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

Erich Battistin

Queen Mary University of London, CEPR, IRVAPP and IZA

Michele De Nadai

University of New South Wales†

Daniela Vuri

University of Rome Tor Vergata, IZA, CESifo and CEIS

May 2017

Abstract

We derive bounds on the distribution of math and language scores of elementary school students in Italy correcting for pervasive manipulation. A natural experiment that randomly assigns external monitors to schools is used to deal with endogeneity of manipulation, as well as its mismeasurement in the data. Bounds are obtained from properties of the statistical model used to detect classes with manipulated scores, and from restrictions on the relationship between manipulation and true scores. Our results show that regional rankings by academic performance are reversed once manipulation is taken into account.

JEL classification: C14; C31; C81; I21; J24.

Keywords: Measurement error; Non-parametric bounds; Partial identification; Score manipulation.

*We are indebted with the Editor and two anonymous referees for constructive comments on previous versions of this manuscript. 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 and Enrico Rettore for helpful discussions and comments and to seminar participants at the 2015 SOLE meeting, the 2015 Laax Labor Economics Workshop, the 2016 IAAE Conference, the 2016 Australasian Meeting of the Econometric Society, the University of Florence, the University of Rome Tor Vergata and the University of Maryland for helpful comments. This research was supported by the Fondazione Bruno Kessler. The views expressed here are those of the authors alone.

†*Corresponding author*: UNSW Business School UNSW Sydney, NSW 2052. Telephone: +61 2 9385 3367. E-mail: m.denadai@unsw.edu.au.

1 Introduction

Cross-national comparisons on student achievement are widely used to gauge the overall performances of a country's school system. Figures obtained from the *Trends in International Mathematics and Science Study* (TIMSS) and the *Progress in International Reading Literacy Study* (PIRLS) show Italian primary schools performing worst than in many European countries, especially in math. These same tests also show Southern Italy well behind the North, not surprisingly in view of the backwardness characterizing Southern regions along many economic dimensions (higher unemployment, lower per-capita income, higher crime rates) but also in terms of financial development (Guiso et al., 2004), political accountability (Nannicini et al., 2013) and workplace productivity (Ichino and Maggi, 2000). At the same time Italy's own accountability system, managed by the *Istituto Nazionale per la Valutazione del Sistema dell'Istruzione* (INVALSI), points to a very different regional pattern, with primary school students in the South doing better than Northerners. The marked regional gradient pictured by INVALSI data can be seen in the left hand side panels of Figure 1, where primary school math and language scores are considered. Moreover, the correlation of raw scores with proxies of school and family inputs unveils patterns in contrast with empirical regularities usually found in the literature. For example we show in Figure 2 that lower per-capita income is associated with higher scores in math, and that public spending is inversely related to achievement. The implications of these results to guide public policy in funding and accountability contradict the need for conspicuous EU investments to support modernization of education in Southern Italy through the Italian National Operative Programme (PON) scheme.[1]

How can these two sets of statistics be reconciled? A key difference between INVALSI tests and TIMSS and PIRLS tests emerges in their administration. INVALSI tests are proctored by local administrators and teachers, whereas in TIMSS and PIRLS scorers are organized into teams and a team leader ensures scoring reliability. We argue here that local administration opens the door to cheating and misreporting, and that it is this sort of manipulation that

[1] See http://www.invalsi.it/invalsi/index.php for a list of PON projects in Italy.

explains surprising patters in INVALSI data. Using a statistical model to detect manipulation, INVALSI identifies about 6% of classes in the country with compromised scores. In the South the proportion of compromised exams averages about 13%, uncovering the substantial regional gradient shown in the right hand side panels of Figure 1. For example, about 16% of classes in Sicily are suspected to have manipulated scores in math compared to less than 1% in Veneto, the Northern region with the most reliable figures according to INVALSI publications.

Angrist et al. (2017) discuss at large the origin of this phenomenon. They argue that local teachers manipulate results by dishonest transcription of students hand-written answer sheets onto machine-readable score report forms. INVALSI itself has acknowledged the problem, and now down-weights schools with suspiciously large results in the derivation of aggregate figures.[2] 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. (2016) 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. Concerns regarding score manipulation have also been raised in Sweden (Böhlmark and Lindahl, 2013 and Diamond and Persson, 2016) and in the United Kingdom, where Key Stage 1 tests at primary school are locally marked[3]. A recent system-wide cheating scandal in Atlanta has raised much interest from the media and several educators have been convicted (Severson 2011, Aviv 2014, Blinder, 2015).

The contamination of INVALSI data raises the problem of uncovering true patterns across Italian regions, which is the objective of this paper. Our analysis develops considering features of the true score distribution, its average being an example. Two main problems challenge identification. As classes with manipulated scores are arguably not representative of the population, selection precludes identification of the counterfactual score for manipulators. Moreover, the manipulation status from the statistical model employed by INVALSI can be misclassified as we do not have direct evidence on who manipulates.

We deal with these two problems using a policy that randomly assigns external monitors

[2]Their correction implicitly assumes that manipulated and honest scores are representative of the same population (INVALSI, 2013), a restriction that we do not impose here.

[3]See https://www.gov.uk/government/collections/national-curriculum-assessments-key-stage-1-tests

to 20% of institutions in the country.[4] Monitors supervise test administration and are responsible for score sheet transcription, as we discuss below. We use the presence of monitors at institution to instrument for manipulation, and show that this is sufficient to bound the distribution of true scores. Allowing for misclassified manipulation widens these bounds. However, if misclassification is independent of the sampling process that assigns monitors to classes and if monitoring prevents manipulation, monitored classes classified as manipulators reveal part of the error. We show how this result, coupled with standard assumptions on the misclassification error (Mahajan 2006, Lewbel 2007, and Hu 2008), yields bounds on the distribution of true scores that allow for endogenous and mismeasured manipulation.[5] Central to the development of our strategy are additional assumptions on the relationship between true scores and the incentives to manipulate. We show that a simple Roy model motivates restrictions on the true underlying achievement patterns and on the extent of misclassification.

The resulting bounds are sufficiently tight to reverse regional differences in raw scores: after correcting for manipulation, students in the North outperform students in the South in both math and language. Looking at a finer geographic disaggregation, we see that bounds in the most problematic regions of the South are dominated by scores in most regions of the North. For example Sicily - the region with the highest presumed incidence of manipulation - is ranked $3rd$ among the 20 Italian regions using raw math scores, and $15th$ at best after our correction. Our conclusions reconcile INVALSI data with evidence from international surveys, both in terms of the regional gradient in achievement and of its relationship with family and school inputs (as shown in Figure 2). We also show that score manipulation in the South is largely independent of the threat of having external monitors at institution, suggesting that dishonesty is widespread.

The remainder of the paper is organized as follows. Section 2 presents the institutional background and data, describes the monitoring experiment and explains the statistical procedure used by INVALSI to detect manipulation. Section 3 shows how to get identification of the parameters of interest when manipulation is measured without error. Section 4 derives the conditions for identification when manipulation is instead misclassified. Section 5 presents

[4]Institutions consist of affiliated schools, not at the same location.

[5]Identification under endogenous and misclassified 'treatment' is also considered in Nicoletti et al. (2011), Kreider et al. (2012), and Battistin et al. (2014).

conclusions on the extent of manipulation in INVALSI data that are robust to measurement error. In addition, we discuss how manipulation affects scores. Section 6 presents bounds on true achievement obtained by imposing restrictions on the behavior of manipulators. Section 7 derives policy implications and 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 academic years. Elementary school lasts 5 years starting from 6 years of age and covers grade 1 to 5. Standardized testing for evaluation purposes is mandatory 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 computed as percentage of correct answers, measured by grade and year of test administration. Our statistical unit of analysis is the class since our manipulation variable varies at class level, as explained below. The working sample includes only public schools (over 90% of primary school students) and consists of about 70,000 classes in each of the two grades covered by three years of data.

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 are selected from 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. The daily salary offered is about 200 euros per class monitored. Monitors supervise test administration and are responsible for score sheet transcription in a limited number of classes which are selected following a two-stage design. First, a sample of institutions strati-

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

fied by region is drawn with probability proportional to grade enrollment; then, in sampled institutions, one or two classes by grade (depending on grade 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 (as evident from descriptives in Bertoni et al. 2013 and discussed in Angrist et al. 2017).

In the absence of external monitoring, tests are proctored by local school staff. Proctors are expected to copy students' original responses onto machine-readable answer sheets (called *scheda risposta*), which are then sent to INVALSI. The transcription procedure is needed because this task is not 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 no further checks are enforced to ensure that students' original responses coincide with information on *scheda risposta* sent to INVALSI. Importantly, the transcription is performed outside official school hours without any monetary incentives for teachers.

Measuring Manipulation

The possibility of score manipulation is acknowledged by INVALSI in their official publications. We build on Angrist et al. (2017), who show that manipulation reflects teacher behavior. Specifically, it follows from dishonest transcription of students hand-written answer sheets onto machine-readable score report forms.

To identify classes with compromised scores, INVALSI adopts a procedure that takes as input within-class information on average and standard deviation of test scores, proportion of missing items, and variability in response patterns (as measured by a Gini index of homogeneity). The 4×4 correlation matrix determined by these indicators is used to extract two principal components, explaining over 90% of total variance for the years considered in our analysis. Cluster analysis is then used to form eight groups of classes from values of their principal components. Fuzzy clustering is adopted, yielding a matrix whose elements are, for each class, eight group membership probabilities. Classes with manipulated scores are identified as those in the group with "extreme" values of the principal components. In

practice, these are classes with abnormally high performance, small dispersion of scores, low proportion of missing items, and high concentration in response patterns relative to the population averages of these indicators. The indicator adopted by INVALSI is the probability of membership to the extreme group resulting from fuzzy clustering. This indicator is subject-specific (math and language), and clustering is stratified by grade and year.[7]

The manipulation indicator used in this paper is obtained replicating the same statistical procedure. However, hard clustering is used instead of fuzzy clustering: only classes in the extreme cluster identified by INVALSI are deemed to have compromised scores. The continuous indicator used by INVALSI is replaced here by a dummy variable aimed at outlier detection. The binary indicator eases interpretation, facilitates the discussion of measurement error in Section 4 and is in the spirit of Jacob and Levitt (2003). The right hand side panels of Figure 1 report the fraction of classes with compromised scores resulting from our indicator. The regional pattern depicted is qualitatively identical to that reported by INVALSI in official reports (e.g., INVALSI, 2010).

Descriptive statistics for the estimation sample are presented in Table A1.[8]

3 Identification When Manipulation is Observed

Let Y_1 and Y_0 be scores with and without manipulation, respectively. The observed score is $Y = Y_0 (1 - M) + Y_1 M$, where M is an indicator for manipulation (the 'treatment'). Class is the unit of analysis. The random variable Z takes value one if the class belongs to a monitored institution, while Q denotes monitored classes; hence $Q = 0$ if $Z = 0$ by design. Class and school level demographics X are also available, the conditioning on which is left implicit.

The monitoring experiment is used to learn about scores for different, latent types of

[7]For additional details, see Quintano et al. (2009). Classes suspected of manipulation are not sanctioned, although from school year 2011/12 INVALSI has used the manipulation indicator to adjust class scores. In classes with values of the indicator above a threshold set by INVALSI, results are not returned to the school. Below this threshold and within a range of values decided by INVALSI, scores are weighted by the value of the manipulation indicator. This procedure was unknown at the time of the test, making it unlikely that score manipulation anticipates the future adjustment.

[8]We will conventionally label as 'North' regions in Northern and Central Italy (Piedmont and Valle d'Aosta, Liguria, Trentino Alto Adige, Veneto, Friuli Venezia Giulia, Emilia Romagna, Tuscany, Umbria, Lazio). These will be contrasted to regions in the South (Abruzzo and Molise, Campania, Pulia, Basilicata, Calabria, Sicily and Sardinia).

teachers in the population. Types are defined by indexing manipulation to $Z = z$, M_z, to express the idea that monitoring should lower manipulation. Combinations of (M_0, M_1) define the four, mutually exclusive groups reported in Table 1. Honest (H) teachers are those who never manipulate. The remaining teachers are classified depending on their behavior in the presence of external monitoring. Complying (C) teachers are those who manipulate only when the threat is low (without monitors). Dishonest (D) teachers are those with positive expected benefit from manipulation even when the threat is high (with monitors). Finally, non-complying dishonest (N) teachers are those who would manipulate only with external monitors.

The incidence of the four groups in the population is ϕ_H, ϕ_C, ϕ_D and ϕ_N, respectively. The monitoring experiment reveals only one of the two potential outcomes (M_0, M_1). The notation for potential scores is also adjusted and indexed to the presence of monitors. The variable Y_{mz} represents class scores when $M = m$ and $Z = z$, where $m = 0, 1$ and $z = 0, 1$. The following assumption will be maintained throughout.

Assumption 1. *(Independence, monitoring effects, exclusion and monotonicity).*

(a) $(Y_{m1}, Y_{m0}, M_0, M_1) \perp Z$*; (b)* $E[M_1 - M_0] \neq 0$*; (c)* $Y_{m1} = Y_{m0}$ *for* $m = 0, 1$*; (d)* $\phi_N = 0$.

Assumption 1.a is implied in our setting by random assignment of monitors to institutions. Assumption 1.b states that institution monitoring is effective in reducing manipulation. Assumption 1.c is an exclusion restriction implying that monitoring lowers scores only by lowering manipulation.[9] Assumption 1.d rules out the presence of non-complying dishonest teachers, imposing that monitoring doesn't cause manipulation. Given these assumptions, we can identify four functionals of score distributions by adapting expressions in Abadie (2002):

$$E[g(Y_1)|D] = \frac{E[g(Y)M|Z = 1]}{E[M|Z = 1]}, \tag{1}$$

[9]Using survey data on exam day experiences and perceptions, Bertoni et al. (2013) find no direct effects of monitors on fifth graders' feelings and motivation. This rules out a possible effect of monitoring on scores over and above the effect on manipulation. Assumption 1.c is violated if the extent of manipulation depends on the presence of monitors at institution, for example because $Y_{11} < Y_{10}$. Assumptions would be needed to sign the role of unobservables that cause such violation, in the spirit of Nevo and Rosen (2012). Point identification of the quantities in (1), (2), (3) and (4) would be lost in this case.

$$E[g(Y_1)|C] = \frac{E[g(Y)M|Z=1] - E[g(Y)M|Z=0]}{E[M|Z=1] - E[M|Z=0]}, \tag{2}$$

$$E[g(Y_0)|C] = \frac{E[g(Y)(1-M)|Z=1] - E[g(Y)(1-M)|Z=0]}{E[(1-M)|Z=1] - E[(1-M)|Z=0]}, \tag{3}$$

$$E[g(Y_0)|H] = \frac{E[g(Y)(1-M)|Z=0]}{E[(1-M)|Z=0]}, \tag{4}$$

where $g(Y)$ is any real function of the observed variable Y such that the moments above are finite. In the empirical analysis the function $g(Y)$ will be assumed non-decreasing in its argument. Since:

$$E[M|Z=z] = \phi_D + (1-z)\phi_C,$$

we have $\phi_D = E[M|Z=1]$, $\phi_C = E[M|Z=0] - E[M|Z=1]$ and $\phi_H = 1 - \phi_D - \phi_C$. Figure 1 motivates the investigation of conditional versions of this parameter by area (e.g., North versus South).

The difference between equations (2) and (3) identifies the effect of manipulation for C teachers. We are interested, however, in the following quantity that includes H and D teachers as well:

$$E[g(Y_0)] = E[g(Y_0)|D]\phi_D + E[g(Y_0)|C]\phi_C + E[g(Y_0)|H]\phi_H. \tag{5}$$

We will use $g(Y_0) = Y_0$ when considering average scores, and $g(Y_0) = 1(Y_0 \geq \theta)$ to learn about classes scoring above a cutoff $\theta \in [0, 100]$. Even if M is not mismeasured, Assumption 1 is not sufficient for point identification as $E[g(Y_0)|D]$ is not identified in general. By varying the latter quantity over the space of its possible values, we obtain the identification region for $E[g(Y_0)]$. As we shall see in Section 4, lack of identification is exacerbated when the indicator M is mismeasured.

The identification region for the quantity of interest is narrowed by assuming that manipulation boosts scores, as seems likely.

Assumption 2. *(Scores and manipulation).* $0 \leq Y_0 \leq Y_1$.

This assumption implies $g(0) \leq E[g(Y_0)|D] \leq E[g(Y_1)|D]$, which narrows the width of the identification region for (5) to $(E[g(Y_1)|D] - g(0))\,\phi_D$. Importantly for what follows, $\phi_D = 0$ implies point identification. We will refer to naive bounds for (5) as those obtained under

Assumption 2.

