Teachers and Cheaters. Just an Anagram?

Santiago Pereda-Fernández*

Banca d'Italia

April 18, 2019

Abstract

I study the manipulation of test scores in Italy. Using an experiment that randomly assigns external monitors to classrooms, I apply a new methodology to study the extent of the manipulation, and propose a correction method. I find that the manipulation was associated with more correlation in the answers after one controls for mean test scores. It was concentrated in the South and Islands region, and it tended to favor female and immigrant students. Finally, the correlation patterns between the amount of manipulation and the number of missing answers suggests that teachers were more responsible for the manipulation than students.

Keywords: Cheating Correction, Copula, Discrimination, Gender, Nonlinear Panel Data, Test Scores Manipulation

JEL classification: C23, C25, I21, I28, J24

*Banca d'Italia, Via Nazionale 91, 00184 Roma, Italy. This paper was previously circulated with the name *A New Method for the Correction of Test Scores Manipulation.* I would like to thank Alessandro Belmonte, Stéphane Bonhomme, Giulia Bovini, Nicola Curci, Domenico Depalo, Patrizia Falzetti, Raquel Fernández, Iván Fernández-Val, Guzmán González-Torres, Caroline Hoxby, Andrea Ichino, Claudio Michelacci, Marco Savegnago, Paolo Sestito, Martino Tasso, Jeffrey Wooldridge, Paolo Zacchia, Stefania Zotteri, two associate editors, two anonymous referees, and seminar participants at Banca d'Italia, EUI, IMT Lucca, Universidad de Alicante, Universidad de Cantabria and the 2nd IAAE for helpful comments and suggestions. All remaining errors are my own. The views presented in this paper do not necessarily reflect those of the Banca d'Italia. I can be reached via email at santiago.pereda@.bancaditalia.it.

1 Introduction

A policy-maker interested in evaluating the education system requires a comparable measure of academic achievement across students. Standardized tests permit the comparison of students' performance, and are often used to evaluate teachers (Hanushek, 1971; Rockoff, 2004; Aaronson et al., 2007) and principals (Grissom et al., 2014), although the reliability of these estimates has been called into question (Rothstein, 2010, 2017; Chetty et al., 2014).

A major threat to the comparability of these tests is the manipulation of the scores, which alters students' recorded performance.[1] There is ample evidence that tests are susceptible of being manipulated, either by teachers grading unfairly (Jacob and Levitt, 2003; Dee et al., 2011; Battistin et al., 2016; Diamond and Persson, 2016), by students copying each other (Levitt and Lin, 2015), or even by principals who alter the pool of students who take the exam (Figlio, 2006; Cullen and Reback, 2006; Hussain, 2015).

Ideally, one would like to correct the test scores to reverse the manipulation. This is challenging, since manipulation of an individual test is not observed and it can be confounded with good performance. However, if the amount of manipulation varies with the class' and students' characteristics, it becomes possible to identify which groups of students benefit most from the manipulation.

In this paper, I study the extent of test scores manipulation, taking advantage of a natural experiment in the Italian education system that randomly assigned external monitors to proctor some tests. In particular, I make the following contributions. First, I propose a new methodology to identify which demographic groups benefited the most from the manipulation. On top of already known results, I find that the manipulation systematically favored female over male students, and immigrants over natives. Second, I propose a method to detect and correct manipulated test scores based on how likely the observed results would be if they were not manipulated.

[1] Throughout this paper I refer to test score manipulation or cheating as any action taken by the students or the teachers that results in a variation of the test scores, usually an increase. This could take place before the test (alteration of the pool of students), during the test (students copying from one another, teachers turning a blind eye or telling the answers), or after the test (unfair grading, including leniency).

Studies on education often rely on raw test scores as a measure of students' achievement and are frequently standardized to have zero mean and unit standard deviation. However, the answers to every single question of the test display a richer correlation structure that can be more informative for detecting manipulation. This correlation stems from factors that operate in a different manner and can be classified into three main categories: individual characteristics, which only affect a single student; class characteristics, which affect every student in the same classroom; and question characteristics, which affect every student, although only in each specific question. Hence, when a question is difficult, a small fraction of students is likely to answer the question correctly, increasing the correlation of students' answers both within and between classrooms. This lends itself to using panel data methods that can accommodate all these types of effects.

