in International Reading Literacy Study - Indagini IEA 2011 PIRLS e TIMSS," *Technical Report*.

Imbens, G., and C. F. Manski (2004): "Confidence Intervals for Partially Identified Parameters," *Econometrica*, 72, 1845–1857.

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

——— (2011): "Le rilevazioni IEA: i risultati degli studenti italiani nelle indagini internazionali PIRLS e TIMSS 2011," *Technical Report*.

——— (2012): "Sistema Nazionale di Valutazione - A.S. 2011/2012, La Rilevazione degli Apprendimenti," *Technical Report*.

——— (2013): "Sistema Nazionale di Valutazione - A.S. 2012/2013, La Rilevazione degli Apprendimenti," *Technical Report*.

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

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

Kreider, B., and J. V. Pepper (2007): "Disability and Employment: Reevaluating the Evidence in Light of Reporting Errors," *Journal of the American Statistical Association*, 102, 432–441.

——— (2008): "Inferring Disability Status From Corrupt Data," *Journal of Applied Econometrics*, 23, 329–349.

——— (2011): "Identification of Expected Outcomes in a Data Error Mixing Model With Multiplicative Mean Independence," *Journal of Business & Economic Statistics*, 29(1), 49–60.

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

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

Manski, C., and J. Pepper (2000): "Monotone instrumental variables: with an application to the returns to schooling," *Econometrica*, 68(4), 997–1010.

Manski, C. F. (1990): "Nonparametric Bounds on Treatment Effects," *American Economic Review Papers and Proceedings*, 80, 319–323.

Molinari, F. (2008): "Partial Identification of Probability Distributions With Misclassified Data," *Journal of Econometrics*, 144, 81–117.

Nannicini, T., A. Stella, G. Tabellini, and U. Troiano (2013): "Social Capital and Political Accountability," *American Economic Journal: Economic Policy*, 5, 222–250.

Nevo, A., and A. M. Rosen (2012): "Identification With Imperfect Instruments," *The Review of Economics and Statistics*, 97(3), 659–671.

PISA (2012): "Results in Focus: What 15-year-olds know and what they can do with what they know: Key results from PISA 2012," *OECD Technical Report*.

Putnam, R., R. Leonardi, and R. Nanetti (1993): *Making Democracy Work*. Princeton University Press, Princeton.

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

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

Table 1: Descriptive Statistics

	Math			Language		
	Italy	North/Centre	South	Italy	North/Centre	South
	(1)	(2)	(3)	(4)	(5)	(6)
	A. Monitored institutions					
Observed score	-0.14	-0.20	-0.04	-0.10	-0.09	-0.11
	(0.91)	(0.74)	(1.15)	(0.92)	(0.78)	(1.13)
Presumed manipulators	0.04	0.01	0.09	0.03	0.01	0.07
	(0.19)	(0.11)	(0.28)	(0.18)	(0.12)	(0.26)
N	33,267	21,589	11,678	33,267	21,589	11,678
	B. Monitored classes					
Observed score	-0.33	-0.30	-0.37	-0.27	-0.18	-0.42
	(0.82)	(0.66)	(1.03)	(0.86)	(0.72)	(1.03)
Presumed manipulators	0.02	0.01	0.04	0.02	0.01	0.03
	(0.13)	(0.08)	(0.19)	(0.12)	(0.07)	(0.18)
N	9,630	6,030	3,600	9,630	6,030	3,600
	C. Unmonitored institutions					
Observed score	0.04	-0.10	0.28	0.03	-0.01	0.09
	(1.02)	(0.80)	(1.27)	(1.02)	(0.83)	(1.27)
Presumed manipulators	0.07	0.02	0.16	0.06	0.03	0.12
	(0.26)	(0.15)	(0.36)	(0.24)	(0.16)	(0.33)
N	106,743	65,909	40,834	106,743	65,909	40,834
	D. All institutions					
Observed score	-0.00	-0.12	0.21	-0.00	-0.03	0.05
	(1.00)	(0.79)	(1.25)	(1.00)	(0.82)	(1.24)
Presumed manipulators	0.07	0.02	0.14	0.06	0.02	0.11
	(0.25)	(0.14)	(0.35)	(0.23)	(0.15)	(0.32)
N	140,010	87,498	52,512	140,010	87,498	52,512

Note. This table shows descriptive statistics for a sample pooling second and fifth graders.