Finally, we are interested in counting 'rotten apples'. The quantities $1 - \phi_H$ (i.e., the fraction of non-honest teachers, consisting of C and D groups) and ϕ_D (i.e., the fraction of D teachers) are considered to this end. These quantities are always point identified if manipulation is not mismeasured. The quantity ϕ_C is of independent interest, as it measures the size of teachers whose behavior is changed by the monitoring experiment. Using the instrumental variables jargon, this is the first stage effect when M is considered as 'treatment'.

Table A2 in the Appendix reports estimates of the quantities above obtained from raw data when $g(Y) = Y$. The counterfactual terms (1), (2), (3) and (4) are estimated from 2SLS regressions which follow from those described in Section 4 imposing no measurement error. The average effect of manipulation for C teachers, $E[Y_1 - Y_0|C]$, is positive and large - see columns (1) and (4). Classes with H teachers appear to have comparable scores everywhere in the country, well above those for classes with C teachers. Since $E[M|Z = 1] - E[M|Z = 0] = -\phi_C$, the table also shows that monitoring reduces manipulation, particularly in the South. Moreover, the size of the complying population among non-honest teachers is below 10% in the South. In the North ϕ_D is negligible and set to zero in estimation for practical purposes, implying that the presence of monitors at institution is sufficient to offset dishonest transcription of scores in all classes. We will soon come back to differences in the prevalence of latent types across areas.

4 Identification With Misclassified Manipulation

Monitoring and misclassification

In practice, we do not know for sure who manipulates. Rather, we observe only a proxy for M, denoted by W, which corresponds to the manipulation indicator described in Section 2.[10] We therefore turn to an analysis that takes misclassification into account. We first assume that class monitoring (that is, monitoring of classes within sampled institutions) prevents manipulation. Here and below the notation employed uses the fact that monitored classes

[10] As previously discussed, W is aimed at detecting outliers along multiple dimensions. However, this indicator is not deterministically related to large class scores. For example, 39% and 67% of classes with scores in the top 10% and 20% of the math distribution, respectively, have $W = 0$.

are only in monitored institutions.

Assumption 3. *(No manipulation in monitored classes).* $E[M_1|Q=1]=0$.

The potential outcome notation is employed here to stress that class monitoring constrains behavior of teachers. The assumption is weaker than assuming that monitoring eliminates manipulation in all classes at institution, and implies:

$$M = M(1-Q). \tag{6}$$

We further assume that the properties of the manipulation indicator do not change across populations for which it is computed. If one maintains $\phi_N = 0$, the probability of detecting manipulation must be the same for D and C teachers. Similarly, detection of honest reporting should be the same for H and C teachers. This is equivalent to assuming classification errors independent of the monitoring experiment.

Assumption 4. *(Misclassification independent of monitoring).* $E[W|M_1 = m, Z = 1] = E[W|M_0 = m, Z = 0]$ *for* $m = 0, 1$.

Assumption 4 is an exclusion restriction implying that the correlation between monitoring and W reflects only the correlation between monitoring and M. The relationship between observed manipulation, W, and latent manipulation, M, can therefore be written as:

$$W = (1-\pi_0) + (\pi_0 + \pi_1 - 1)M + \eta, \tag{7}$$

the terms π_1 and π_0 denoting probabilities of correct detection of manipulated and honest scores, respectively:[11]

$$\pi_m \equiv Pr[W = m|M = m], \qquad m = 0, 1.$$

Our characterization is completed by assuming non-differential misclassification, which is a standard assumption in the measurement error literature (see Carroll et al., 2006, and Chen et al., 2011). It qualifies W as a surrogate of M, in the sense that the latter variable is finer than the former in the relationship between manipulation, outcomes and Q.

Assumption 5. *(Non-differential misclassification). For* $z = 0, 1$ *and* $m = 0, 1$*:*

[11] Since by definition of (M_0, M_1) we have $E[W|M = m, Z = z] = E[W|M_z = m, Z = z]$, Assumption 4 is equivalent to $E[W|M = m, Z = 1] = E[W|M = m, Z = 0]$ for $m = 0, 1$.

$$(Y_0, Y_1, Q) \perp W | M_z = m, Z = z.$$

This again is an exclusion restriction, implying that W does not have any residual correlation with (Y_0, Y_1, Q) when the manipulation status M is revealed. Importantly, it does not rule out the likely correlation between manipulation, Y and Q, and can be stated conditional on X. Also, Assumption 5 lends itself to an interpretation that stems from the definition of latent types. It says that errors in detection of misconduct (of D and C teachers) or of honest behavior (of H and C teachers) are independent of outcomes and class monitoring.

We can now derive the quantities (1)-(4), which are functionals of (Y, Q, M, Z), as functionals of (Y, Q, W, Z). This allows us to establish an identifying map between quantities of interest and their analogues computed from raw data. Define:

$$\Lambda \equiv [W - (1 - \pi_0)] (1 - Q), \qquad \Psi(\pi_1) \equiv (\pi_0 + \pi_1 - 1) - \Lambda.$$

The dependence on π_0 is left implicit here, as this parameter will not be relevant for the derivation of bounds in what follows. The main results are presented in propositions, whose proof follows from calculations reported in the Appendix.

Proposition 1. *(Potential distributions under misclassification). Under Assumptions 1, 3, 4 and 5:*

$$E[g(Y_1)|D] = \frac{E[g(Y)\Lambda|Z = 1]}{E[\Lambda|Z = 1]}, \tag{8}$$

$$E[g(Y_1)|C] = \frac{E[g(Y)\Lambda|Z = 1] - E[g(Y)\Lambda|Z = 0]}{E[\Lambda|Z = 1] - E[\Lambda|Z = 0]}, \tag{9}$$

$$E[g(Y_0)|C] = \frac{E[g(Y)\Psi(\pi_1)|Z = 1] - E[g(Y)\Psi(\pi_1)|Z = 0]}{E[\Psi(\pi_1)|Z = 1] - E[\Psi(\pi_1)|Z = 0]}, \tag{10}$$

$$E[g(Y_0)|H] = \frac{E[g(Y)\Psi(\pi_1)|Z = 0]}{E[\Psi(\pi_1)|Z = 0]}. \tag{11}$$

The quantities above are partially identified by letting the probabilities of correct classifications, π_0 and π_1, vary over their support. Importantly, quantities involving outcomes under manipulation depend only on π_0. The equations presented are derived imposing the constrain in (6). When $Q = 0$ for all classes, Proposition 1 provides expressions for function-

als of potential outcomes in an instrumental variables setting when the treatment status is misclassified.

Proposition 2. *(Rotten apples under misclassification). Under Assumptions 1, 3, 4 and 5:*

$$1 - \phi_H = \frac{E[\Lambda|Z=0]}{(\pi_0 + \pi_1 - 1)}, \qquad \phi_D = \frac{E[\Lambda|Z=1]}{(\pi_0 + \pi_1 - 1)}.$$

Note that the ratio of these two quantities depends only on π_0:

$$\frac{\phi_D}{1-\phi_H} = \frac{E[\Lambda|Z=1]}{E[\Lambda|Z=0]}, \tag{12}$$

and represents the share of D teachers among the pool of non-honest (D or C) teachers. It also follows that:

$$\phi_C = \frac{E[\Lambda|Z=1] - E[\Lambda|Z=0]}{(\pi_0 + \pi_1 - 1)}.$$

Proposition 3. *(Identification of π_0). Under Assumptions 3, 4 and 5 the misclassification probability π_0 is identified from the monitoring experiment:*

$$(1 - \pi_0) = E[W|Q=1].$$

The last result is central for what follows, implying that the quantities (8), (9) and (12) are point identified from the data. It follows that identification region for the parameter in (5) becomes wider when $W \neq M$ because π_1 is unknown. At the same time, point identification of ϕ_D, ϕ_C and ϕ_H is lost.[12]

The difference between (9) and (10) identifies the causal effect of manipulation for C teachers. Using the expression for $(1 - \pi_0)$ in Proposition 3, it follows that:

$$E[g(Y_1) - g(Y_0)|C] = (\pi_1 - E[W|Q=1]) \frac{E[g(Y)|Z=1] - E[g(Y)|Z=0]}{E[W|Z=1] - E[W|Z=0]},$$

the last term on the right hand side representing the causal effect obtained from raw data (Table A2 in the Appendix reports the effect when $g(Y) = Y$). A sufficient condition to ensure that misclassification preserves the sign of this causal effect is $\pi_1 > E[W|Q=1]$. The

[12] Define:

$$S_C(a) \equiv \{E[g(Y_0)|C] : \pi_1 \geq a\}, \qquad S_H(a) \equiv \{E[g(Y_0)|H] : \pi_1 \geq a\},$$

as the sets of values taken by (10) and (11), respectively, when $\pi_1 \geq a$. Calculations available on request show that, for $\delta > 0$ and $a + \delta < 1$, we have $S_C(a+\delta) \subset S_C(a)$ and $S_H(a+\delta) \subset S_H(a)$. This implies that the identification regions for (10) and (11) shrink as π_1 increases to one.

latter assumption also implies that manipulation biases upward causal effects for complying teachers estimated from raw data (Aigner, 1973).[13]

Finally the following assumption is also imposed. It is convenient in our setting, and is not rejected in the data when combined with Assumption 8 below. This is more than the minimum requirement to maintain positive correlation between M and W, and imposes a lower bound on the correlation between W and M.

Assumption 6. *(Informational content of W).* $\pi_1 \geq 0.5$.

Estimation

The quantities in Proposition 1 are estimated from a sequence of 2SLS regressions using Z as instrument at selected values of the probability π_1 (for an application of the same method see Angrist et al. 2013). For C teachers, the following equations are considered:

$$g(Y)\Lambda = \alpha_1^C + \beta_1^C \Lambda + \zeta_1^C,$$

$$g(Y)\Psi(\pi_1) = \alpha_0^C + \beta_0^C \Psi(\pi_1) + \zeta_0^C, \tag{13}$$

and the coefficients β_1^C and β_0^C are used to estimate (9) and (10), respectively. For D teachers, the quantity (8) is obtained by considering the coefficient β_1^D from:

$$g(Y)\Lambda Z = \alpha_1^D + \beta_1^D \Lambda Z + \zeta_1^D. \tag{14}$$

Finally, the coefficient β_0^N from:

$$g(Y)\Psi(\pi_1)(1-Z) = \alpha_0^N + \beta_0^N \Psi(\pi_1)(1-Z) + \zeta_0^N,$$

is used to estimate the quantity (11). We estimate separate regressions by area (Northern versus Southern regions), controlling for grade and year effects and using sampling probability weights constructed from the stratification variables in the monitoring experiment (region, grade enrollment at institution and their interactions).[14] Standard errors are obtained using

[13] Under Assumption 6 below, Assumption 2 doesn't restrict the range of values of the probability π_1.

[14] All control variables here are categorical. To ease the computational burden, and since the identification result abstracts from parametric restrictions, we impose that covariates enter linearly the various conditional expectations. We checked the sensitivity of our conclusions to this restriction by implementing the estimator proposed by Frölich (2007), which uses non-parametric estimates over cells defined by the cross-tabulation of

100 bootstrap replications clustering on institution. We will use these as point-wise standard errors (i.e., standard errors computed for a fixed value of π_1) in presenting some of the results. The quantities in Proposition 2 are estimated from their sample analogues, using sampling probability weights, and their standard errors computed via bootstrap.[15]

Finally, the value $(1-\pi_0)$ is estimated by taking the empirical analogue of $E[W|Q=1]$ using all monitored classes in the sample. In particular, we impose the same value of $(1-\pi_0)$ across areas, as the regression of W on X for monitored classes ($Q=1$) did not yield important differences over time, grades and areas. The resulting estimate of π_0 is 98%, implying that 2% of classes are erroneously classified as manipulators.

5 Counting Rotten Apples

Using raw INVALSI data we compute the fraction of honest (ϕ_H), complying (ϕ_C) and dishonest teachers (ϕ_D) from Proposition 2. Results are presented in Figure 3 by varying π_1 over the interval $[0.5, 1]$, separately for Northern and Southern regions. Shaded areas represent 95% confidence intervals obtained at each value of π_1. The left hand side panels show that the fraction of honest teachers is almost 100% in the North, and that it uniformly dominates, at all values of π_1, the fraction of honest teachers in the South. Manipulation appears to be more pronounced for math.

The fraction of C teachers is reported on the right hand side panels of Figure 3. The incidence of D teachers can be mechanically obtained as a residual term $\phi_D = 1 - \phi_H - \phi_C$, and is not presented. The striking feature about manipulation in the North is that all dishonest teachers are compliers, implying $\phi_D \simeq 0$. The important policy conclusion to draw is that institution monitoring in the North annihilates manipulation. Complying teachers in the South are roughly half the pool of non-honest (C or D) teachers. Depending on the value of π_1, the size of the non-honest group varies between 11% and 23% for language and between 15% and 30% for math. All D teachers are located in the South, suggesting that manipulation is an area-specific phenomenon. The veil of ignorance about the probability

X and Z. In our data, the average sample size across 240 cells is 583 classes. Results from this alternative estimation strategy are reported in Figures B1-B4 of the on-line Appendix, and are qualitatively similar to those presented below.

[15] See Arcones and Giné (1992) and Hahn (1996) for the validity of bootstrap for just identified 2SLS models with an i.i.d. sample from (Y, W, Z).

π_1 limits our ability to count rotten apples. However, our conclusions on peculiarities of the two areas are not precluded by this limitation.

The evidence documented poses the question of how score distributions are affected by such pervasive manipulation. The issue is addressed in Figure 4, where densities of Y_1 for D and C teachers are reported. In light of the information conveyed by Figure 3, only results for the South are presented. For a grid of values θ in the support of the outcome variable, we estimate equations (13) and (14) using $g(Y) = 1(Y \geq \theta)$. The resulting estimates of β_0^C and β_0^D represent cumulative distributions for C and D teachers, respectively, at all values θ. These are combined using the isotonic regression smoother (Brunk, 1958) to impose non-decreasing curves, which we then use to plot densities in Figure 4. Shaded areas represent 95% confidence intervals constructed from bootstrap standard errors using 100 replications.[16]

Teachers manipulate scores to obtain almost perfect results (i.e., 100% correct). This is consistent with manipulation boosting scores on all items regardless of their difficulty. Wholesale curbstoning, a strategy that minimizes transcription or grading effort while maintaining high levels of achievement, has been identified as the primary force behind score manipulation in Angrist et al. (2017). The distributions presented for math and language are substantially identical across latent types.

6 Bounds on True Scores

Naive bounds

Bounds on $E[g(Y_0)]$ are obtained by varying π_1 over the range implied by Assumption 6. For all admissible values of this probability, we compute (5) by retrieving the relevant quantities from Proposition 1 and Proposition 2. Imposing Assumption 2, the counterfactual term $E[g(Y_0)|D]$ is bounded from above by (8), yielding the following bounds for a known value

[16] Density estimation may be carried out using:

$$g(Y) = \frac{1}{h} K(\frac{a - Y}{h}),$$

where the term on the right hand side is a kernel function with bandwidth h (see Angrist et al. 2013 for a similar idea). The resulting estimates of β_0^C and β_0^D represents densities at a for C and D teachers, respectively. The approach taken in this section ensures some symmetry with density estimation at the end of Section 6, and yields qualitatively identical results.

of π_1:

$$E[g(Y_0)] \leq E[g(Y_1)|D]\phi_D + E[g(Y_0)|C]\phi_C + E[g(Y_0)|H]\phi_H,$$

$$E[g(Y_0)] \geq g(0)\phi_D + E[g(Y_0)|C]\phi_C + E[g(Y_0)|H]\phi_H,$$

where $g(0)$ is the value of $g(Y_0)$ at the lowest possible value of the score ($Y_0 = 0$).[17] Consistently with the evidence presented in the previous section, we impose $\phi_D = 0$ in the North implying that in this area $E[g(Y_0)]$ is point identified at all values of π_1. Shaded areas, here and in what follows, represent point-wise 95% confidence intervals (i.e., confidence intervals computed for a known value of π_1) constructed using the bootstrap procedure by Horowitz and Manski (2000) with 100 replications.