To overcome these challenges, the method I propose compares the likelihood of the results of two groups: one in which test scores are assumed to be fair (treatment group) and another in which they might have been manipulated (control group), analogously to a comparison between blind graded and non-blind graded exams (e.g., Lavy (2008) or Hinnerich et al. (2011)). The results in the treatment group are used to estimate the probability of obtaining the observed test scores at random without manipulation. An excessive amount of unlikely results in the control group indicates the existence of manipulated test scores. The larger the difference, the more widespread the manipulation.

The likelihood function accounts for all the previously mentioned effects that create correlation patterns in students' answers without manipulation, as well as the information provided by the students' observable characteristics. Under the assumption that the estimates of the group with an external monitor are not manipulated, differences between the two sets of estimates reflect the amount of manipulation for each demographic group and question.[2] The estimates from the treatment group are subsequently used to calculate the probability of obtaining the observed results without manipulation. This constitutes the basis for the correction method, which estimates the expected amount of manipulation conditional on

[2]This does not imply that the test scores of every student in the control group were manipulated or that the manipulation was of the same magnitude for students with the same characteristics.

how likely the results would have been in the absence of manipulation.

There are two types of misclassification when one attempts to detect manipulated test scores: mistaking fair tests for manipulated (false positives), and mistaking manipulated tests for fair (false negatives). Methods that try to detect cheating are more frequently focused on reducing the number of false negatives. Furthermore, they are often based on the value of some test score statistics that are similar for scores of high-achieving students and manipulated scores. For example, they are both likely to have high class means, or correlated test scores, which could merely reflect effective teaching practices, potentially yielding an excessive amount of false positives. Hence, both potential sources of misclassification should be taken into consideration.

The data I use stem from a set of low stakes standardized tests in the Italian education system. Students in primary, lower secondary, and upper secondary education take two tests in mathematics and Italian language in their own schools, proctored by a teacher from their own school. They are responsible for grading, transcribing the test scores, and sending them back to the National Institute for the Evaluation of the Education System (INVALSI). However, a set of randomly selected classrooms has an external monitor who is responsible for the same tasks, but had no prior connection to the school. This constitutes a large scale natural experiment to study test score manipulation in the absence of an external monitor.

Previous work used the results from preceding years of the primary and lower secondary tests.[3] They found that having an internal monitor is associated with higher, more correlated test scores (Bertoni et al., 2013), which could be the result of students' interactions (Lucifora and Tonello, 2015) or of teachers' shirking at grading (Battistin et al., 2016). Moreover, the amount of manipulation is much larger in the South & Islands of Italy, which is greatly correlated with other measures of social capital (Paccagnella and Sestito, 2014).

I find substantial manipulation in the test scores that is heterogeneous across various dimensions. Apart from the already known geographical patterns, I find that female and

[3]In particular, Bertoni et al. (2013) focused on grades 2 and 5 for the 2010 tests, Battistin et al. (2016) and Battistin et al. (2017) on grades 2 and 5 for the 2010-12 tests, and Lucifora and Tonello (2015) on grade 6 for the 2010 tests.

immigrant students benefited from this manipulation more than their male and native peers. Specifically, the manipulation was up to 1.7% of the maximum score higher for females relative to males in mathematics exams, whereas in Italian exams, it was at most 0.4%. Regarding differences between different ethnic groups, I find that immigrant students in Italy tend to be favored relative to natives, mostly in Italian language exams, in which it can be up to 1.8% of the maximum score.

If students had been responsible for the manipulation, the correlation in their answers would have increased. However, after controlling for the mean scores, this correlation is roughly the same for students in the same classroom, regardless of the monitor type. On the other hand, the correlation is larger for students in the same classroom than for students in different classrooms. Hence, rather than manipulation, the correlation reflects a combination of teacher quality, peer effects, and sorting of students. Also, open-ended questions were more manipulated, and the amount of manipulation was negatively correlated to the fractions of missing open-ended questions relative to multiple choice questions. These patterns are the opposite of what would have arisen if students had copied each other during the exam.

Even though these exams had no formal consequences to teachers (e.g., their wages are not linked to the results), they may have had incentives to manipulate the results if they perceived that they could be evaluated in the future, e.g., if they were to be paid based on the performance of their students, or if principals used the results internally.[4] Hence, manipulation could be a means to invalidate the comparability of the results to prevent their students' test scores from being used to evaluate them.