Table 2: Monitoring Effects on Score Manipulation

	Math			Language		
	Italy	North/Centre	South	Italy	North/Centre	South
	(1)	(2)	(3)	(4)	(5)	(6)
Monitor at institution (Z_j)	-0.024***	-0.008***	-0.051***	-0.021***	-0.010***	-0.039***
	(0.002)	(0.001)	(0.006)	(0.002)	(0.002)	(0.005)
Monitor in class (D_{ij})	-0.081***	-0.029***	-0.163***	-0.070***	-0.036***	-0.124***
	(0.008)	(0.005)	(0.018)	(0.007)	(0.005)	(0.016)
N	139,996	87,491	52,505	140,003	87,493	52,510

Note. This table shows the effect of monitors on score manipulation. Class-level effects are 2SLS estimates using the presence of institutional monitors as an instrument. Robust standard errors, clustered on institution, are shown in parentheses. All regressions include year and grade fixed effects and sampling strata controls (grade enrollment at institution, region dummies and their interactions). * significant at 10%; ** significant at 5%; *** significant at 1%.

Table 3: Monitoring Effects on Test Scores

	Math			Language		
	Italy	North/Centre	South	Italy	North/Centre	South
	(1)	(2)	(3)	(4)	(5)	(6)
Monitor at institution (Z_j)	-0.179***	-0.118***	-0.289***	-0.158***	-0.103***	-0.256***
	(0.012)	(0.011)	(0.027)	(0.012)	(0.012)	(0.025)
Monitor in class (D_{ij})	-0.609***	-0.416***	-0.919***	-0.538***	-0.365***	-0.816***
	(0.041)	(0.039)	(0.084)	(0.040)	(0.040)	(0.080)
N	140,010	87,498	52,512	140,010	87,498	52,512

Note. This table shows the effect of monitors on test scores. Class-level effects are 2SLS estimates using the presence of institutional monitors as an instrument. Robust standard errors, clustered on institution, are shown in parentheses. All regressions include year and grade fixed effects and sampling strata controls (grade enrollment at institution, region dummies and their interactions). * significant at 10%; ** significant at 5%; *** significant at 1%.

Table 4: Bounds on Scores from the Corrupt Sampling Model

	Math		Language	
	North/Centre	South	North/Centre	South
	(1)	(2)	(3)	(4)
	A. Monitored institutions			
Observed scores	-0.20	-0.04	-0.09	-0.11
Naive	[-2.73, -0.22]	[-4.57, -0.05]	[-3.73, -0.11]	[-5.08, -0.19]
Corrupt sampling	[-0.27, -0.22]	[-0.84, -0.05]	[-0.20, -0.11]	[-0.86, -0.19]
Corrupt sampling with verification	[-0.27, -0.22]	[-0.81, -0.05]	[-0.20, -0.11]	[-0.84, -0.19]
	B. Unmonitored institutions			
Observed scores	-0.10	0.28	-0.01	0.09
Naive	[-3.14, -0.11]	[-3.82, 0.29]	[-4.42, -0.02]	[-6.05, 0.08]
Corrupt sampling	[-0.26, -0.11]	[-1.24, 0.29]	[-0.25, -0.02]	[-1.52, 0.08]
Corrupt sampling with verification	[-0.26, -0.11]	[-1.22, 0.29]	[-0.24, -0.02]	[-1.51, 0.08]
	C. All institutions			
Observed scores	-0.12	0.21	-0.03	0.05
Naive	[-2.73, -0.22]	[-3.82, -0.05]	[-3.73, -0.11]	[-5.08, -0.19]
Corrupt sampling	[-0.26, -0.22]	[-0.84, -0.05]	[-0.20, -0.11]	[-0.86, -0.19]
Corrupt sampling with verification	[-0.26, -0.22]	[-0.81, -0.05]	[-0.20, -0.11]	[-0.84, -0.19]