Bounds for average math and language scores are presented in Figure 5, separately for the two areas, using $g(Y_0) = Y_0$. The discussion of results for other functionals of the score distribution is deferred to the final part of this section. Math scores in the North are centered at about 61% of correct answers, a value included in the corresponding bounds computed for the South. Confidence intervals in the South shrink as π_1 grows to one, ranging from $[51, 64]$ for $\pi_1 = 0.5$ to $[56, 63]$ for $\pi_1 = 1$. Results are not conclusive about the ranking of areas with respect to performance in math. However, the bottom panel of Figure 5 tells a different story for language scores. Scores in the North virtually bound from above the range of admissible values for scores in the South. The average difference between areas computed from raw data is reversed once manipulation is taken into account. Confidence intervals for average scores in the South shrink from $[60, 71]$ when $\pi_1 = 0.5$ to $[65, 71]$ when $\pi_1 = 1$.

Behavioral restrictions

Restrictions on the origin of manipulation can be used to tighten naive bounds. The following assumption is reasonable for the case at hand, implying that dishonest teachers have the lowest scores. 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 as we discuss further below.

[17] This choice is rather conservative. The lowest scores are 22% and 30% for math and language, respectively, in monitored institutions in Sicily, which is the Italian region with the highest presumed manipulation rate.

Assumption 7. *(Origin of manipulation). The following inequality holds for $z = 0, 1$:*

$$E[g(Y_0)|M_z = 0] \geq E[g(Y_0)|M_z = 1].$$

In potential outcome notation, the assumption is equivalent to stating that D teachers have the worst scores:

$$E[g(Y_0)|\bar{D}] \geq E[g(Y_0)|D], \tag{15}$$

and that H teachers have better scores than do D and C teachers:

$$E[g(Y_0)|H] \geq E[g(Y_0)|\bar{H}], \tag{16}$$

where the notation $\bar{D}$ and $\bar{H}$ denotes non-D and non-H teachers, respectively. As we show in the Appendix, these assumptions are expected to tighten the naive upper bound in our application. However, they do not necessarily make the naive lower bound more informative. To ease presentation, we state the bounds implied by Assumption 7 in the following proposition.

Proposition 4. *(Bounds under behavioral restrictions). Under Assumptions 1, 3, 4, 5 and 7, the following bounds are defined for a known value of π_1:*

$$E[g(Y_0)] \leq \min\left\{E[g(Y_0)|C]\frac{\phi_C}{1-\phi_D} + E[g(Y_0)|H]\frac{\phi_H}{1-\phi_D}, E[g(Y_0)|H]\right\},$$

$$E[g(Y_0)] \geq g(0)\frac{\phi_D}{1-\phi_H} + E[g(Y_0)|C]\frac{\phi_C}{1-\phi_H}.$$

Figure 6 shows how naive bounds for average scores are refined by imposing Assumption 7. Results are obtained by taking the intersection with bounds in the previous section. We find that (15) and (16) unveil geographic differences in math scores and reinforce the ranking in language scores already pictured by naive bounds. Math scores in the North now bound from above admissible values for scores in the South. Confidence intervals in the South change from $[51, 58]$ for $\pi_1 = 0.5$ to $[56, 61]$ for $\pi_1 = 1$. Assumption 7 implies a sizable improvement for the width of bounds compared to the naive case discussed above: about 46% at $\pi_1 = 0.5$ and 38% at $\pi_1 = 1$. Confidence intervals for language scores in the South are now $[60, 68]$ when $\pi_1 = 0.5$, and $[65, 70]$ when $\pi_1 = 1$ (with a 17% and 27% width improvement, respectively), strengthening geographic differences already evident with naive bounds.

A Roy model for manipulation

Assumption 7 yields a partial ordering of $E[g(Y_0)]$ across latent groups, and can be motivated using a Roy model for the decision to manipulate. Write potential outcomes as $Y_1 = \mu_1 + \varepsilon_1$ and $Y_0 = \mu_0 + \varepsilon_0$, and let $\gamma_1 Z + \tau$ be the manipulation cost. The latter varies across classes and areas through the random term τ, and increases for everyone in the presence of external monitors ($\gamma_1 > 0$). By letting $V \equiv -\varepsilon_1 + \varepsilon_0 + \tau$ and $\gamma_0 \equiv \mu_1 - \mu_0$, the decision to manipulate is modeled as:

$$M = 1(Y_1 - Y_0 - \gamma_1 Z - \tau > 0) = 1\left(\gamma_0 - \gamma_1 Z \geq V\right), \tag{17}$$

where V is unobservable and assumed continuous with a strictly increasing distribution function. In this setting, manipulation occurs if the expected benefit is positive. The assumption that Z is independent of the triple $(\varepsilon_0, \varepsilon_1, V)$, which in our setting follows from the monitoring experiment, together with the latent index equation (17) imply, and are implied, by Assumption 1 (Vytlacil, 2002). Additive separability between Z and V plays an essential role in this result. It follows that latent groups in the population are identified from the value of V:

$$D:\ \gamma_0 - \gamma_1 \geq V, \qquad C:\ \gamma_0 \geq V \geq \gamma_0 - \gamma_1, \qquad H:\gamma_0 \leq V.$$

This representation implies that scores for D, C and H teachers are stochastically ordered if $E[\varepsilon_0|V = v]$ is not decreasing in v.[18] It therefore follows that Assumption 7 is implied by non-decreasing class performance in manipulation cost.

The same Roy model also implies the full ordering of $E[g(Y_0)]$ across latent groups for any non-decreasing function $g(Y_0)$. In general this is neither implied by, or implies, Assumption 7. This paves the way for the following additional assumption used in the derivation of bounds.

Assumption 8. *(Ranking of scores across latent types). The following inequality holds:*

$$E[g(Y_0)|H] \geq E[g(Y_0)|C] \geq E[g(Y_0)|D].$$

This requirement refines the identification region of the parameter of interest by changing the upper bound and by restricting the parameter space for π_1. To see this, notice that the first inequality in Assumption 8 implies:

[18] Joint normality of (ε_0, V) with positive correlation of the two components is also sufficient for the result.

$$E[g(Y_0)|H] - E[g(Y_0)|C] \geq 0, \tag{18}$$

for all values of π_1. Given that both $E\left[g\left(Y_0|H\right)\right]$ and $E\left[g\left(Y_0|C\right)\right]$ are identified from equations (10) and (11) up to knowledge of π_1, we could use equation 18 to rule out values of π_1 yielding a negative difference. Conditional on values of π_1 for which (18) is not violated, the following proposition presents the lower and upper bounds from imposing that scores are ranked across latent types of teachers. We show in the Appendix that the new upper bound is more informative than the upper bound derived in Proposition 4.

Proposition 5. *(Bounds under ranking of scores). Under Assumptions 1, 3, 4, 5 and 8, the following bounds are defined for a known value of π_1 when $E[g(Y_0)|H] \geq E[g(Y_0)|C]$:*

$$E[g(Y_0)] \leq E[g(Y_0)|C](1-\phi_H) + E[g(Y_0)|H]\phi_H, \tag{19}$$

$$E[g(Y_0)] \geq g(0)\phi_D + E[g(Y_0)|C](1-\phi_D). \tag{20}$$

Bounds for average math and language scores obtained by imposing Assumption 7 and Assumption 8 are shown in Figure 7. We set $\pi_1 \geq 0.56$ and $\pi_1 \geq 0.55$ for math and language, respectively, as these are the critical values ensuring the validity of (18).[19] Classes in the North outperform classes in the South in both math and language. For math, scores in the South range from $[52, 59]$ when π_1 is at its minimum to $[56, 60]$ for $\pi_1 = 1$ (implying a 20% improvement on the width of bounds in Proposition 4 when $\pi_1 = 1$). Similar results are obtained for language, the width of confidence interval at $\pi_1 = 1$ now being $[65, 68]$, with an improvement of about 40%.

The discussion so far has addressed the problem of bounding average scores. The same methodology can be used to provide bounds for distributions, by replicating the same analysis using $g(Y_0) = 1(Y_0 \geq \theta)$. Bounds (19) and (20) define mixtures bracketing the true, unknown distribution of Y_0 for a known value of π_1. For a grid of values θ in the support of the outcome variable, we compute the complement to the cumulative distributions for complying teachers, $E[1(Y_0 \geq \theta)|C]$, and honest teachers, $E[1(Y_0 \geq \theta)|H]$. The isotonic regression smoother is then applied as in Figure 4, and bounds are computed by combining smoothed distributions

[19] Figure B5 in the on-line Appendix presents the profile of this difference with respect to π_1, together with confidence intervals obtained by bootstrapping the difference between (10) and (11) which we estimate from 2SLS regressions as explained above.

using mixture weights ϕ_H and ϕ_D. Results are presented in Figure 8 for scenarios corresponding to 'small' ($\pi_1 = 0.95$), 'moderate' ($\pi_1 = 0.90$) and 'large' ($\pi_1 = 0.80$) misclassification of true manipulators.

Results show that the ranking of average scores does not follow from stochastic dominance across areas. The lower tail of the math distribution is thicker in the South, implying a larger number of classes with problematic performance compared to the North. However, the fraction of classes scoring above 60% (approximately the average score for Northerners, as shown in Figure 7) is comparable across areas, and the upper tail in the South is significantly thicker than in the North for scores between 70% and 90%. A similar comment applies to language scores: the fraction of classes scoring above 70%, the average score for Northerners, is comparable across areas, but the lower tail of the distribution is much ticker in the South. This conclusion holds up at all values of π_1 considered, implying that poor performance in the South is driven by a large number of classes lagging behind academic standards of the best classes in the area, which are instead comparable to the North.

Regional rankings

The same analysis can be carried out at a finer geographic level considering the 20 Italian administrative regions. This disaggregation is important as reflects political divisions responsible for the administration of local resources, including those assigned to schools. We start by deriving, for each region, bounds on the incidence of manipulation $E[M]$. These are obtained from (7) by varying π_1 over its support. Pragmatically, we impose $E[M] = 0$ in those regions (6 for math and 5 for language, all in the North) where the upper bound is below 1%. It follows that in these regions raw scores can be treated as true scores ($Y = Y_0$). Maintaining the assumption that all manipulators in the North are compliers, we compute bounds on regional scores as in Figure 7. Our correction heavily affects the national ranking because the effects of manipulation on scores are large, as it is shown in Figure 9. Here bounds are presented for $\pi_1 = 0.9$, as the general conclusions are not sensitive to this choice (see Figures B6 and B7 in the on-line Appendix which report results for $\pi_1 = 0.80$ and $\pi_1 = 0.95$, respectively). Dots in the figure represent average scores computed from raw data. Continuous lines are confidence intervals for bounds on true scores obtained from our correction. The vertical axis reports names of all regions, which are ranked clockwise from

North to South (Lazio is the last Northern region considered in the main analysis). Upper bounds in the South are dominated by scores for most regions in the North. For example, Sicily is ranked $3rd$ according to raw math scores, and at best $15th$ after our correction. Moreover, the figure allows to establish at least a partial ordering of Southern regions.

Not only our correction changes the regional gradient in scores, but also affects its relationship with family and school inputs. Figure 2 presents the association between scores and per-capita income (left panel) and pupil-to-teacher ratio (right panel) across the 20 administrative regions of the peninsula. Dots in the figure are average scores obtained from raw data, which are interpolated to obtain the downward-sloping line for per-capita income and the upward-sloping line for pupil-to-teacher ratio, respectively. Superimposed are labels for the regions with lowest (Veneto) and highest (Sicily) incidence of presumed manipulation according to the indicator W. Results show that regions with low per-capita income and high pupil-to-teacher ratio have the highest scores, a fact hard to reconcile with evidence from the international literature. The figure also presents linear fits once manipulation is taken into account. Adjusted scores, represented with crosses, coincide with regional upper bounds in Figure 9, thus considering the most conservative scenario for the relationship with the socio-economic indicators considered.

We find that score manipulation reverses the sign of the correlation. This finding has important implications for empirical analyses using INVALSI data, as the effects of manipulation when true scores, Y_0, are used as dependent variable in the relationship with inputs, X, is not innocuous as it is in the case of classical measurement error. Although our discussion was not centered around differences between $\frac{\partial}{\partial x}E[Y_0|X=x]$ and $\frac{\partial}{\partial x}E[Y|X=x]$, the Appendix presents a simple setup showing under which conditions the latter term can be wrongly signed because of score manipulation.

7 Conclusions and policy implications

Our findings have important policy implications. The first result is that manipulation is widespread only in some areas of the country (see Figure 3). Scores in the North are reported correctly for at least 98% of classes. The fraction of classes with compromised math scores in the South is instead at least 15%, but can be almost 30% depending on the assump-

tions made on the extent of misclassification. Manipulation reflects dishonesty of teachers (Angrist et al. 2017) which, in the South, is a widespread attitude rather than an opportunity to cut corners: approximately half of dishonest teachers manipulate scores regardless of the threat of having an external monitor at institution. Italian teachers work in a highly regulated public sector, with virtually no risk of termination, and are subject to a pay and promotion structure that are largely independent of performance. As the available resources are inadequate to increase the number of monitored classes or to provide sensible monetary incentives to teachers, INVALSI should improve reliability of the information collected either by sanctioning dishonest behavior or by enforcing high quality standards in the transcription process. From 2013/14 multiple versions of the same test are employed with items randomly ordered, making the mechanical transcription of correct answers into *scheda risposta* more difficult. Whether or not this measure has been successful in limiting manipulation is still an open issue.

Measuring manipulation and purging data from its distortive effects is of primary interest for the redistribution of resources and the design of education policies. All indicators of score manipulation are error prone, and the procedure followed by INVALSI is no exception. We have shown that the INVALSI monitoring experiment can be used to unveil part of the error. If the properties of the manipulation indicator are independent of how monitors are assigned to classes, the fraction of monitored classes classified as honest must equal π_0. This probability in our data is approximately 98%, with little variability across grades, areas and time. As INVALSI flags as suspicious classes with a distribution of answers unusually concentrated around high scores, 2% of truly exceptional classes may be erroneously deemed to have manipulated results. The possibility of misclassification should be acknowledged in the publication of official reports.

Our approach offers an alternative to the correction used by INVALSI until 2013 (Falzetti, 2013). Their method employed the class-level probability of manipulation derived from fuzzy clustering (as described in Section 2). Average figures in the country were obtained by down-weighting classes with abnormally high values of the indicator with respect to Veneto, a region in the North where scores are viewed as the most accurate. Classes with a probability value below the median for Veneto were given weight one; all remaining classes were weighted one minus this probability. This adjustment affected marginally the regional gra-

dient obtained from raw data, as can be seen from official reports published by INVALSI.[20] Starting from 2013, this procedure was refined by INVALSI using weights constructed with a different methodology. Both corrections are not uncontroversial, as they implicitly assume that manipulators are a random sample from the population (as well as that they can be detected for sure). Our approach overcomes such limitations.

Learning about the incidence of score manipulation allows to rank Italian regions in terms of performance at national tests (see Figure 9). If the propensity to manipulate decreases with true scores, an assumption consistent with the implications of a simple Roy model, bounds on true scores are tight enough to reverse the evidence from raw data. Classes in the South underperform with respect to the rest of Italy, and differences are particularly pronounced in the most problematic regions. This conclusion aligns well with that from international surveys like TIMMS and PIRLS. Interestingly, a closer look at score distributions reveals higher inequality in the South and thick tails at the lower end. Besides, the best classes in the South have scores comparable to the best classes in the North. It follows that differences in score distributions between areas do not result from a location shift, and poor average achievement in the South can be ascribed to a disproportionally large number of low performing classes. These are the learning environments that should be primary target of policy interventions in Southern Italy, for example through the National Operative Programme (PON) scheme.

Why is the fact that score manipulation distorts regional rankings 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. Manipulation explains the puzzling, negative relationship between scores and family and school inputs that researchers would measure from raw data, as we have shown in Figure 2. The association between achievement and inputs is reversed by the correction, as better endowed regions are now characterized by higher scores. Ignoring manipulation, at least for the case of primary schools considered here, would heavily bias results of empirical analyses using micro data on scores. This finding has important implications for public policy in funding and accountability.

Our findings raise a number of additional questions, including why teacher manipulation

[20] We document this in Figure B8 of the on-line Appendix. The correlation between regional ranks before and after the correction is 99% and 66% for math and language scores, respectively. A variant to this procedure is also considered by INVALSI, and assigns weight zero to classes with a probability value above 50%. The resulting regional ranking is comparable to that in Figure B8 (results are available upon request).

is so much more prevalent in the South, and what can be done to enhance accurate assessment in Italy and elsewhere. Similar concerns have been raised in regard to the consequences of local proctoring and grading of tests in Britain and New York. For example, local teachers mark 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 monitoring and external marking (grading).[21] It's also worth asking what are the determinants of low performance of students in the South of Italy, 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.

[21]See documents and links at http://www.education.gov.uk/sta/assessment, and the evidence of manipulation in Battistin and Neri (2015).