The rest of the paper is organized as follows. The institutional details of the test and some descriptive statistics are presented in Section 2. The empirical strategy and the correction methods are explained in Section 3. Section 4 shows the results of the estimation, while Section 5 shows the class-level correction in practice. In Section 6 I analyze the possible mechanism behind the results and assess the consequences of the manipulation. Section 7

[4]These concerns, among others, have led to important boycotts of the 2014-15 and 2015-16 tests: in some of the exams, up to 10% of the students did not participate. See `http://www.invalsi.it/invalsi/doc_evidenza/2015/Comunicato_stampa_Prove_INVALSI_2015_07_05.pdf` `http://www.invalsi.it/invalsi/doc_evidenza/2016/Com_Stampa_INVALSI_II_SEC_SEC_GRADO.pdf`.

relates the results to what was previously found in the literature and Section 8 concludes.

2 Italian National Evaluation Test

INVALSI is the institute responsible for the design and administration of standardized tests in Italy. Since the academic year 2008-09, all students enrolled in certain grades are required to take one test in mathematics and another in Italian language. Although the Italian Ministry of Education stated the necessity of establishing a system of evaluation of teachers and schools based on students' performance, the tests have been low stakes for all grades, with the exception of the 8th (*III media*). The latter coincides with the end of the compulsory secondary education, and the results of the test account for a sixth of their final marks.

The exams are taken in classrooms, and students are proctored by either an internal or an external monitor who is also responsible for grading, transcribing the result of each student to a sheet and sending it to INVALSI. Internal monitors are teachers from the same school who were not the students' teacher during the academic year of the test. On the other hand, external monitors are teachers and principals who had not worked in the town of the school they are assigned to for at least two years before the test, while internal monitors are teachers from the same school who were not the teachers of the students taking the test.[5]

External monitors are randomly assigned to classes with the same selection mechanism used by the IEA-TIMSS survey. In a first stage, a fixed number of schools from each region is selected at random. In a second stage, the external monitors are assigned to either one or two classrooms selected at random by INVALSI, depending on the total number of classrooms in the school.[6] Students in these classes constitute the treatment group.

Teachers, unlike external monitors, may have incentives to manipulate test scores. Even though the exams are low stakes for both students and teachers, they may perceive that

[5]Some external monitors are retired teachers, while others are *precari*, *i.e.* teachers with no tenured position. They are paid between 100 and 200 EUR for the job, and can be asked in the subsequent years to monitor more exams, giving them incentives to grade fairly.

[6]The 2013 tests were the first in which INVALSI did the assignment by public procedure. Previously, it was done internally by the selected schools. The recent changes in the assignment of external monitors have allowed reducing the number of classrooms with an external monitor.

they are evaluated based on the test results. To understand this, notice that INVALSI sends the results to principals, who can make them public to entice parents to enroll their children in their school. Furthermore, anecdotal evidence suggests that the results are often discussed in front of all teachers, which could have an effect on them, such as the assignment of troublesome students. This, coupled with the possibility that principals might be able to pay teachers based on their students' performance in the future, may give them incentives to manipulate test scores.[7]

2.1 Data and Descriptive Statistics

As shown in Table 1, over 2.3 million students were tested during the academic year 2012-13. Over 143,000 of them were assigned an external monitor, and the mean number of correct answers of those students was smaller than for those whose monitor was internal. This difference was larger in mathematics exams, but there was a lot of variability across grades.

Table 1: Size of the groups, academic year 2012-13

	2nd grade		5th grade		6th grade		8th grade		10th grade	
	EX	IN	EX	IN	EX	IN	EX	IN	EX	IN
N	25070	437479	24773	424046	27504	410332	28153	360528	38273	270262
C	1424	25346	1426	25559	1457	21756	1464	19041	2203	15339
S	737	6451	736	6422	732	5143	1416	4537	1094	3276
% Correct	53.87	61.20	54.79	59.52	44.53	45.25	50.83	52.48	42.09	45.13
(Math)	(20.68)	(21.58)	(18.87)	(19.25)	(16.80)	(16.70)	(18.98)	(19.02)	(17.72)	(18.39)
% Correct	59.90	64.76	74.36	76.82	64.25	64.40	72.44	73.12	64.20	65.92
(Ita)	(17.39)	(17.84)	(16.12)	(15.52)	(16.74)	(16.87)	(14.96)	(14.78)	(16.20)	(17.00)

Notes: N, C and S respectively denote the number of students, classrooms and schools, and EX and IN respectively denote the groups with the external and the internal monitor. Classes with an internal monitor in schools that had at least one class with an external monitor are excluded. Standard deviations in parentheses.