Note. This table shows bounds obtained by imposing the restrictions discussed in Section 4. Panel A refers to monitored institutions. Panel B refers to unmonitored institutions. Panel C reports bounds constructed from the intersection of bounds in the first two panels. Naive bounds are defined from Assumption 4. Bounds from corrupt sampling are calculated under Assumptions 1-4, and follow from Horowitz and Manski (1995). Bounds from corrupt sampling with verification are calculated under Assumptions 1-4, and follow from Dominitz and Sherman (2006).

Table 5: Bounds on Scores using the Measurement Model and Monotonicity Restrictions

	Math		Language	
	North/Centre	South	North/Centre	South
	(1)	(2)	(3)	(4)
	A. Monitored institutions			
Observed scores	-0.20	-0.04	-0.09	-0.11
Non-differential misclassification	[-0.31, -0.23]	[-1.03, -0.10]	[-0.23, -0.12]	[-1.02, -0.22]
Manipulation decreasing with true scores	[-0.31, -0.26]	[-1.03, -0.28]	[-0.23, -0.15]	[-1.02, -0.38]
	B. Unmonitored institutions			
Observed scores	-0.10	0.28	-0.01	0.09
Non-differential misclassification	[-0.29, -0.11]	[-1.45, 0.35]	[-0.28, 0.00]	[-1.71, 0.14]
Manipulation decreasing with true scores	[-0.29, -0.18]	[-1.45, -0.07]	[-0.28, -0.07]	[-1.71, -0.22]
	C. All institutions			
Observed scores	-0.12	0.21	-0.03	0.05
Non-differential misclassification	[-0.29, -0.23]	[-1.03, -0.10]	[-0.23, -0.12]	[-1.02, -0.22]
Manipulation decreasing with true scores	[-0.29, -0.26]	[-1.03, -0.28]	[-0.23, -0.15]	[-1.02, -0.38]

Note. This table shows bounds obtained by imposing the restrictions discussed in Section 5. Panel A refers to monitored institutions. Panel B refers to unmonitored institutions. Panel C reports bounds constructed from the intersection of bounds in the first two panels. Bounds from non-differential misclassification are calculated under Assumptions 1-5. Bounds assuming manipulation decreasing with true scores are calculated under Assumptions 1-6.

Table 6: Summary of Main Results: Bounds on Scores and Manipulation Rates

	Math		Language	
	North/Centre	South	North/Centre	South
	(1)	(2)	(3)	(4)
	A. All institutions			
Observed scores	-0.12	0.21	-0.03	0.05
Scores disclosed by INVALSI	-0.15	-0.02	-0.06	-0.13
True scores	[-0.26, -0.26]	[-0.81, -0.28]	[-0.20, -0.15]	[-0.84, -0.38]
	B. Monitored institutions			
Presumed manipulators	0.01	0.09	0.01	0.07
True manipulators	[0.01, 0.01]	[0.05, 0.11]	[0.01, 0.02]	[0.04, 0.09]
Effect of manipulation on scores	[5.18, 7.65]	[4.24, 8.52]	[4.22, 7.80]	[4.33, 9.03]
	C. Unmonitored institutions			
Presumed manipulators	0.02	0.16	0.03	0.12
True manipulators	[0.02, 0.03]	[0.12, 0.26]	[0.02, 0.04]	[0.10, 0.20]
Effect of manipulation on scores	[3.41, 6.37]	[2.90, 6.64]	[2.95, 7.31]	[3.12, 8.95]

Note. This table shows bounds obtained by imposing Assumptions 1-6. Panel A presents bounds on scores, which are derived as intersection between bounds in the last row of Table 4 and Table 5. The table also reports scores calculated from raw data (observed scores) and scores adjusted using the procedure employed by INVALSI (scores disclosed by INVALSI). Panel B and Panel C refer to monitored and unmonitored institutions, respectively. The latter two panels report the percentage of true manipulators, as well as the effect of manipulation on scores.