References

Abadie, A. (2002): "Bootstrap Tests for Distributional Treatment Effects in Instrumental Variable Models," *Review of Economic Studies*, 97, 284–292.

Aigner, D. (1973): "Regression with a binary independent variable subject to errors of observation," *Journal of Econometrics*, 1, 49–60.

Angrist, J. D., E. Battistin, and D. Vuri (2017): "In a Small Moment: Class Size and Moral Hazard in the Italian Mezzogiorno," *American Economic Journal: Applied Economics*, forthcoming.

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

Arcones, M., and E. Giné (1992): "On the bootstrap of M-estimators and other statistical functionals," in *Exploring the Limits of Bootstrap*, ed. by R. LePage, and L. Billard. Wiley, New York.

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., M. De Nadai, and B. Sianesi (2014): "Misreported Schooling, Multiple Measures and Returns to Educational Qualifications," *Journal of Econometrics*, 181(2), 136–150.

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.

Battistin, E., and L. Neri (2015): "Manipulation of Internally and Externally Assessed Evaluations of Students: Evidence from the UK," Queen Mary University of London, Unpublished mimeo.

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

Blinder, A. (2015): "Atlanta Educators Convicted in School Cheating Scandal," *New York Times, April 1*, Accessed at: https://www.nytimes.com/2015/04/02/us/verdict-reached-in-atlanta-school-testing-trial.html.

Böhlmark, A., and M. Lindahl (2013): "Independent Schools and Long-Run Educational Outcomes - Evidence from Sweden's Large Scale Voucher Reform," forthcoming, *Economica*.

Brunk, H. D. (1958): "On the Estimation of Parameters Restricted by Inequalities," *The Annals of Mathematical Statistics*, 29(2), 437–454.

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.

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

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

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.

Frölich, M. (2007): "Nonparametric IV Estimation of Local Average Treatment Effects with Covariates," *Journal of Econometrics*, 139(1), 35–75.

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

Hahn, J. (1996): "A Note on Bootstrapping Generalized Method of Moments Estimators," *Econometric Theory*, 12, 187–197.

Horowitz, J. L., and C. F. Manski (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.

INVALSI (2010): "Sistema Nazionale di Valutazione - A.S. 2009/2010, 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.

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

Kreider, B., J. V. Pepper, C. Gundersen, and D. Jolliffe (2012): "Identifying the Effects of SNAP (Food Stamps) on Child Health Outcomes When Participation Is Endogenous and Misreported," *Journal of the American Statistical Association*, 107(499), 958–975.

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.

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.

Nicoletti, C., F. Peracchi, and F. Foliano (2011): "Estimating Income Poverty in the Presence of Missing Data and Measurement Error," *Journal of Business and Economic Statistics*, 29(1), 61–72.

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.

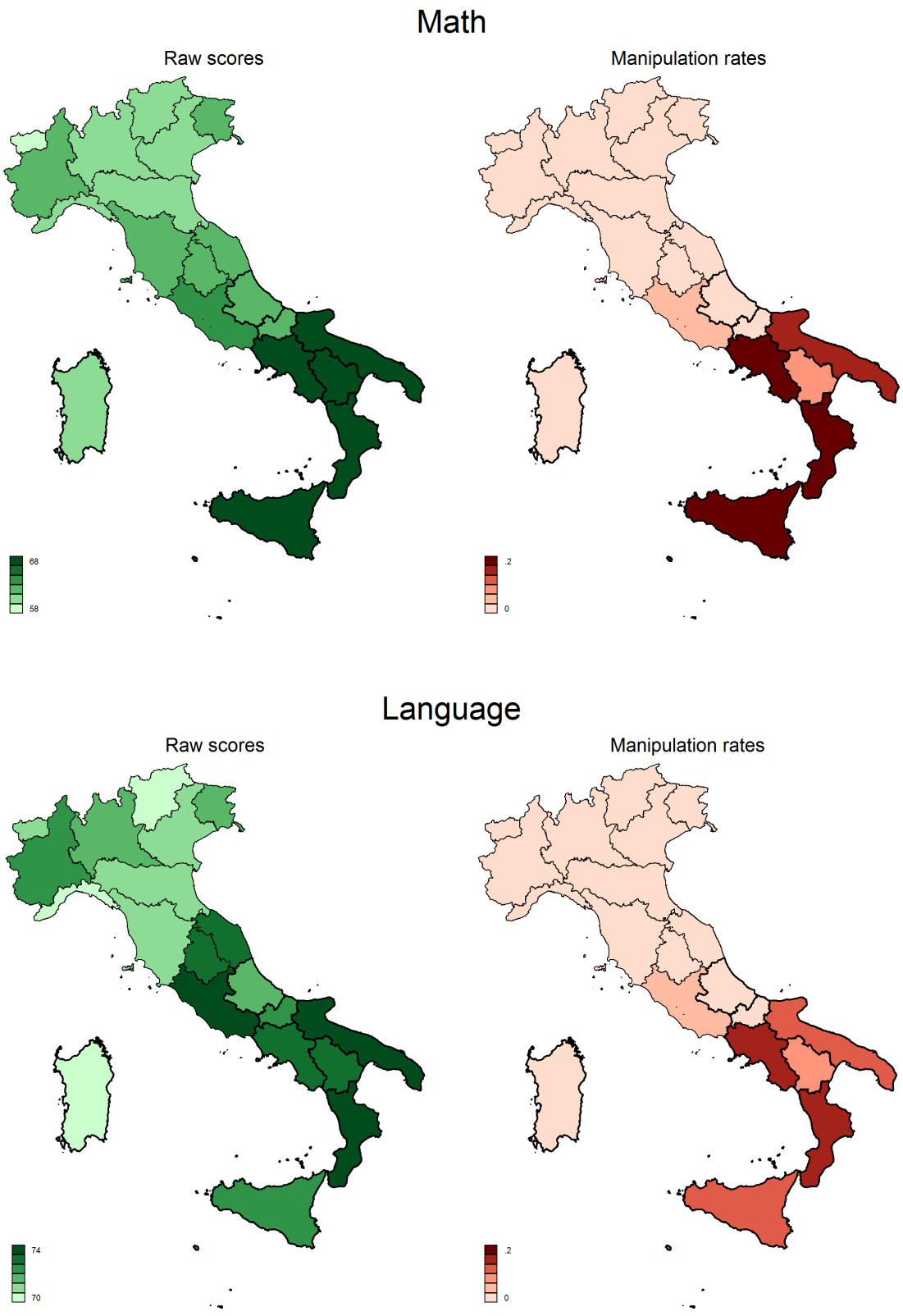
Vytlacil, E. J. (2002): "Independence, Monotonicity, and Latent Index Models: An Equivalence Result," *Econometrica*, 70(1), 331–341.

Table 1: Compliance Types

	$M_1 = 1$	$M_1 = 0$
$M_0 = 1$	Dishonest (D)	Complying (C)
$M_0 = 0$	Non-Complying Dishonest (N)	Honest (H)

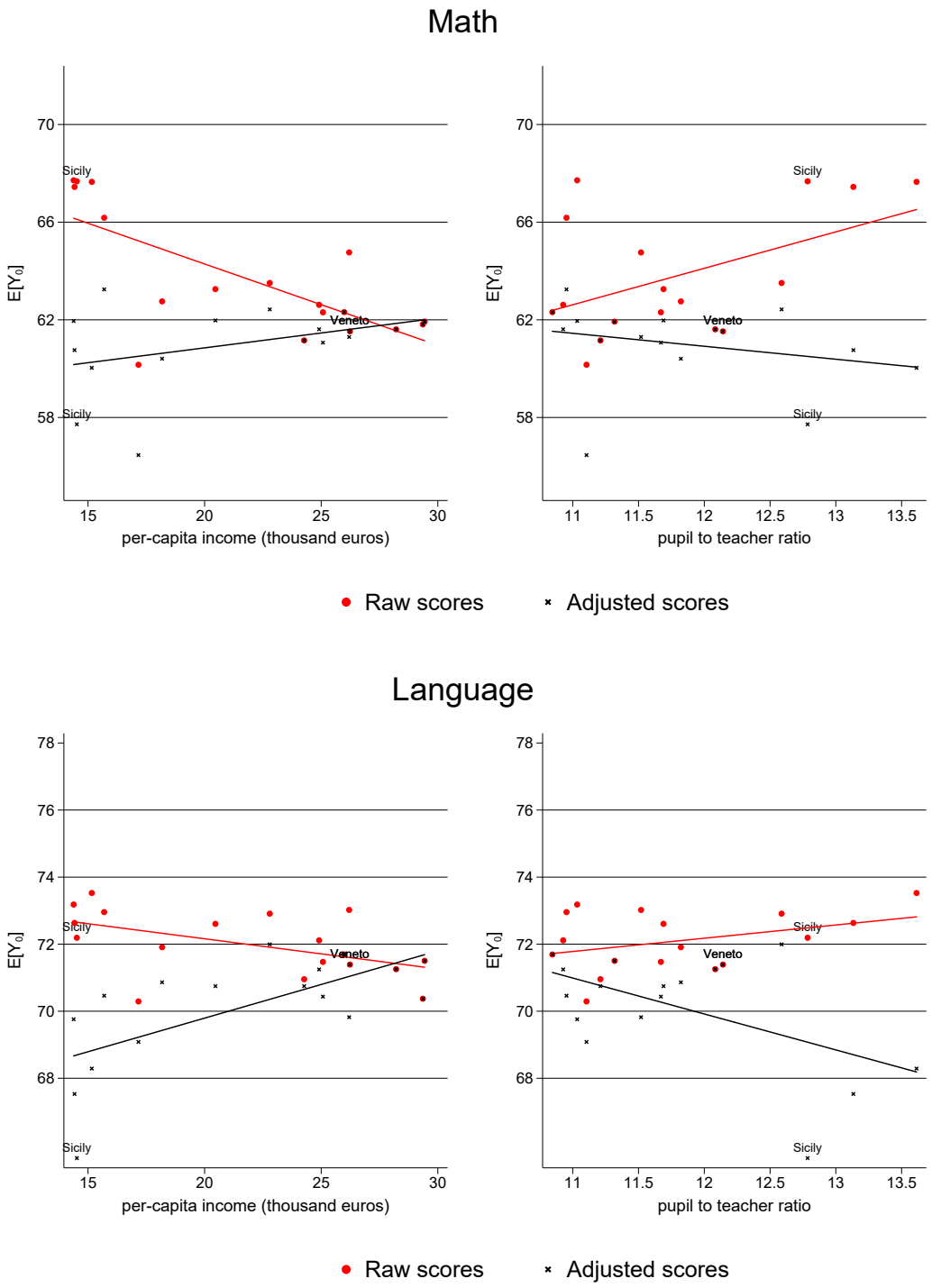
Note. This table defines the four compliance types implied by the monitoring experiment. Types refer to teacher behavior. See Section 3 for details.

Figure 1: Raw Scores and Manipulation Rates



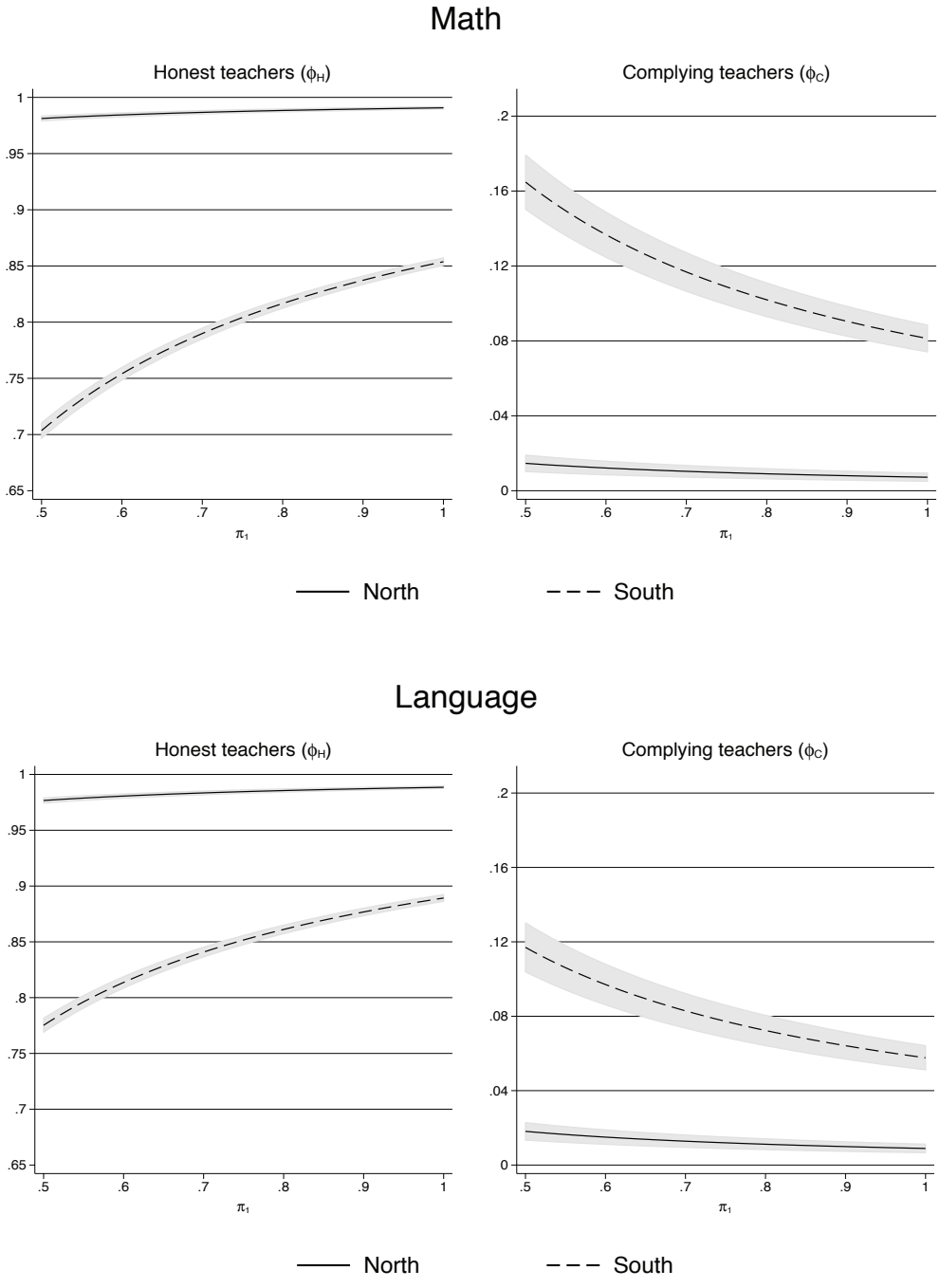
Note. These figures are obtained from INVALSI data pooling second and fifth grade students for the school years 2009-2011.