Table 2 shows the mean and standard deviation of the covariates I use in this paper. As in previous years, some of the variables were not perfectly balanced across the two groups. In particular, the mean class size was slightly larger in classrooms with an external monitor in the majority of the exams, which also had a slightly higher presence of male and

[7]Two hundred million euros have been assigned to principals to distribute among their teachers. The criteria to distribute this money includes teaching quality, which could be measured by the results of the INVALSI tests. See `https://labuonascuola.gov.it/documenti/LA_BUONA_SCUOLA_SINTESI_SCHEDE.pdf?v=0b45ec8`.

immigrant students in the upper secondary tests. Also, the geographic stratification led to an over-representation of students from regions in which test scores were more manipulated in previous years (Bertoni et al., 2013).[8]

Table 2: Mean and standard deviation of covariates

	2nd grade		5th grade		6th grade		8th grade		10th grade	
	EX	IN	EX	IN	EX	IN	EX	IN	EX	IN
Class size	17.61*	17.26	17.37*	16.59	18.88	18.86	19.23*	18.93	17.37	17.62
	(4.69)	(5.22)	(4.75)	(5.04)	(4.20)	(4.54)	(4.49)	(4.54)	(5.49)	(5.97)
Male	0.51	0.51	0.50	0.50	0.51	0.51	0.51	0.50	0.51*	0.49
	(0.50)	(0.50)	(0.50)	(0.50)	(0.50)	(0.50)	(0.50)	(0.50)	(0.50)	(0.50)
Native	0.95	0.95	0.94	0.94	0.92	0.92	0.91*	0.92	0.90*	0.91
	(0.21)	(0.21)	(0.24)	(0.24)	(0.26)	(0.26)	(0.28)	(0.27)	(0.29)	(0.28)
North	0.39*	0.46	0.38*	0.44	0.43*	0.45	0.41*	0.43	0.41*	0.45
	(0.49)	(0.50)	(0.49)	(0.50)	(0.49)	(0.50)	(0.49)	(0.50)	(0.49)	(0.50)
Center	0.19*	0.18	0.19*	0.18	0.19*	0.17	0.20*	0.18	0.18*	0.16
	(0.40)	(0.39)	(0.39)	(0.38)	(0.39)	(0.38)	(0.40)	(0.38)	(0.39)	(0.37)
South & Isles	0.42*	0.36	0.43*	0.38	0.38	0.39	0.39	0.39	0.40*	0.38
	(0.49)	(0.48)	(0.50)	(0.48)	(0.49)	(0.49)	(0.49)	(0.49)	(0.49)	(0.49)

Notes: EX and IN respectively denote the groups with the external and the internal monitor. Standard deviations in parentheses. The asterisk denotes that difference between the two groups is significantly different from zero at the 95% confidence level.

For expositional brevity, I focus the analysis on the 10th graders' mathematics exam: 10th graders' constitute the largest treatment group, and the difference in the percentage of correct answers between the two groups is larger in the mathematics exam.[9] In all, 38,273 students in 2,203 classes were assigned an external monitor, whereas 270,262 students in 15,339 classrooms were assigned an internal monitor in schools without external monitors.[10]

There were 50 questions in the 10th graders' mathematics test. The left graph in Figure 1 shows the proportion of students who answered each question correctly. Students proctored by external monitors had lower scores in all but three of the questions. The difference between the two groups is slightly larger in difficult questions, *i.e.* questions in which the proportion