Figure 1: Observed Scores, Adjusted Scores and Manipulation Rates for Math

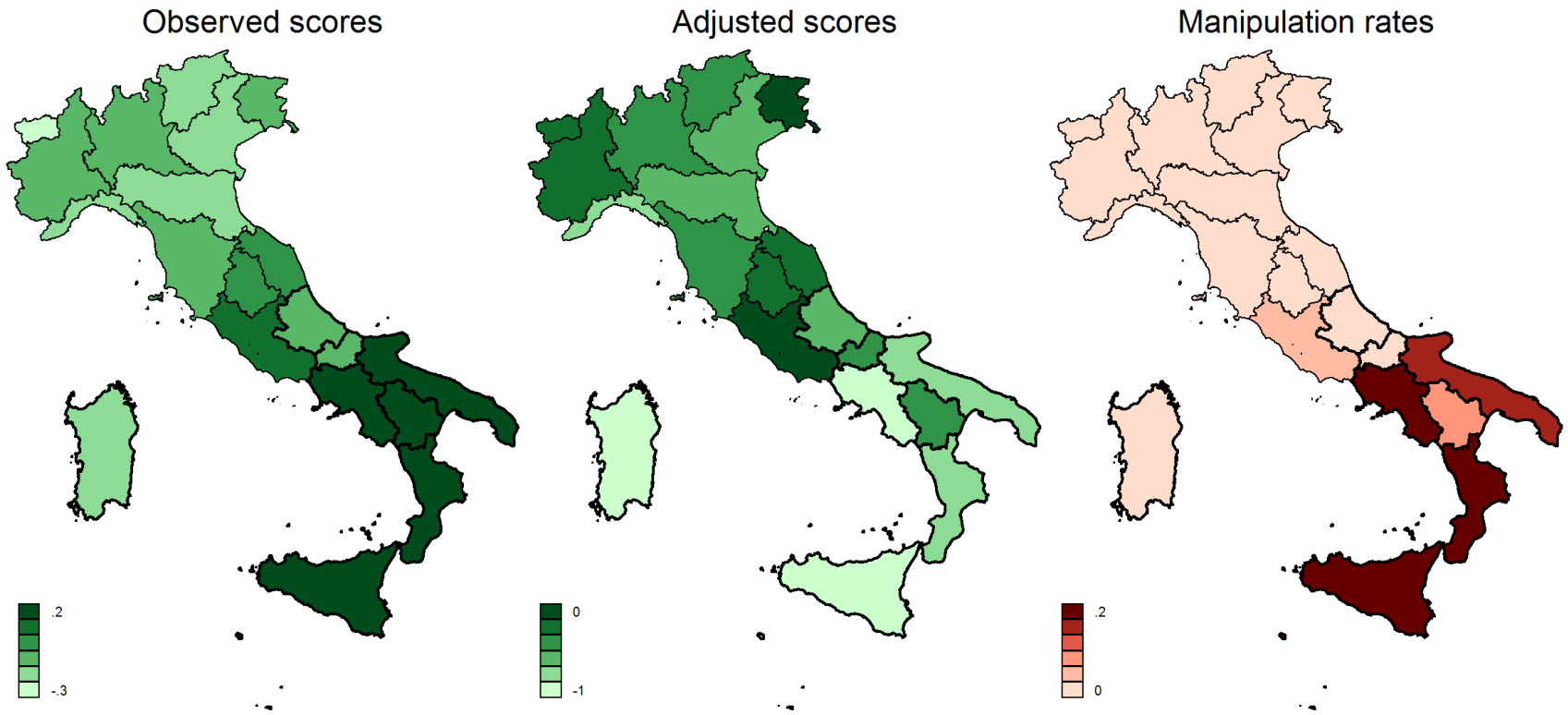


Figure 2: Observed Scores, Adjusted Scores and Manipulation Rates for Language