Figure 2: Raw and Adjusted Scores against School and Family Inputs



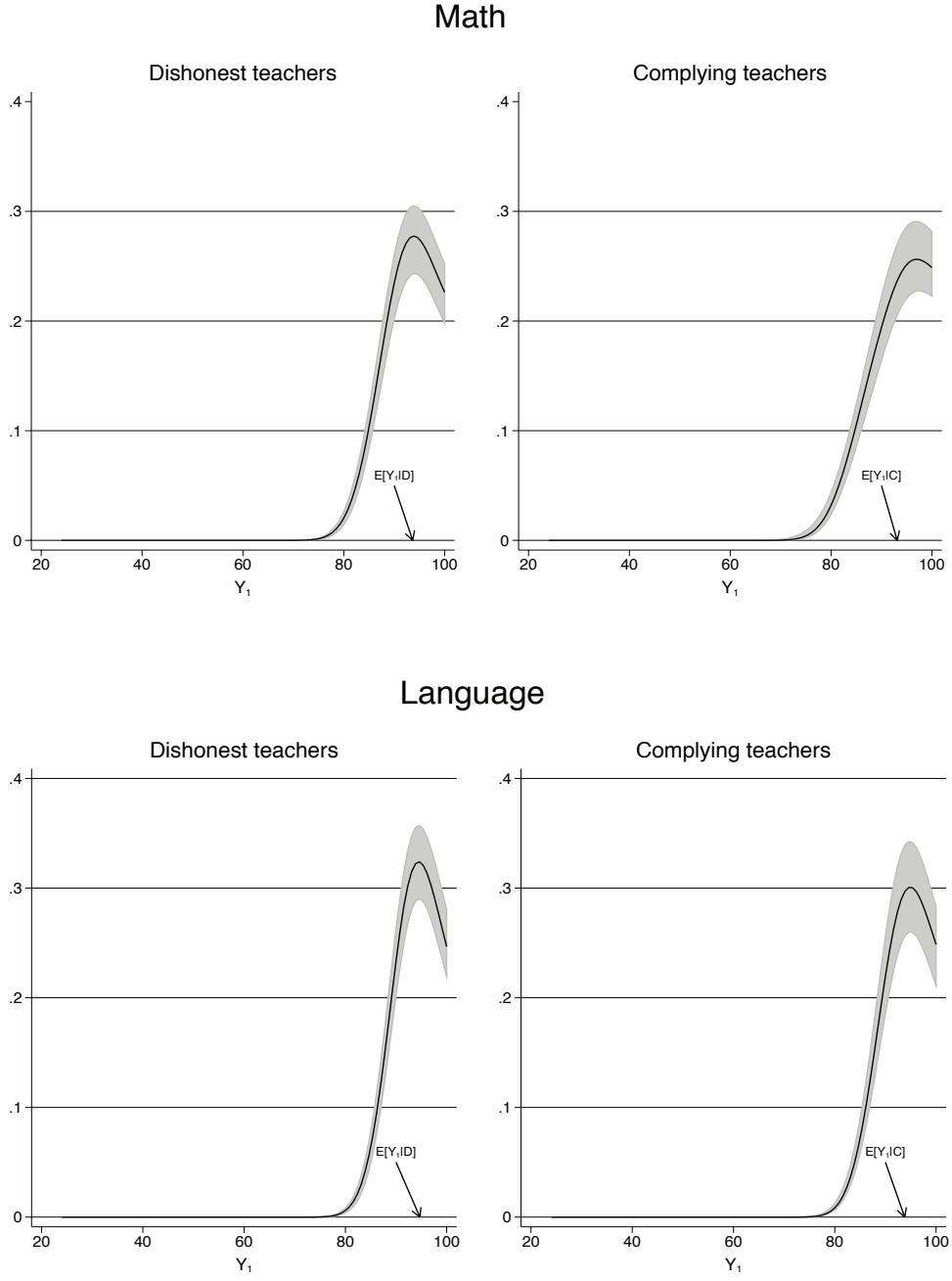
Note. The figure plots regional math and language scores against regional per-capita income (left panel) and pupil to teacher ratio (right panel). Points with • refer to raw scores and points with x refer to adjusted scores. Labeled in the figure are regions with the lowest (Veneto) and highest (Sicily) incidence of presumed manipulation. Data on per-capita income are obtained from *Istat, Conti economici regionali 2012*. Data on the pupil to teacher ratio are from the Ministry of Education, *La scuola statale - sintesi dei dati 2009-2010*.

Figure 3: Percentages of Honest and Complying Teachers



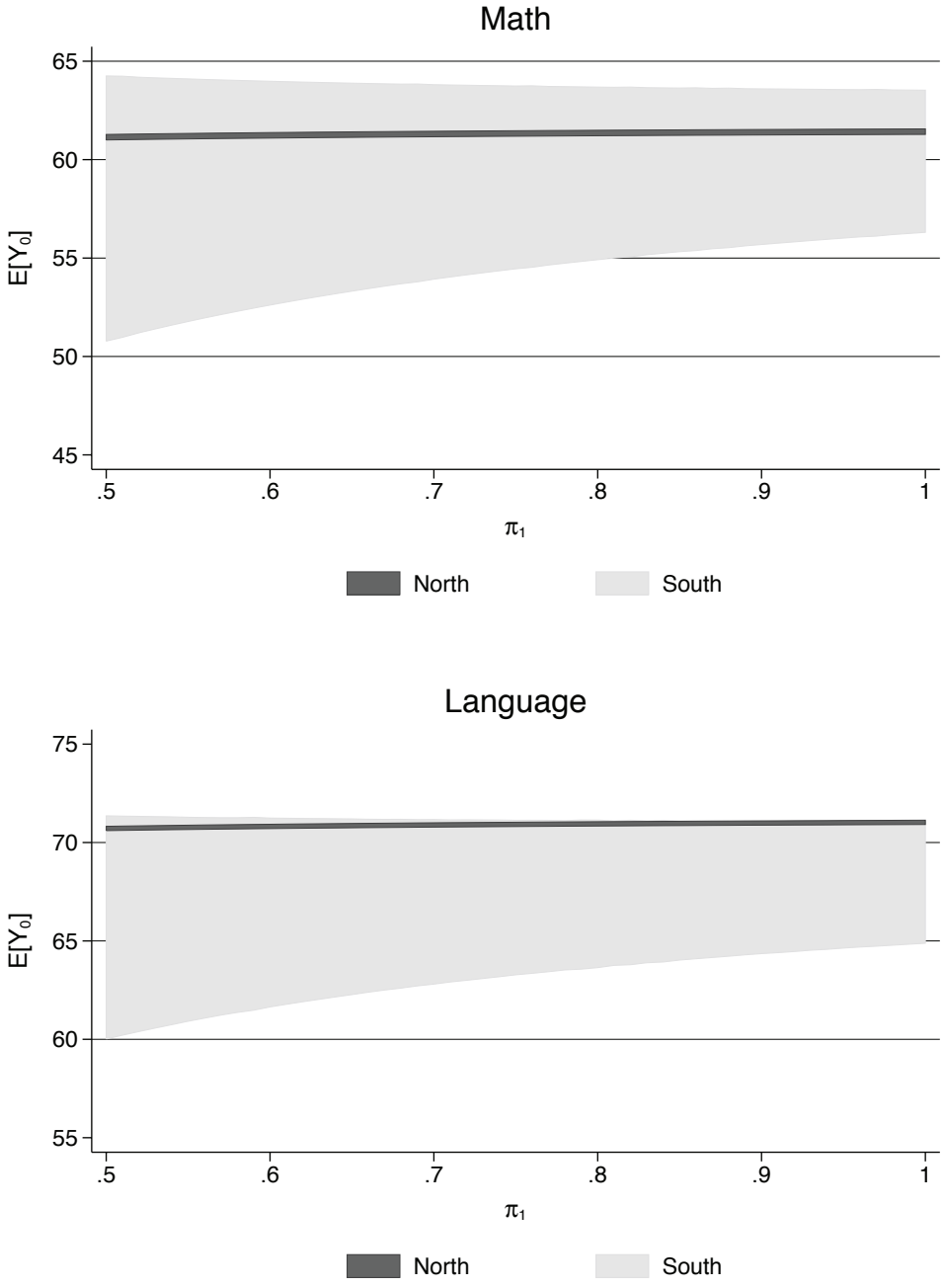
Note. This figure reports the percentage of honest (ϕ_H) and complying (ϕ_C) teachers. Results are presented for the interval $\pi_1 \geq 0.5$, separately for the North (continuous line) and the South (dashed line). Shaded areas are 95% bootstrap confidence intervals obtained at each value of π_1.

Figure 4: Manipulated Score Distributions by Compliance Type



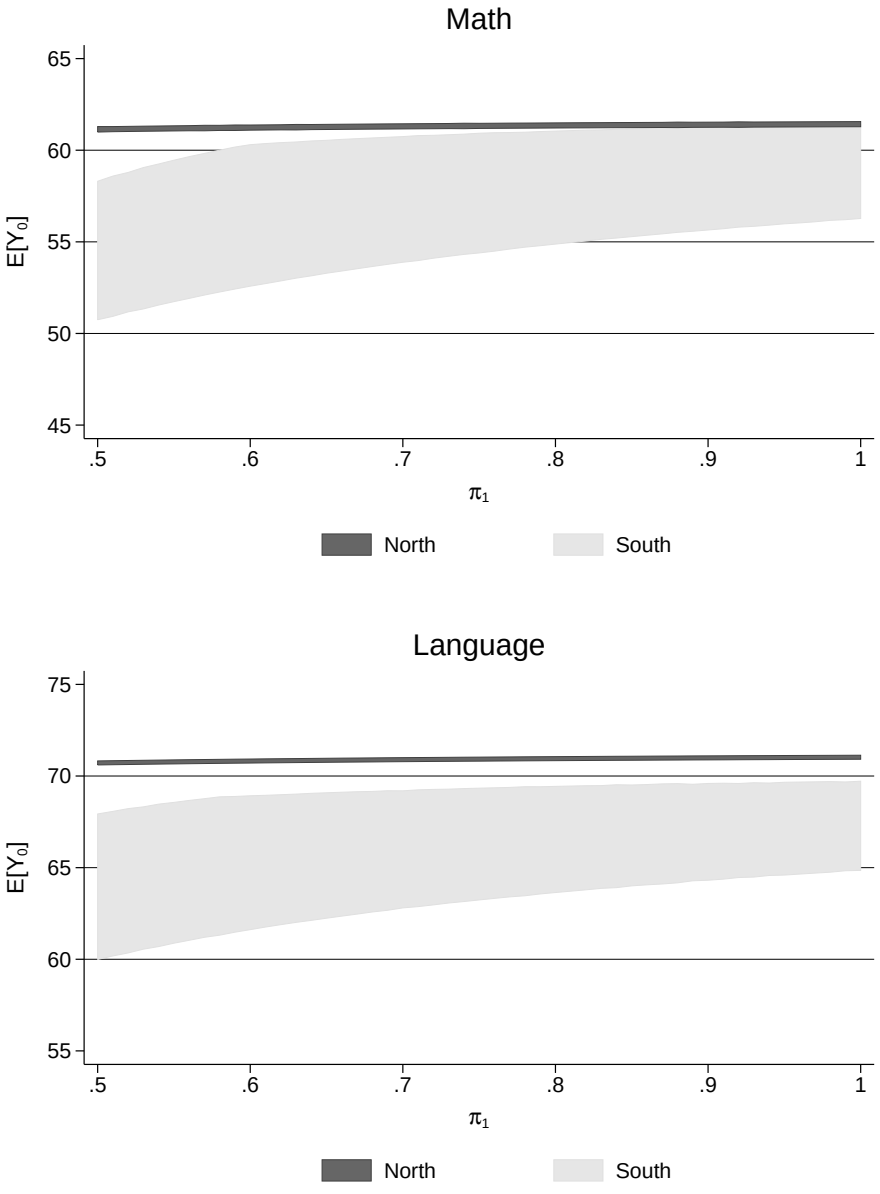
Note. This figure shows the score distribution for math and language under manipulation for dishonest (left hand side panels) and complying (right hand side panels) teachers. Only classes in the South are considered. Shaded areas are bootstrap 95% confidence intervals (see Section 4 for details).

Figure 5: Naive Bounds on $E[Y_0]$



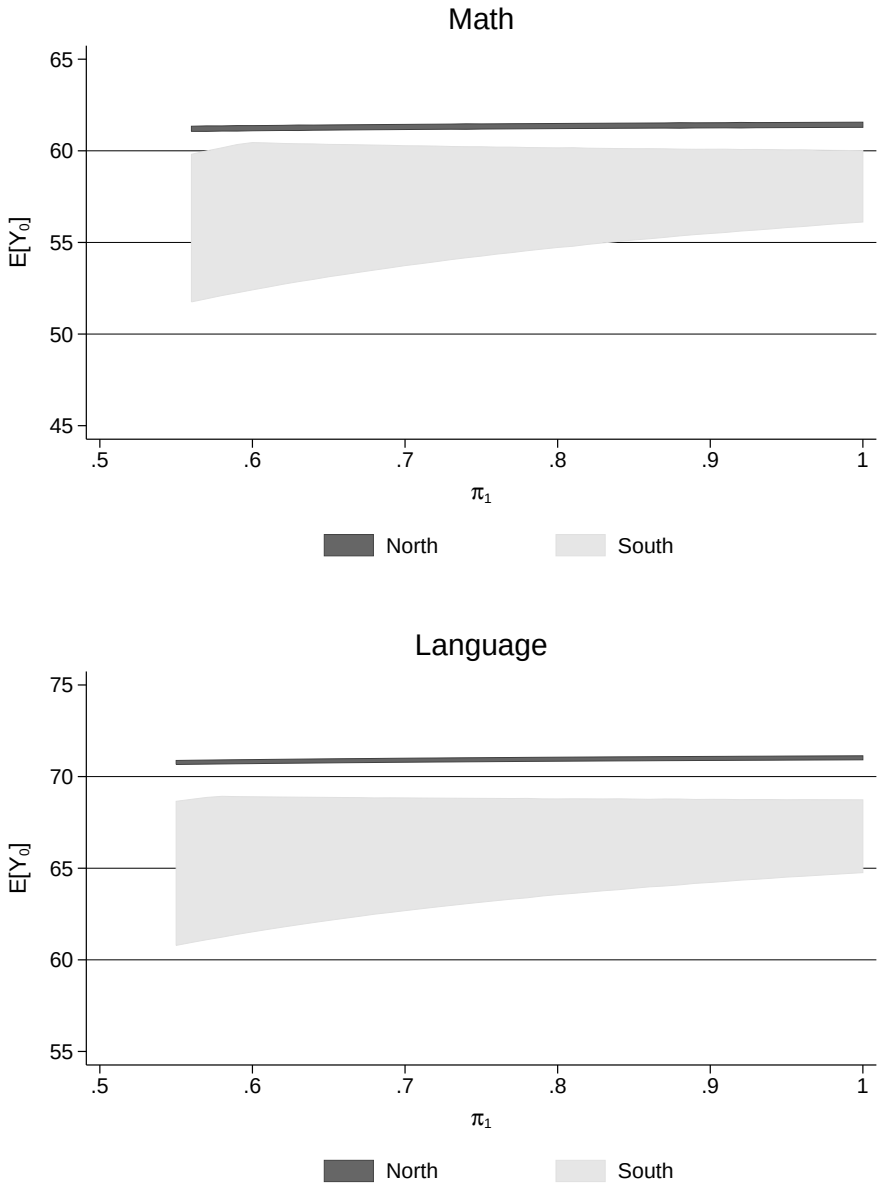
Note. This figure shows naive bounds for average math and language scores, separately for North and South, for $\pi_1 \geq 0.5$. We impose $\phi_D = 0$ in the North. Shaded areas are 95% confidence intervals at each value of π_1 using the procedure by Horowitz and Manski (2000).

Figure 6: Bounds on $E[Y_0]$ using Behavioral Restrictions



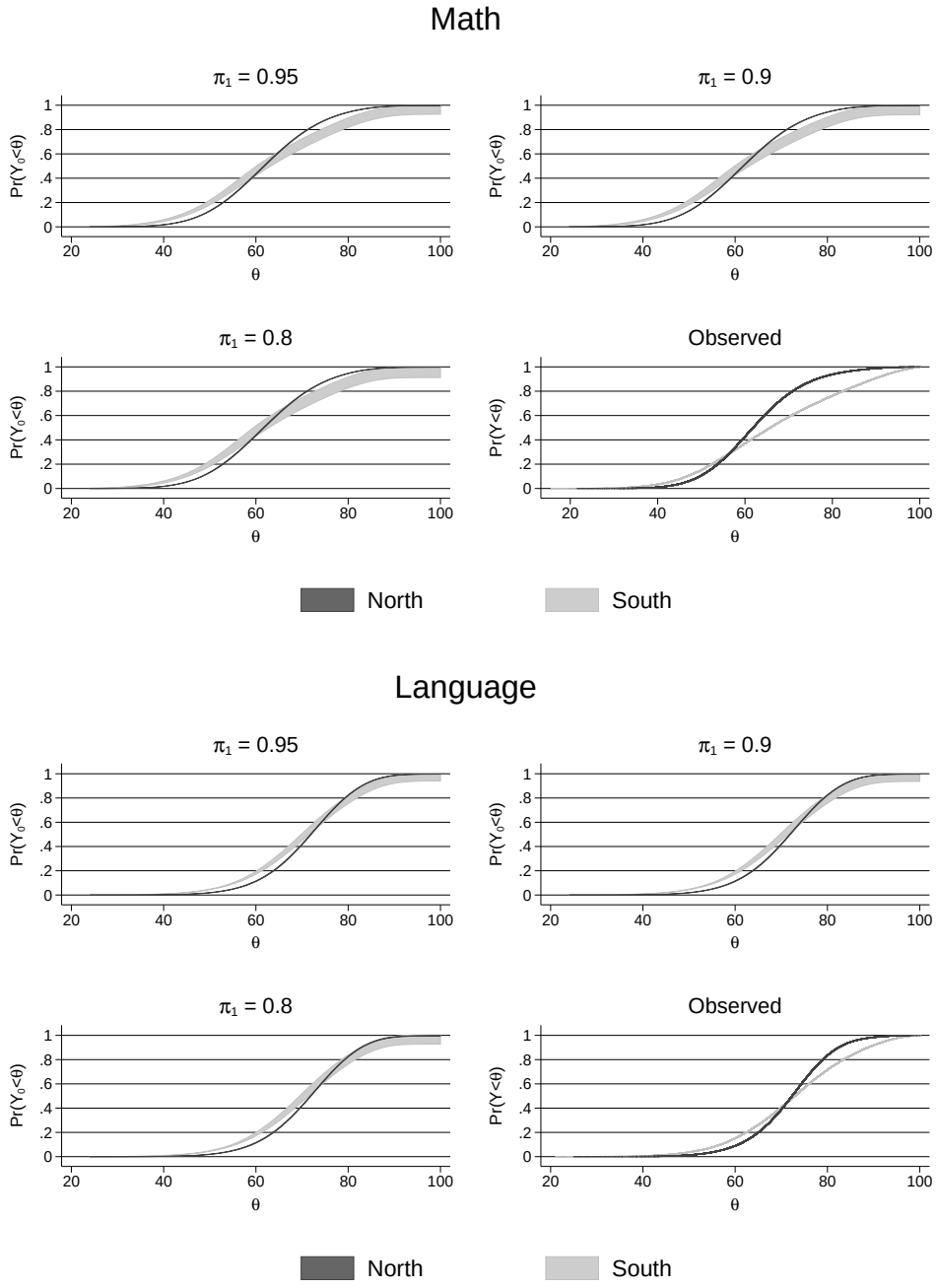
Note. This figure shows bounds for average math and language scores when Assumption 7 is imposed, separately for North and South, for $\pi_1 \geq 0.5$. We impose $\phi_D = 0$ in the North. Shaded areas are 95% confidence intervals at each value of π_1 using the procedure by Horowitz and Manski (2000).