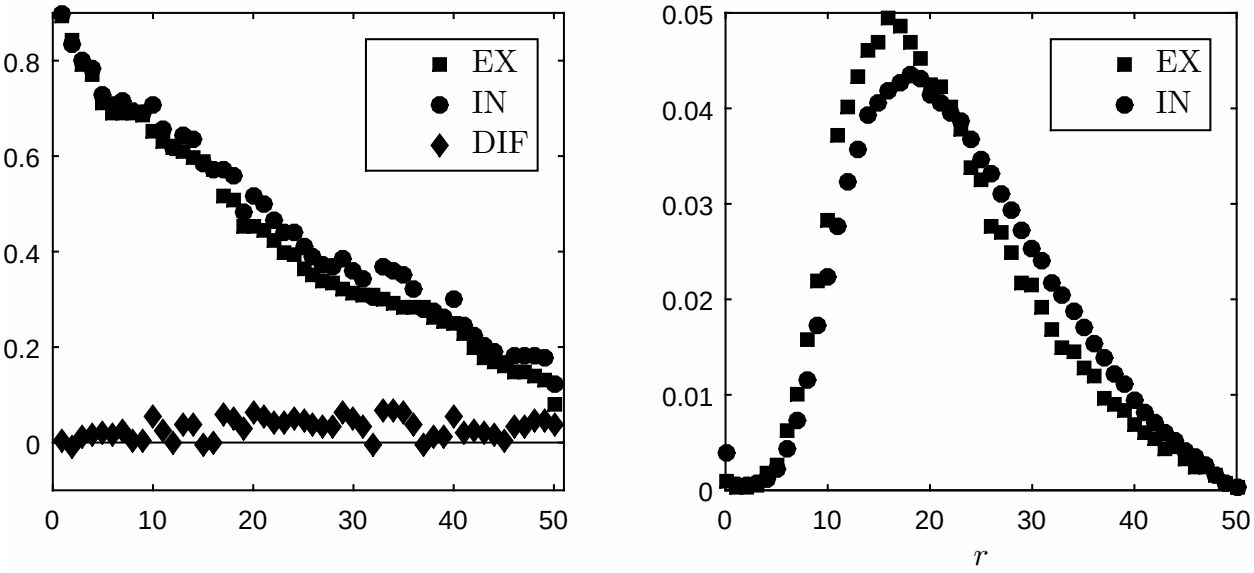
[8]Italy is split into three macro regions: North (Emilia Romagna, Friuli-Venezia Giulia, Liguria, Lombardia, Piemonte, Trentino-Alto Adige, Valle d'Aosta, and Veneto), Center (Lazio, Marche, Toscana, and Umbria), and South and Islands (Abruzzo, Basilicata, Calabria, Campania, Molise, Puglia, Sardegna, and Sicilia).

[9]Because the amount of manipulation substantially varied by exam, pooling all the exams together to detect cheating patterns would be counter-productive, because it would cover several manipulation patterns. Regardless, the regressions for all exams are shown in Appendix S1, and the results that are different across exams are also highlighted in the paper.

[10]Since Bertoni et al. (2013) found that the manipulation was less severe in non-treated classrooms in treated schools, I exclude them from the main analysis.

of correct answers for the treatment group is small, although in some exams there is no correlation between the amount of manipulation and difficulty. Similarly, the distribution of the total number of correct answers is different for both groups (right graph), and the mean, the median, and the mode are smaller for the group with the external monitor. The difference is larger around the center of the distribution, and it is much smaller at the tails. Since this is a low stakes exam, there are no jumps at a cut-off grade and the change is quite smooth.

Figure 1: 10th grade mathematics exam results



The left graph depicts the proportion of correct answers by question (questions are sorted by how frequently they were correctly answered by students proctored by an external monitor); the right graph depicts the students' distribution of test scores. EX, IN, DIF, and r respectively denote the groups with the external and the internal monitor, the difference between them, and the number of correct answers.

Some correction methods use the correlation in the answers to identify manipulation of the scores (Jacob and Levitt, 2003; Quintano et al., 2009). However, if the two groups have different mean test scores, the correlation in the answers will be different by construction, even in the absence of manipulation.[11] This poses a comparability problem that requires appropriately controlling for the mean test scores.

To illustrate this point, consider an alternative statistic to the within-class correlation of

[11]For example, if every student in a class got the maximum grade, then the correlation in the answers would be one. On the other hand, if every student answered one half of the answers correctly, the correlation could be equal to one, but also equal to zero.