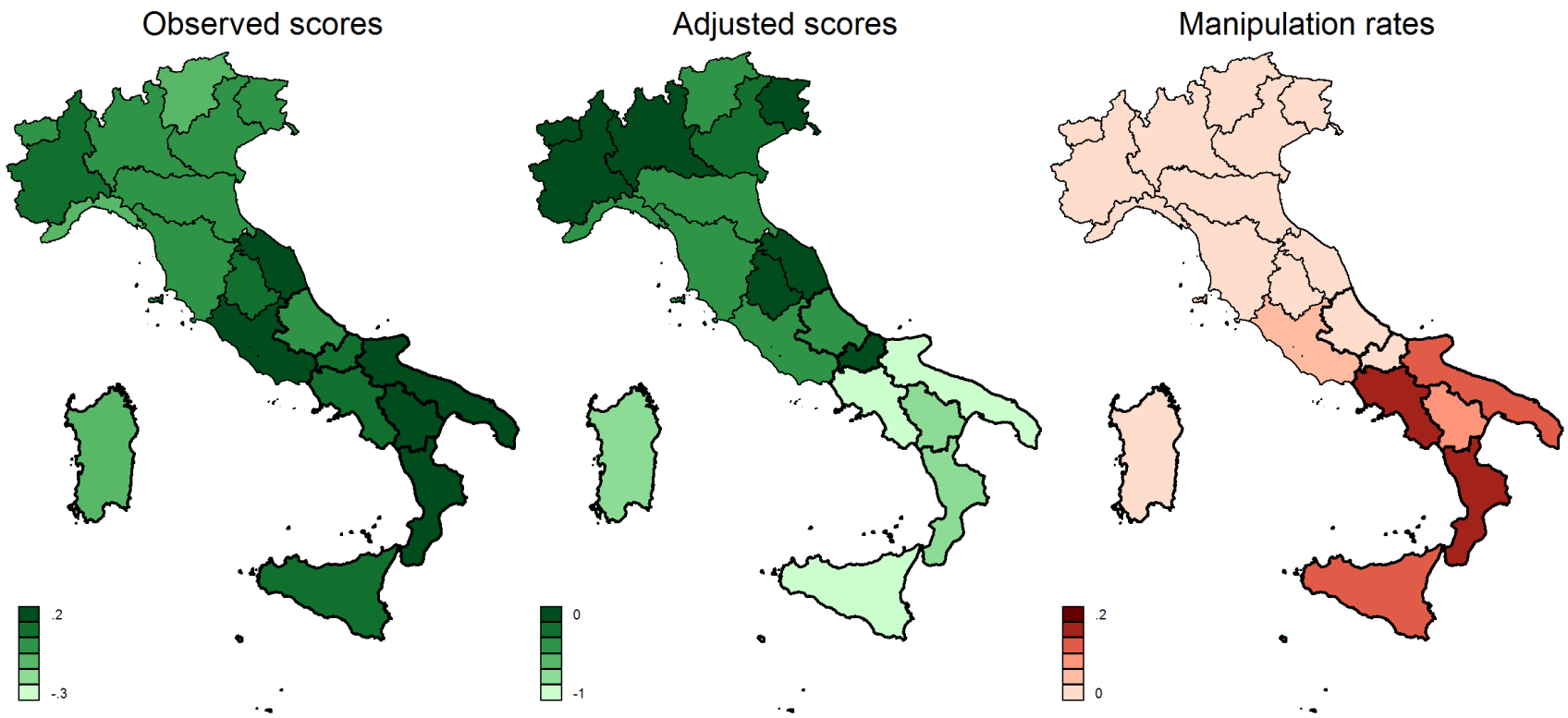
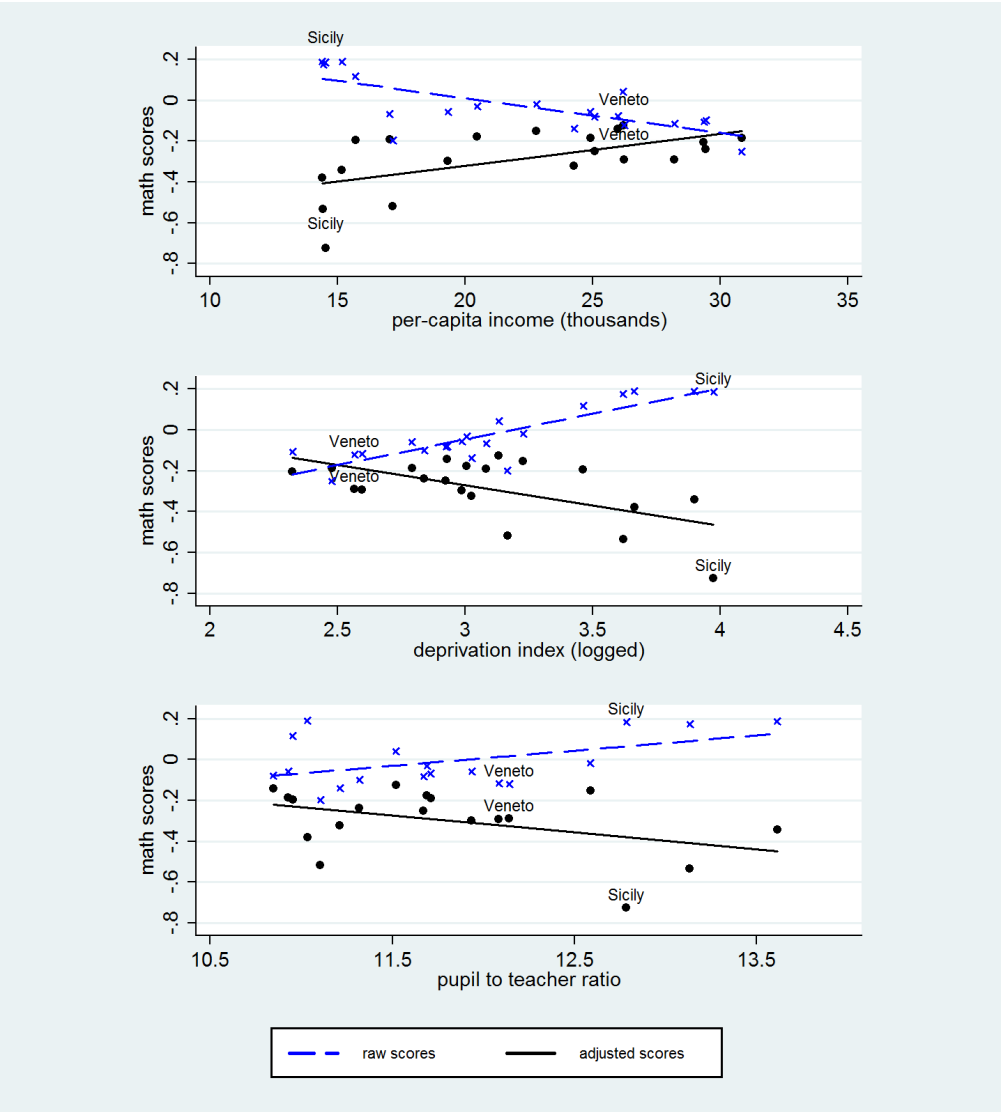