Figure 7: Bounds on $E[Y_0]$ using Full Ranking of Types



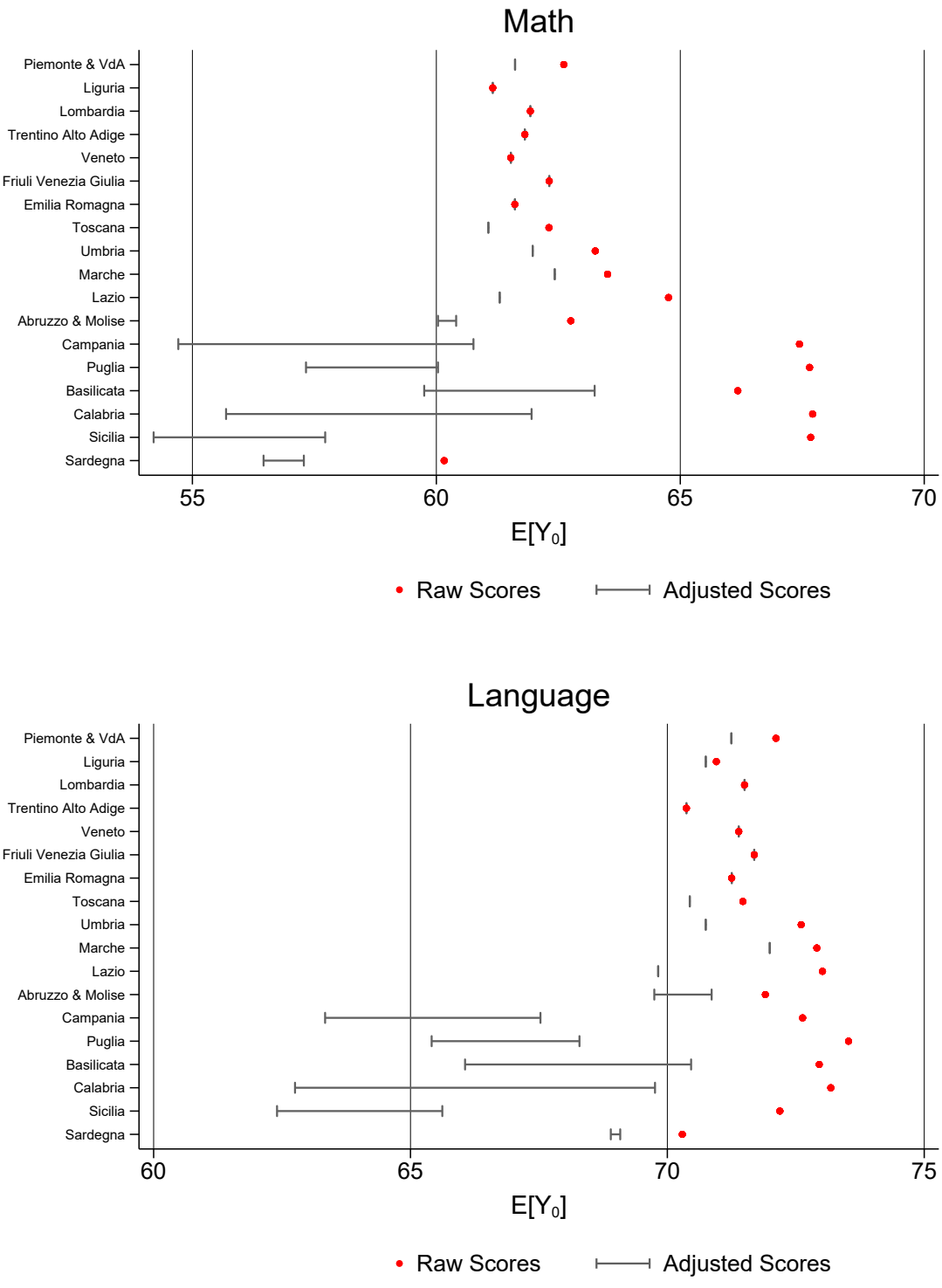
Note. This figure shows bounds for average math and language scores when Assumption 7 and Assumption 8 are imposed, separately for North and South, for $\pi_1 \geq 0.5$. We impose $\phi_D = 0$ in the North. Shaded areas are 95% confidence intervals at each value of π_1 using the procedure by Horowitz and Manski (2000).

Figure 8: Bounds on Score Distributions



Note. This figure shows bounds on math and language scores distributions when Assumption 7 and Assumption 8 are imposed, separately for North and South, at selected values of π_1. We impose $\phi_D = 0$ in the North. Shaded areas are 95% confidence intervals using the preocedure by Horowitz and Manski (2000).

Figure 9: Regional Rankings using Raw and Adjusted Scores



Note. This figure shows average scores from raw data and bounds on the average of true scores, the latter being obtained under Assumption 7 and Assumption 8 when $\pi_1 = 0.9$ (see Section 6 for details).

Appendix

Relationship between scores and socio-economic characteristics

We use a simple model in an effort to explain how manipulation may reverse the relationship between scores and socio-economic indicators (see the discussion at the end of Section 6). This is motivated by results in Figure 2. Assume the following relationship between true scores, Y_0, and school inputs, X:

$$E[Y_0|X = x, g] = \alpha_{0g} + \beta_0 x,$$

where $E[X] = 0$ and $g = C, D, H$. Assume $\beta_0 \geq 0$ without loss of generality, so that inputs above the average are associated with better scores. Assume also:

$$\alpha_{0D} \simeq \alpha_{0C} \equiv \alpha_{0\bar{H}} < \alpha_{0H},$$

implying that, at common values $X = x$, scores for D and C teachers are approximately equal and below scores for H teachers. This setting is consistent with Assumption 8. Manipulation takes the form:

$$E[Y_1|X = x, g] = \alpha_1 \simeq 100,$$

and follows from nearly perfect curbstoning (see Figure 4; for a discussion on the anatomy of manipulation see Angrist et al., 2017).

Consider institutions without monitors ($Z = 0$), where all complying teachers manipulate. The following quantities are defined (the conditioning on $Z = 0$ is left implicit throughout):

$$E[Y_0|X = x] = (\alpha_{0H} + \beta_0 x)\phi_H(x) + (\alpha_{0\bar{H}} + \beta_0 x)[1 - \phi_H(x)],$$

$$E[Y|X = x] = (\alpha_{0H} + \beta_0 x)\phi_H(x) + \alpha_1[1 - \phi_H(x)],$$

where $\phi_H(x)$ is the fraction of H teachers at $X = x$. The expressions above imply:

$$E[Y|X = x] = E[Y_0|X = x] + \alpha_1[1 - \phi_H(x)] - (\alpha_{0\bar{H}} + \beta_0 x)[1 - \phi_H(x)].$$

and the following expression for the covariance between Y and X:

$$Cov[Y, X] = Cov[Y_0, X] - (\alpha_1 - \alpha_{0\bar{H}})Cov[\phi(X), X] - \beta_0 \left\{Var[X] - E[X^2\phi_H(X)]\right\}.$$

Dividing both sides by $Var(X)$ we obtain:

$$\frac{Cov[Y,X]}{Var[X]} = \frac{Cov[Y_0,X]}{Var[X]} - (\alpha_1 - \alpha_{0\bar{H}})\delta_H - \beta_0 \{1-\kappa_H\},$$

where $\kappa_H = E[X^2\phi_H(X)]/Var(X) \in (0,1)$ and $\delta_H = Cov(\phi_H(X),X)/Var(X)$. This expression relates the slope of the linear regression of Y on X, on the left hand side (see Figure 2), to the slope of the linear regression of Y_0 on X, on the right hand side. Manipulation boosts scores, implying $\alpha_1 - \alpha_{0\bar{H}} \geq 0$. Notice that, if manipulation is nearly perfect curbstoning, the term $\alpha_1 - \alpha_{0\bar{H}}$ can be fairly large. If $\delta_H \geq 0$, an assumption likely to hold in our data, we have $Cov[Y_0,X] \geq 0$ and that the second and third terms on the right hand side of the last expression are negative. This may yield sign reversion, for example if the gradient of $\phi_H(X)$, amplified by $(\alpha_1 - \alpha_{0\bar{H}})$, exceeds the gradient of Y_0.

Now consider institutions with monitors ($Z=1$), where all complying teachers transcribe scores honestly. The quantity $E[Y_0|X=x]$ coincides with that for unmonitored ($Z=0$) institutions. By denoting with $\phi_C(x)$ the fraction of C teachers at $X=x$, we have (the conditioning on $Z=1$ is left implicit throughout):

$$E[Y|X=x] = \alpha_1 + (\alpha_{0H} - \alpha_1)\,\phi_H(x) + (\alpha_{0\bar{H}} - \alpha_1)\,\phi_C(x) + \beta_0 x\,[\phi_H(x) + \phi_C(x)],$$

which, by re-arranging terms, implies:

$$\frac{Cov[Y,X]}{Var\,[X]} = \frac{Cov[Y_0,X]}{Var\,[X]} - (\alpha_1 - \alpha_{0\bar{H}})\,\widetilde{\delta}_H - \beta_0 \{1-\widetilde{\kappa}_H\},$$

where $\widetilde{\kappa}_H = E\left[X^2\left\{\phi_H(X) + \phi_C(X)\right\}\right]/Var(X) \in (0,1)$ and $\widetilde{\delta}_H = Cov\left[\left\{\phi_H(X) + \phi_C(X)\right\},X\right]/Var(X)$. If $\widetilde{\delta}_H \geq 0$, sign reversion can follows from arguments similar to those discussed above.

Proofs of Propositions

Proposition 1

Assumption 5 implies:

$$E[W(1-Q)|Z=z] = (1-\pi_0)E[1-Q|Z=z] + (\pi_0+\pi_1-1)E[M|Z=z],$$

$$E[g(Y)W(1-Q)|Z=z] = (1-\pi_0)E[g(Y)(1-Q)|Z=z] + (\pi_0+\pi_1-1)E[g(Y)M|Z=z].$$

These expressions can be solved for the unknowns (i.e., quantities that depend on M) and substituted into equations (1)-(4). Start from:

$$E[M|Z=z] = \frac{E[\Lambda|Z=z]}{(\pi_0+\pi_1-1)},$$
$$E[g(Y)M|Z=z] = \frac{E[g(Y)\Lambda|Z=z]}{(\pi_0+\pi_1-1)}.$$

It follows that:

$$E[g(Y_1)|D] = \frac{E[g(Y)\Lambda|Z=1]}{E[\Lambda|Z=1]}.$$

The ratio of the quantities:

$$E[M|Z=1] - E[M|Z=0] = \frac{E[\Lambda|Z=1] - E[\Lambda|Z=0]}{(\pi_0+\pi_1-1)},$$

$$E[g(Y)M|Z=1] - E[g(Y)M|Z=0] = \frac{E[g(Y)\Lambda|Z=1] - E[g(Y)\Lambda|Z=0]}{(\pi_0+\pi_1-1)},$$

implies:

$$E[g(Y_1)|C] = \frac{E[g(Y)\Lambda|Z=1] - E[g(Y)\Lambda|Z=0]}{E[\Lambda|Z=1] - E[\Lambda|Z=0]}.$$

Now consider:

$$E[g(Y)(1-M)|Z=z] = E\left[g(Y)\left\{1 - \frac{\Lambda}{(\pi_0+\pi_1-1)}\right\}|Z=z\right],$$

and:

$$E[(1-M)|Z=z] = E\left[1 - \frac{\Lambda}{(\pi_0+\pi_1-1)}|Z=z\right].$$

Since we can write:

$$1 - \frac{\Lambda}{(\pi_0+\pi_1-1)} = \frac{\Psi(\pi_1)}{(\pi_0+\pi_1-1)},$$

it is:

$$E[g(Y_0)|C] = \frac{E[g(Y)\Psi(\pi_1)|Z=1] - E[g(Y)\Psi(\pi_1)|Z=0]}{E[\Psi(\pi_1)|Z=1] - E[\Psi(\pi_1)|Z=0]}.$$

The same calculations also imply:

$$E[g(Y_0)|H] = \frac{E[g(Y)\Psi(\pi_1)|Z=0]}{E[\Psi(\pi_1)|Z=0]}.$$

Proposition 2

Using the representations in the proof of Proposition 1 we can write:

$$\phi_D = E[M|Z=1] = \frac{E[\Lambda|Z=1]}{(\pi_0+\pi_1-1)},$$
$$1-\phi_H = E[M|Z=0] = \frac{E[\Lambda|Z=0]}{(\pi_0+\pi_1-1)}.$$

Proposition 3

Under Assumptions 3, 4 and 5 we have:

$$E[W|Q=1,Z=1] = (1-\pi_0) + (\pi_0+\pi_1-1)\,E[M|Q=1,Z=1] = 1-\pi_0.$$

Proposition 4

Assume that π_1 is known. Inequality (15) implies:

$$E[g(Y_0)|D] \leq E[g(Y_0)|C]\frac{\phi_C}{1-\phi_D} + E[g(Y_0)|H]\frac{\phi_H}{1-\phi_D},$$

which when substituted into (5) yields:

$$E[g(Y_0)] \leq E[g(Y_0)|C]\frac{\phi_C}{1-\phi_D} + E[g(Y_0)|H]\frac{\phi_H}{1-\phi_D}.$$

Inequality (16) implies:

$$E[g(Y_0)|D] \leq E[g(Y_0)|H]\frac{(1-\phi_H)}{\phi_D} - E[g(Y_0)|C]\frac{\phi_C}{\phi_D}.$$

Substituting into (5) we have:

$$E[g(Y_0)] \leq E[g(Y_0)|H](1-\phi_H) + E[g(Y_0)|H]\phi_H = E[g(Y_0)|H].$$

The two inequalities derived, taken jointly, imply:

$$E[g(Y_0)] \leq \min\left\{E[g(Y_0)|C]\frac{\phi_C}{1-\phi_D} + E[g(Y_0)|H]\frac{\phi_H}{1-\phi_D}, E[g(Y_0)|H]\right\},$$

which is the expression for the upper bound. The naive upper bound can be written as:

$$E[g(Y_1)|D]\phi_D + E[g(Y_0)|C]\phi_C + E[g(Y_0\}|H]\phi_H. \tag{21}$$

If $E[g(Y_1)|D]$ is larger than $E[g(Y_0)|C]$ and $E[g(Y_0)|H]$, as it is likely to be the case in our application, the new upper bound is more informative than the naive upper bound.

Inequality (16) acts on the lower bound by imposing the following restriction:

$$E[g(Y_0)] \geq E[g(Y_0)|D]\frac{\phi_D}{1-\phi_H} + E[g(Y_0)|C]\frac{\phi_C}{1-\phi_H}.$$

The expression for the lower bound follows by bounding $E[g(Y_0)|D]$ from below:

$$E[g(Y_0)] \geq g(0)\frac{\phi_D}{1-\phi_H} + E[g(Y_0)|C]\frac{\phi_C}{1-\phi_H}.$$

The naive lower bound can be written as:

$$g(0)\phi_D + E[g(Y_0)|C]\phi_C + E[g(Y_0)|H]\phi_H,$$

which can be smaller or larger than the bound in this proposition.