the answers: the mean number of correct answers in common between two students, denoted by $\mathbb{E}(s)$, where s is the correct number of answers in common. This mean depends on the distribution of the number of correct answers, which is different for students in the treatment and the control groups. Formally, if the first student got $\overline{r}$ questions correct, and the second student got $\underline{r}$ questions correct, $\mathbb{E}(s) = \sum_{\underline{r}=0}^{Q} \sum_{\overline{r}=0}^{Q} \mathbb{E}(s|\overline{r}, \underline{r}) \mathbb{P}(\underline{r}, \overline{r})$. Because $\mathbb{P}(\underline{r}, \overline{r})$ differs for each group, the mean number of correct answers in common is not comparable between the two groups. Instead, define the following counterfactual conditional mean:

$$\mathbb{E}_{j,h}^{cf}(s) \equiv \sum_{\underline{r}=0}^{Q} \sum_{\overline{r}=0}^{Q} \mathbb{E}_j(s|\overline{r}, \underline{r}) \mathbb{P}_h(\overline{r}) \mathbb{P}_h(\underline{r}) \tag{1}$$

where $j, h = \{EX, IN\}$ and $\mathbb{P}_h(r)$ is the unconditional distribution of the total number of correct answers for students in group h. The first term of Equation 1 captures the raw difference in correlation between the two groups, regardless of the distribution of correct answers in the overall population. When $j = h$, the second and third terms differ for each group, preventing a fair comparison, but if both groups use the same probability weights, then it is possible to assess the effect of cheating on the correlation in answers.

Table 3 shows the counterfactual values for each exam in percentage terms (to make them comparable across exams). When each group uses its own distribution (top two rows), the difference between the two groups in the 10th graders' mathematics exam equals 3% of the total number of answers. However, the difference shrinks to 0.3% if one uses the overall distribution (third and fourth rows). This result holds in all exams, and the largest reduction is attained in the 2nd graders' mathematics exam, in which the original difference of 8% completely vanishes after controlling for the mean test scores.

However, the counterfactuals show the existence of some excess within-class correlation. If one uses the conditional mean for students in different classrooms, observe that the percentage of correct answers in common is smaller than for students in the same classroom. Hence, the correlation in students' answers mostly reflects factors other than manipulation, such as teacher effects (Hanushek, 1971).

Table 3: Percentage of correct answers in common

		2nd grade		5th grade		6th grade		8th grade		10th grade	
		M	I	M	I	M	I	M	I	M	I
DIF	EX	32.3	38.2	33.7	57.3	24.0	45.1	28.6	55.4	22.9	44.4
DIST	IN	40.7	44.2	39.3	60.9	24.7	45.2	30.5	56.4	25.9	46.7
SAME	EX	40.2	43.8	38.7	60.7	24.6	45.3	30.2	56.3	25.3	46.1
DIST	IN	40.2	43.9	38.9	60.7	24.6	45.2	30.4	56.3	25.6	46.4
	IND	39.3	43.2	37.8	60.2	24.0	44.6	29.0	55.6	24.3	45.5

Notes: EX, IN and IND respectively denote the mean number of correct answers in common of two students with $\overline{r}$ and $\underline{r}$ correct answers (Equation 1) when they are in the same class and the examiner is external, in the same class and the examiner is internal, or in different classes in either group. DIF DIST and SAME DIST respectively denote that the weighting function $P_h(r)$ was each group's own distribution, or the overall distribution. I and M respectively denote the Italian and mathematics exams.

3 Empirical Methodology

Let y_{icq} equal one if student i in classroom c answered question q correctly, and zero otherwise. This variable can be modeled with a latent variable, y^*_{icq}, that depends on three effects: a student-class effect, η_{ic}, a question effect, ξ_q, and a specific student-class-question *iid* shock, ε_{icq}. The student-class effects measures the ability of a student, whereas the question effects measure the difficulty of each particular question. [12] Formally,

$$y_{icq} = \mathbf{1}\left(y^*_{icq} \geq 0\right) \tag{2}$$

$$y^*_{icq} = x'_{ic}\beta + \eta_{ic} + \xi_q + \varepsilon_{icq} \tag{3}$$

where, from an econometric standpoint, the number of questions (Q) is fixed, the number of classrooms (C) is large, and the number of students per classroom (N_c) is small but not fixed. Because of the incidental parameter problem, it is impossible to obtain consistent estimates of the student-class effects, but it is possible to consistently estimate the question effects.[13] The latter are parameters in the regression, while the former are treated as random effects. Denote by $y_c \equiv (y_{1c1}, ..., y_{1cQ}, ..., y_{N_ccQ})$ the vector with the results of all students in classroom c. Assume that the distribution of the unobservables is given by $\varepsilon_{icq} \sim Logistic(0, 1)$ and

[12] The question effect may also capture the location of the question in the exam. There were several versions of each exam, the only difference among them being the order of the questions. Unfortunately, the version assigned to each student is not recorded in the dataset.