Figure 3: Observed and Adjusted Scores against School and Family Inputs



Note. The figure plots regional average math scores against per-capita income (top panel), an index of deprivation (central panel) and the pupil to teacher ratio (bottom panel). Points plotted with a "x" refer to observed scores, points plotted with a "•" refer to adjusted scores. Labeled in the figure are regions with the lowest (Veneto) and highest (Sicily) incidence of manipulation. Data on per-capita income are obtained from Istat, Conti economici regionali 2012. Data on the deprivation index are from Istat, Indagine sul reddito e condizioni di vita (Eu-Silc) 2012. Data on the pupil to teacher ratio are from Ministry of Education, La scuola statale - sintesi dei dati 2009-2010.

Appendix

Figure A1: Adjusted Scores (alternative method)

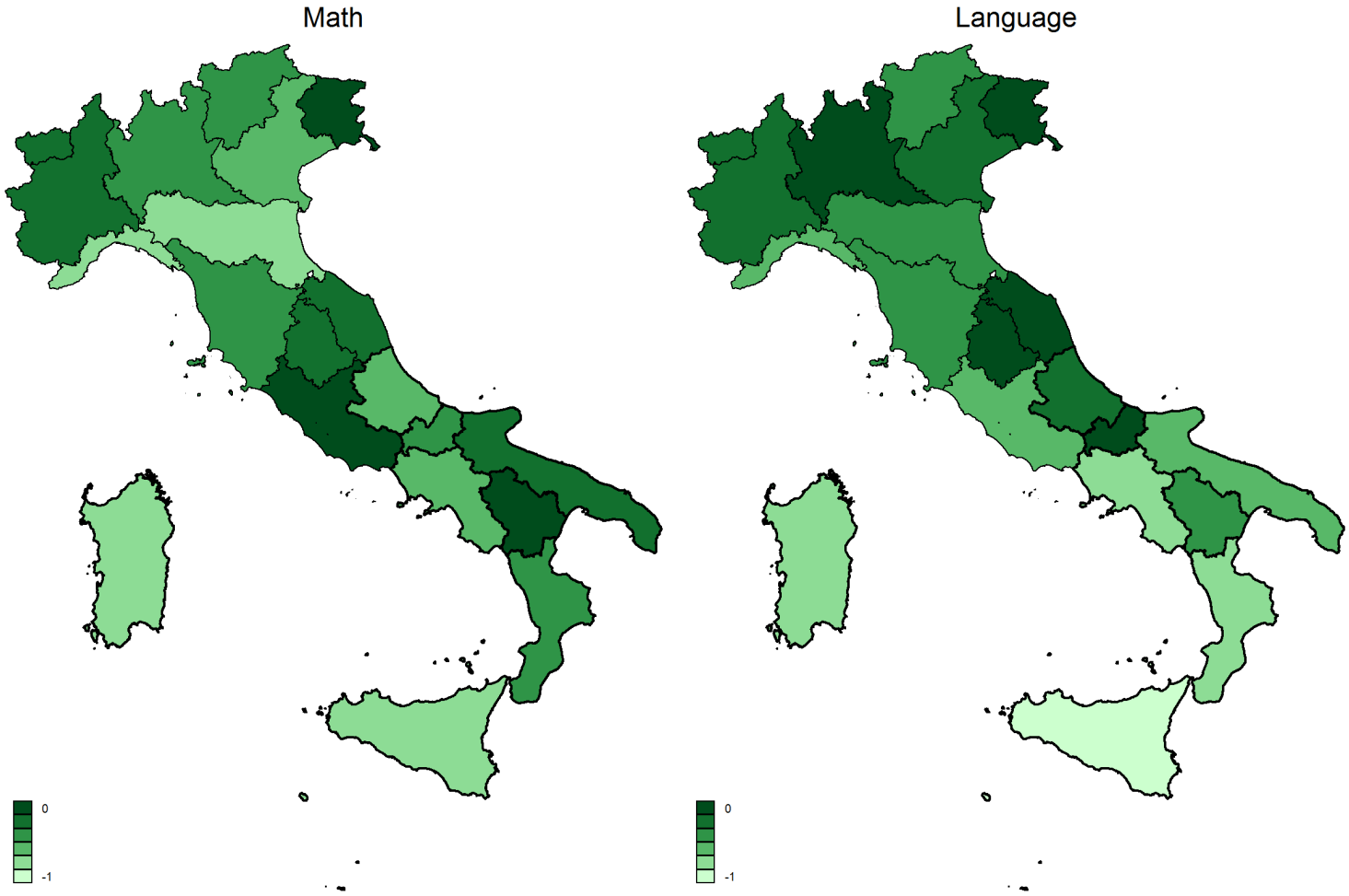


Figure A2: Adjusted Scores Using the INVALSI Methodology

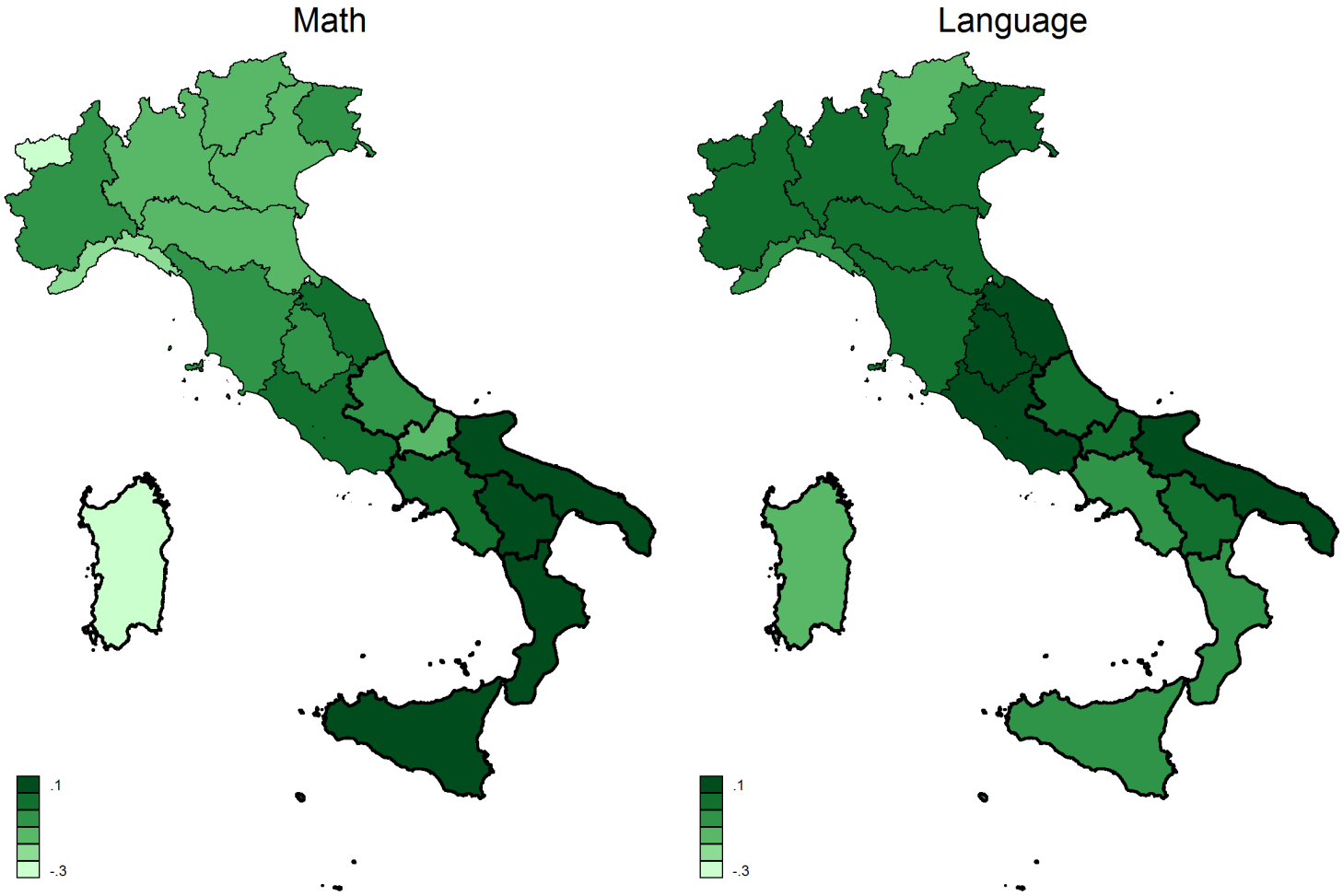


Figure A3: Scores Adjusted Using the INVALSI Methodology (alternative method)