Proposition 5

Assume that π_1 is known. The second inequality in Assumption 8 bounds from above the counterfactual term $E[g(Y_0)|D]$. If substituted into (5), this yields:

$$E[g(Y_0)] \leq E[g(Y_0)|C](1-\phi_H) + E[g(Y_0)|H]\phi_H,$$

which is the expression for the upper bound. Notice that the ranking across types implies that the difference between upper bounds from Proposition 5 and Proposition 4 is equal to:

$$\{E[g(Y_0|H] - E[g(Y_0)|C]\}\,\frac{\phi_H\phi_D}{1-\phi_D} \geq 0,$$

implying that the upper bound here is tighter than the upper bound in Proposition 4.

As for the lower bound, the first inequality in Assumption 8 implies:

$$E[g(Y_0)] \geq E[g(Y_0)|D]\phi_D + E[g(Y_0)|C](1-\phi_D) \geq g(0)\phi_D + E[g(Y_0)|C](1-\phi_D).$$

This defines a new lower bound on (5). Remember that the naive lower bound is: $g(0)\phi_D + E[g(Y_0)|C]\phi_C + E[g(Y_0)|H]\phi_H$. Therefore the difference between the latter and the one obtained imposing Assumption 8 is:

$$E[g(Y_0)|C](\phi_C - \phi_C - \phi_H) + E[g(Y_0)|H]\phi_H = (E[g(Y_0)|H] - E[g(Y_0)|C])\phi_H$$

which is always positive.

Table A1: Descriptive Statistics

	Italy	North	South
	(1)	(2)	(3)
	Panel A. Math		
Raw score	64.042	62.419	66.747
	(13.027)	(10.628)	(15.892)
Presumed manipulators	0.066	0.020	0.142
	(0.248)	(0.141)	(0.349)
Monitored classes	0.069	0.069	0.069
	(0.253)	(0.253)	(0.253)
	Panel B. Language		
Raw score	72.077	71.784	72.565
	(10.172)	(8.730)	(12.187)
Presumed manipulators	0.056	0.023	0.112
	(0.231)	(0.150)	(0.316)
Monitored classes	0.069	0.069	0.069
	(0.253)	(0.253)	(0.253)
	Panel C. Other covariates		
Monitored institutions	0.238	0.247	0.222
	(0.426)	(0.431)	(0.416)
Second grade	0.482	0.489	0.471
	(0.500)	(0.500)	(0.499)
2009 survey	0.343	0.341	0.347
	(0.475)	(0.474)	(0.476)
2010 survey	0.329	0.330	0.329
	(0.470)	(0.470)	(0.470)
2011 survey	0.327	0.329	0.324
	(0.469)	(0.470)	(0.468)
Number of classes	140,010	87,498	52,512

Note. This table presents descriptive statistics from INVALSI data pooling second and fifth grade students for the years 2009-2011. Standard deviations in parentheses.

Table A2: Estimates of Average Counterfactual Scores by Compliance Types and Percentages of Honest, Complying and Dishonest Teachers

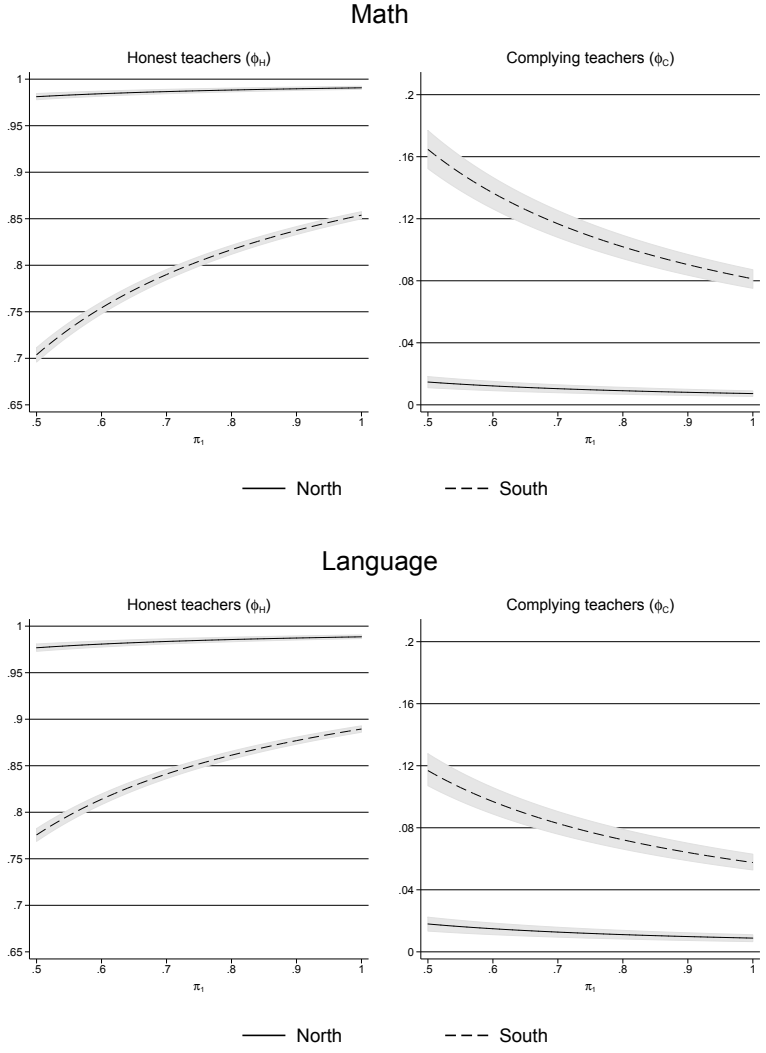
	North			**South**		
	Complying	Dishonest	Honest	Complying	Dishonest	Honest
	(1)	(2)	(3)	(4)	(5)	(6)
			Panel A. Math			
$E[Y_1]$	90.469	-	-	91.866	90.423	-
	(0.210)	-	-	(0.483)	(0.090)	-
$E[Y_0]$	27.226	-	62.156	24.194	-	63.133
	(3.895)	-	(0.071)	(4.194)	-	(0.140)
ϕ	0.020	-	0.980	0.072	0.085	0.842
	(0.000)	-	(0.000)	(0.004)	(0.003)	(0.002)
			Panel B. Language			
$E[Y_1]$	91.249	-	-	92.373	91.201	-
	(0.149)	-	-	(0.417)	(0.073)	-
$E[Y_0]$	57.846	-	71.480	39.176	-	70.429
	(2.424)	-	(0.056)	(4.087)	-	(0.106)
ϕ	0.023	-	0.977	0.051	0.072	0.876
	(0.001)	-	(0.001)	(0.003)	(0.003)	(0.002)

Note. This table shows estimates of average counterfactual scores by compliance types and percentages of honest, complying and dishonest teachers. All terms are obtained from 2SLS regressions similar to those described in Section 3, assuming that classes with manipulated scores are correctly classified.

Additional Materials (for on-line publication)

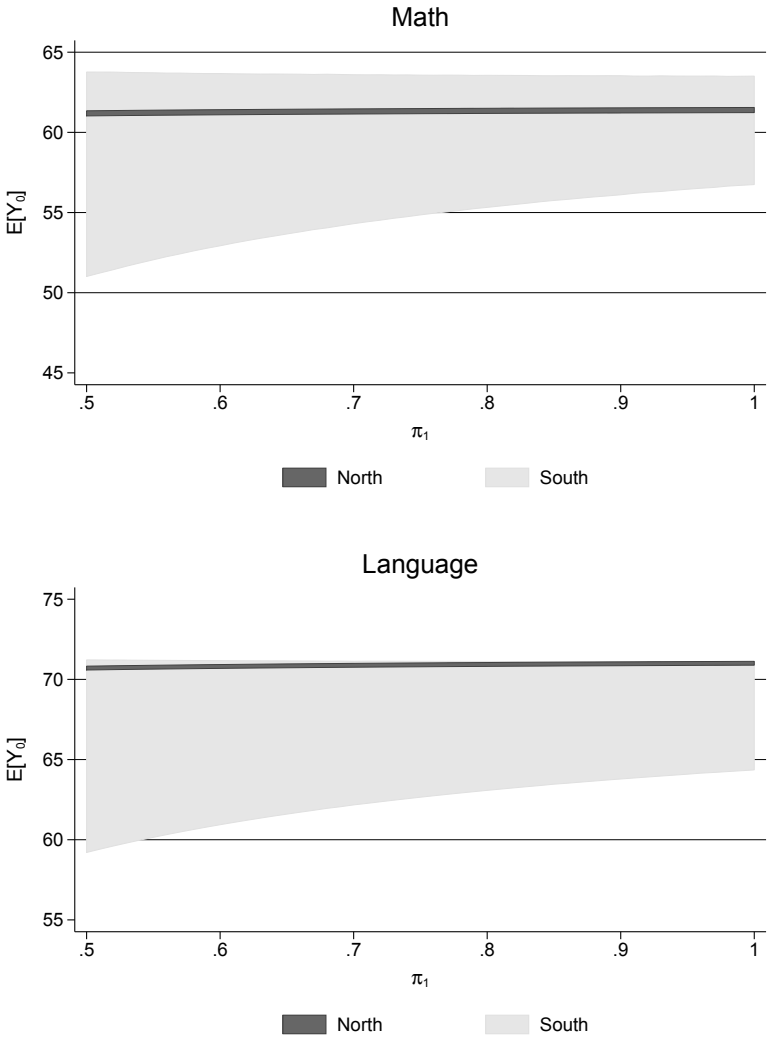
On-line Appendix

Figure B1: Percentage of Honest and Complying Teachers



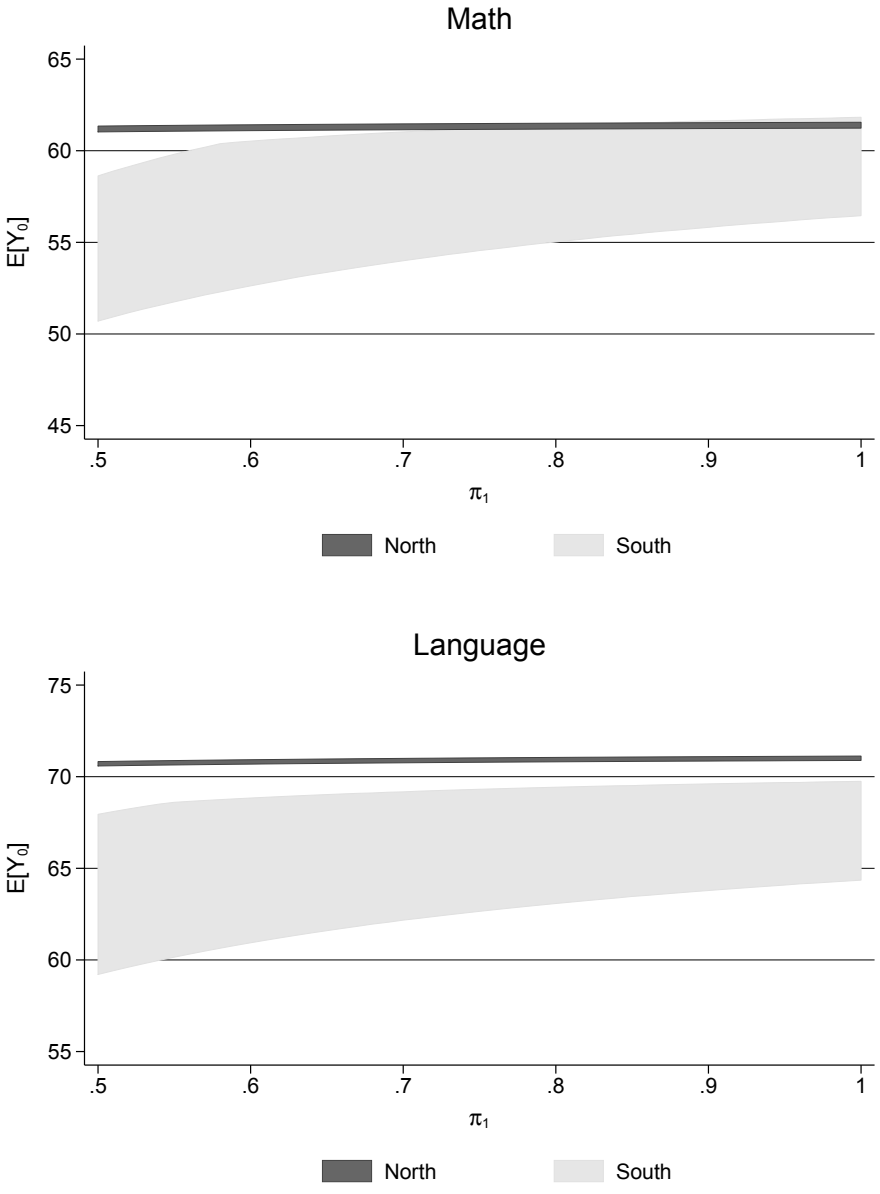
Note. Estimation results from non-parametric IV estimation (Frölich 2007). The figure reports the percentage of honest (ϕ_H) and complying (ϕ_C) teachers. Results are presented for the interval $\pi_1 \geq 0.5$, separately for the North (continuous line) and the South (dashed line). Shaded areas are 95% bootstrap confidence intervals obtained at each value of π_1.

Figure B2: Naive Bounds on $E[Y_0]$



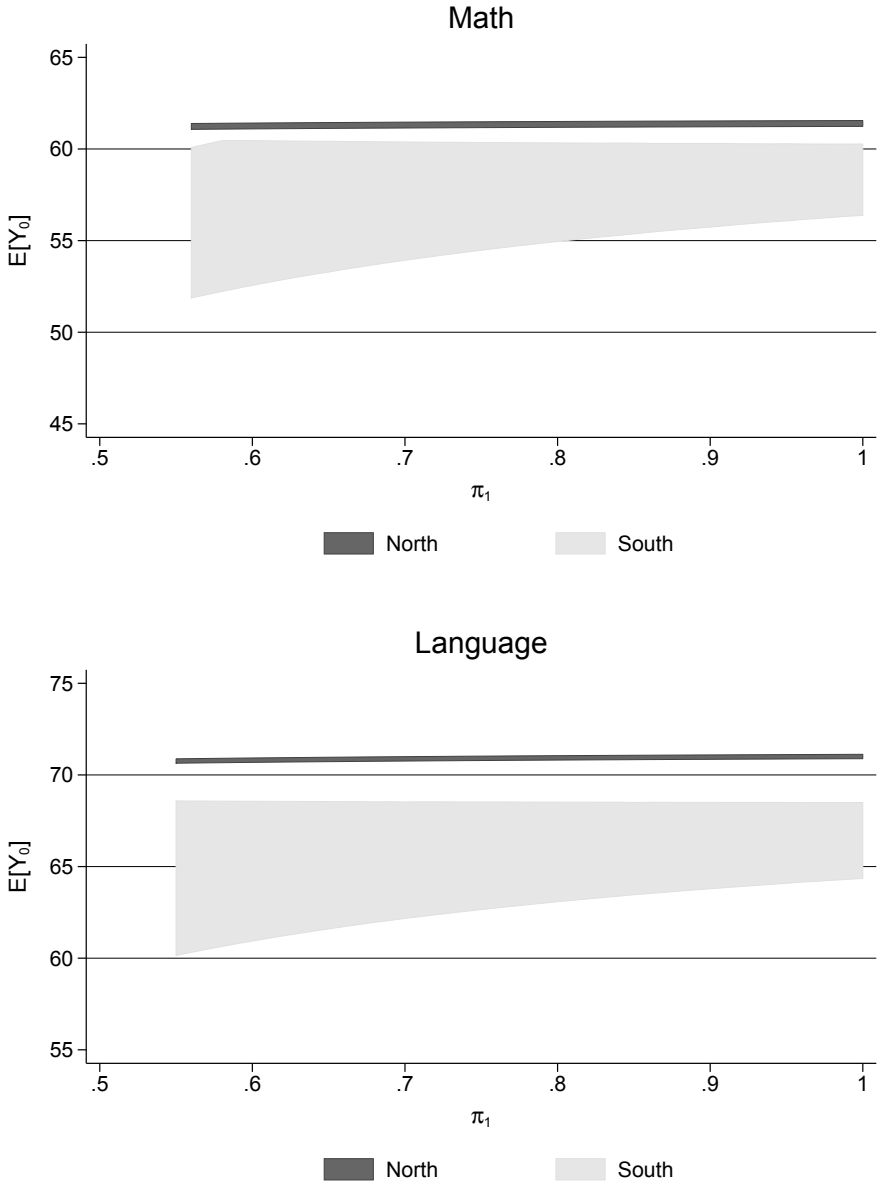
Note. Estimation results from non-parametric IV estimation (Frölich 2007). The figure shows naive bounds for $\pi_1 \geq 0.5$. We impose $\phi_D = 0$ in the North. Shaded areas are 95% confidence intervals at each value of π_1 using the procedure by Horowitz and Manski (2000).

Figure B3: Bounds on $E[Y_0]$ using Behavioral Restrictions



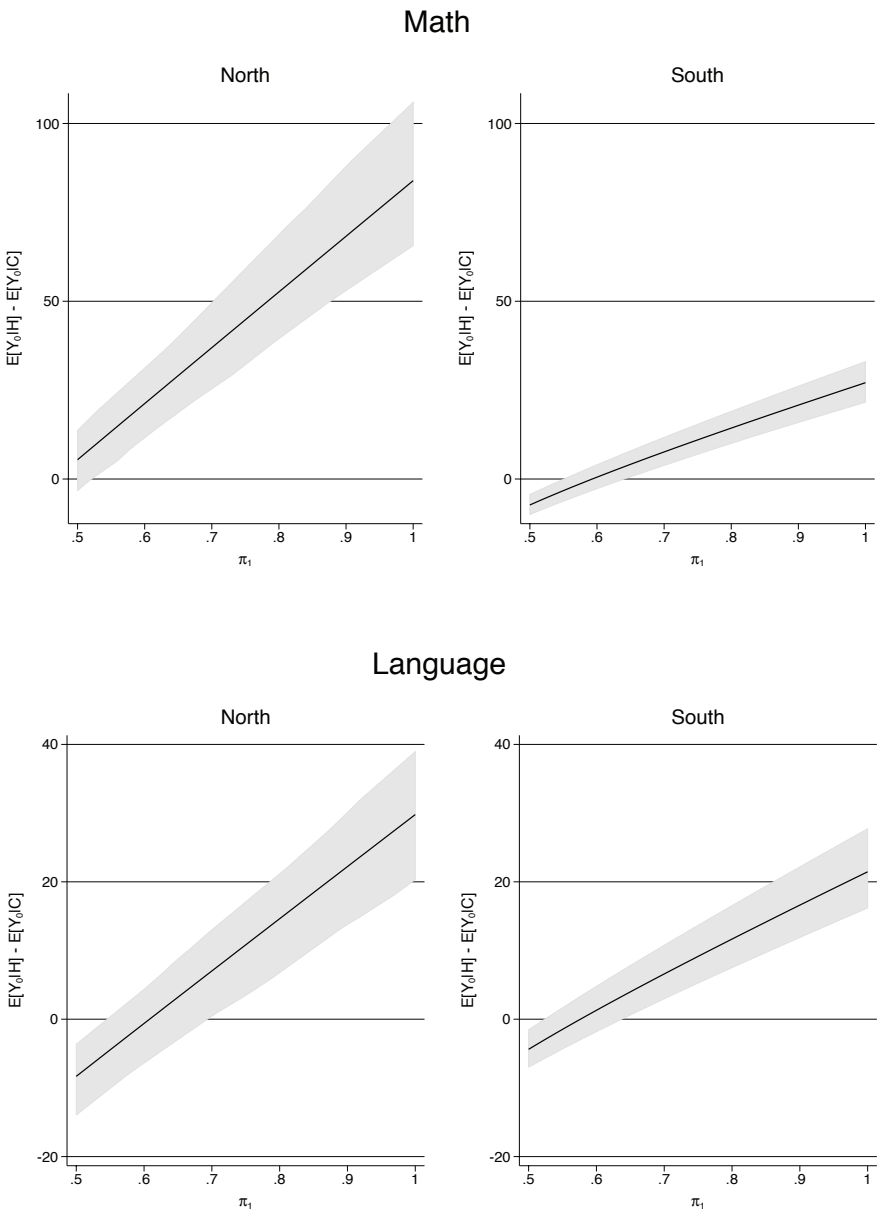
Note. The figure shows bounds for average math and language scores when Assumption 7 is imposed, separately for North and South, for $\pi_1 \geq 0.5$. We impose $\phi_D = 0$ in the North. Shaded areas are 95% confidence intervals at each value of π_1 using the procedure by Horowitz and Manski (2000).

Figure B4: Bounds on $E[Y_0]$ using Full Ranking of Types



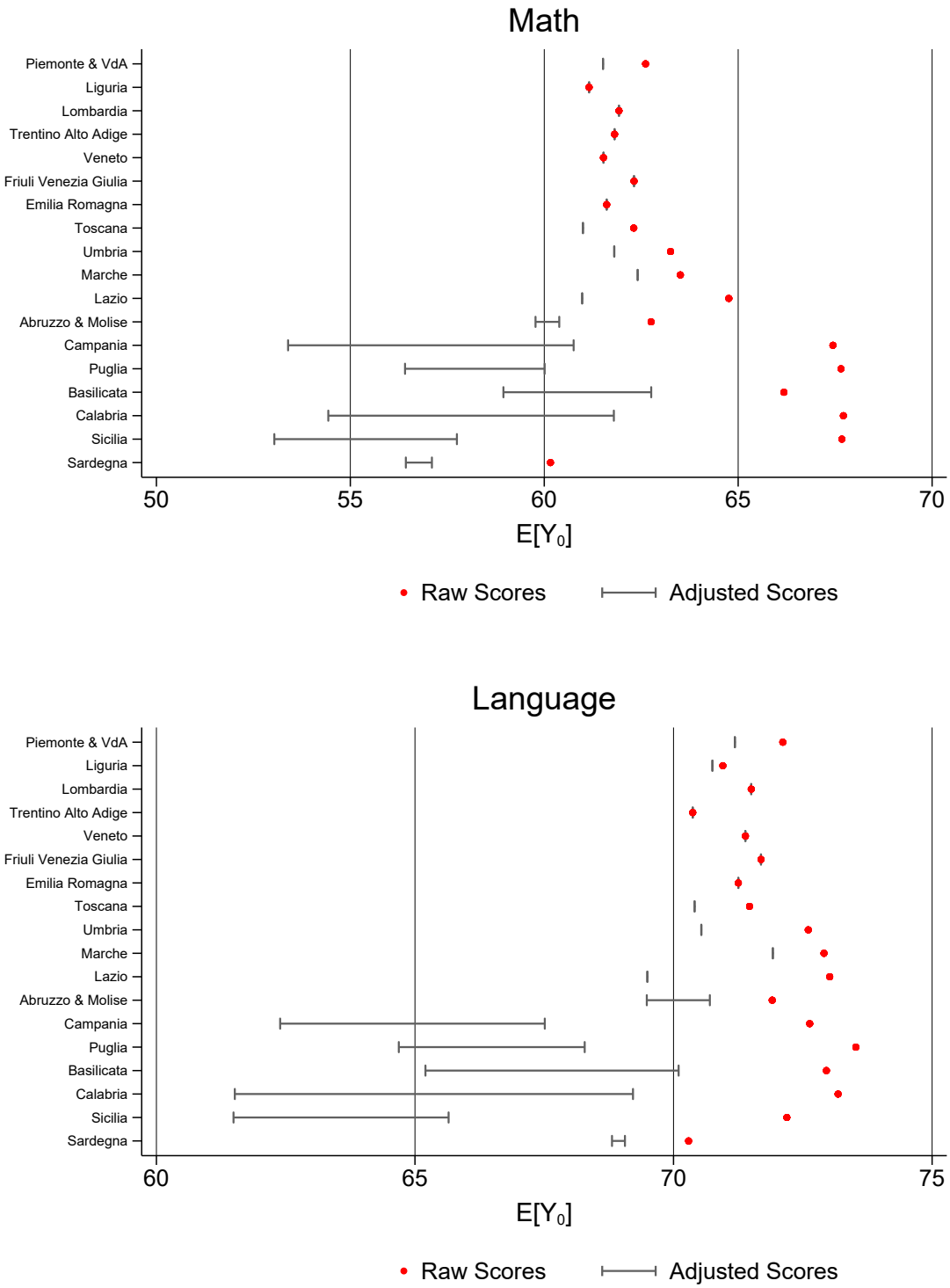
Note. The figure shows bounds for average math and language scores when Assumption 7 and Assumption 8 are imposed, separately for North and South, for $\pi_1 \geq 0.5$. We impose $\phi_D = 0$ in the North. Shaded areas are 95% confidence intervals at each value of π_1 using the procedure by Horowitz and Manski (2000).

Figure B5: Confidence Intervals for $E[Y_0|H] - E[Y_0|C]$



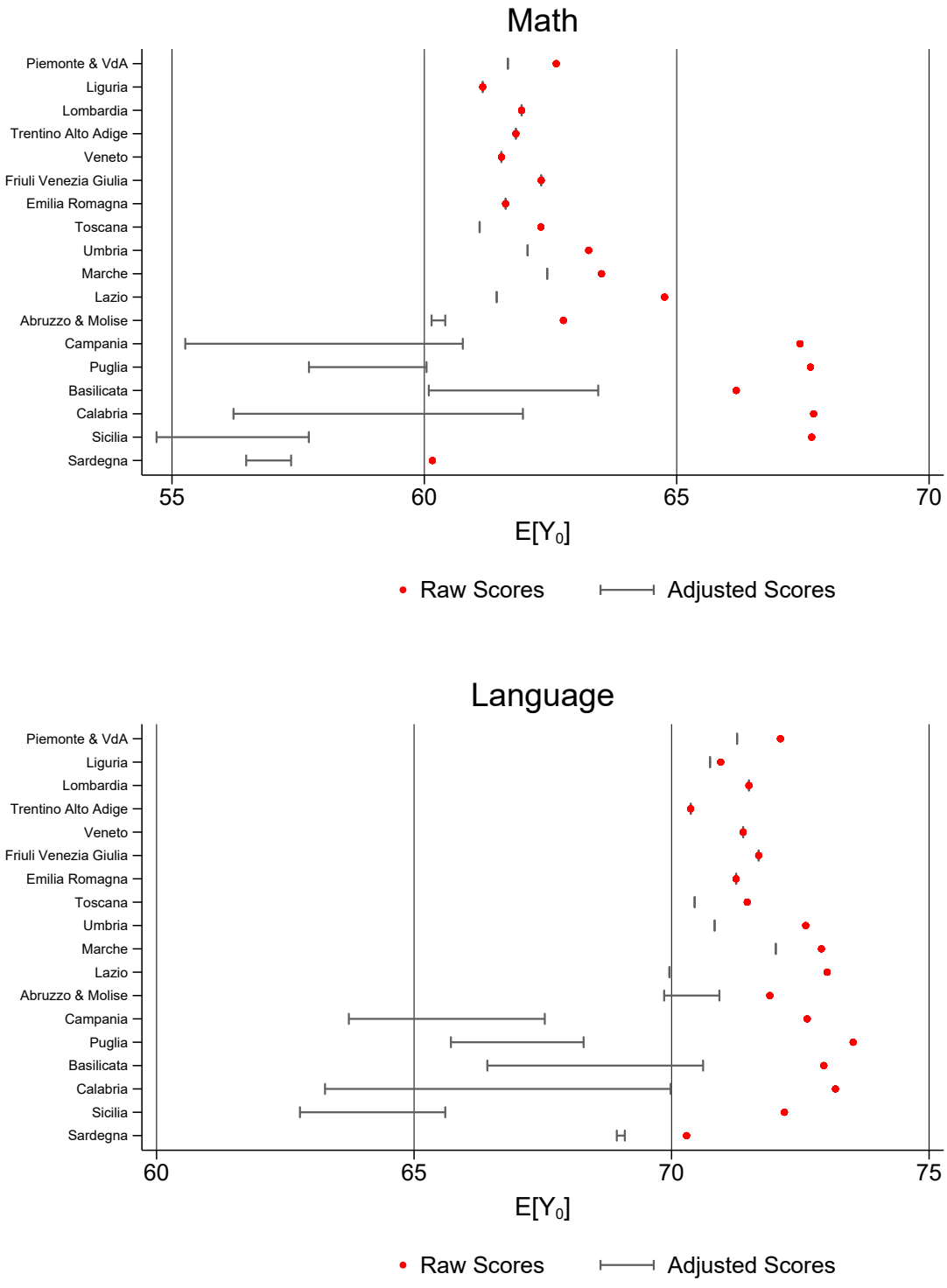
Note. This figure reports $E[Y_0|H]$-$E[Y_0|C]$ at different values of π_1. Shaded areas represent 95% boostrap confidence intervals.

Figure B6: Regional Rankings using Raw and Adjusted Scores for $\pi_1 = 0.8$



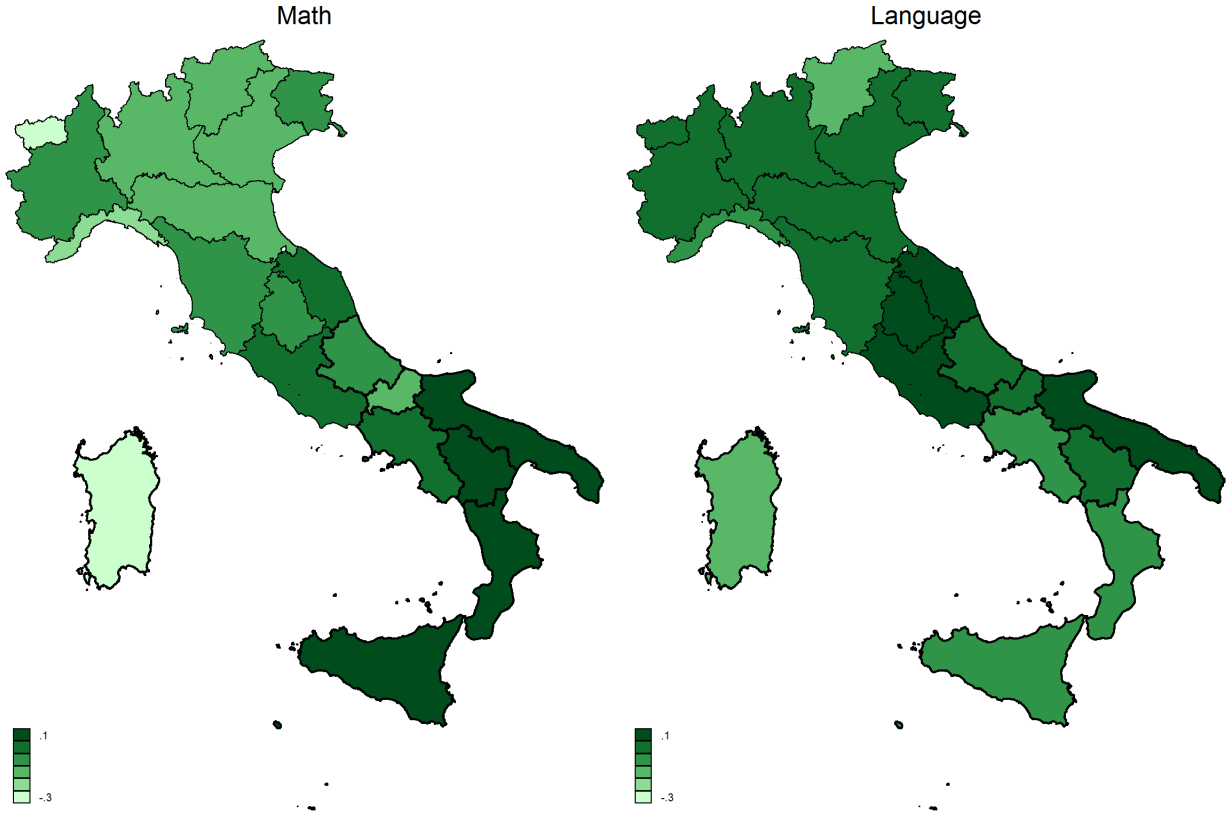
Note. This figure shows average scores from raw data and bounds on the average of true scores, the latter being obtained under Assumption 7 and Assumption 8 when $\pi_1 = 0.8$ (see Section 6 for details).

Figure B7: Regional Rankings using Raw and Adjusted Scores for $\pi_1 = 0.95$



Note. This figure shows average scores from raw data and bounds on the average of true scores, the latter being obtained under Assumption 7 and Assumption 8 when $\pi_1 = 0.95$ (see Section 6 for details).

Figure B8: Adjusted Scores using the INVALSI Methodology



Note. These figures are obtained from INVALSI data pooling second and fifth grade students for the school years 2009-2011, for details about correction see Falzetti (2013).