[13] The setup is similar to those considered in Item Response Theory: they model the result to each question using an individual latent trait that is constant across questions, and questions are allowed to vary in difficulty. See Bacci et al. (2014) for an example applied to the INVALSI tests.

$\eta_{ic} \sim \mathcal{N}\left(0, \sigma_\eta^2\right)$, and let $\theta \equiv \left(\xi', \sigma_\eta^2\right)'$. If the student-class effects were independent of each other, the system 2-3 would be a random effects logit with normally distributed random effects. Its likelihood is given by

$$\mathcal{L}(\theta) = \sum_{c=1}^{C} \sum_{i=1}^{N_c} \log \left(\int_{\mathbb{R}} \frac{\exp\left(\sum_{q=1}^{Q} y_{icq} \left(x_{ic}'\beta + \eta_{ic} + \xi_q\right)\right)}{\prod_{q=1}^{Q} \left(1 + \exp\left(x_{ic}'\beta + \eta_{ic} + \xi_q\right)\right)} d\Phi\left(\frac{\eta_{ic}}{\sigma_\eta}\right) \right) \tag{4}$$

The evidence found in Section 2.1 does not support the independence of the student-class effect. Still, Equation 4 can be used to consistently estimate the vector of parameters θ, although not efficiently (Pereda-Fernández, 2017). However, it is necessary to estimate the correlation of the student-class effects to consistently estimate joint events. A convenient way to model the correlation of the student-class effects is using a copula, *i.e.* a multivariate function that captures the correlation structure of a vector of random variables.[14] Copulas depend on the ranks of the individual effects, $u_{ic} \equiv \Phi\left(\frac{\eta_{ic}}{\sigma_\eta}\right)$, which are invariant to the parameters of the marginal distribution of η_{ic}.[15]

Denote by η_c and u_c the N_c-dimensional vectors of the individual effects and their ranks in class c. I model their correlation with a Clayton copula, denoted by $C(u_c; \rho)$, where ρ is the parameter that models the correlation intensity. The Copula-Based Random Effects (CBRE, Pereda-Fernández, 2017) estimator maximizes the following likelihood function:

$$\mathcal{L}(\theta) = \sum_{c=1}^{C} \log \left(\int_{[0,1]^{N_c}} \frac{\exp\left(\sum_{i=1}^{N_c} \sum_{q=1}^{Q} y_{icq} \left(x_{ic}'\beta + \eta_{ic} + \xi_q\right)\right)}{\prod_{i=1}^{N_c} \prod_{q=1}^{Q} \left(1 + \exp\left(x_{ic}'\beta + \eta_{ic} + \xi_q\right)\right)} dC(u_c; \rho) \right) \tag{5}$$

Remark 1: In principle, it would be possible to use a different copula and select the one that has the best fit. However, simulation results in Pereda-Fernández (2017) suggest that the largest improvement in fit comes from using a copula, and estimated probabilities are roughly the same regardless of the exact parametric copula. Among these, the Clayton copula is convenient from a computational standpoint.

Remark 2: The covariates used in this paper have finite support. Hence, neither the marginal distribution nor the copula of student-class effects is nonparametrically identified

[14] As proved by Sklar (1959), any multivariate cdf can be written as a copula the arguments of which are the marginal distributions, *i.e.* $\mathbb{P}(X_1 \leq x_1, ..., X_d \leq x_d) = C(F_1(x_1), ..., F_{N_c}(x_{N_c}))$. See Nelsen (2006) for an introduction to copulas.

[15] Equation 4 implicitly assumes that the copula of the individual effects for students in the same classroom is independent, *i.e.* $C\left(\Phi\left(\frac{\eta_{1c}}{\sigma_\eta}\right), ..., \Phi\left(\frac{\eta_{N_c c}}{\sigma_\eta}\right)\right) = \prod_{j=1}^{N_c} \Phi\left(\frac{\eta_{ic}}{\sigma_\eta}\right)$.

(Chernozhukov et al., 2013; Pereda-Fernández, 2017). There exist estimators that do not impose distributional assumptions, but because they do not estimate the distribution of these effects (Chamberlain, 1980; Manski, 1975), they cannot be used to estimate joint probabilities.

Remark 3: I am ruling out class-question effects, which would matter if some teachers were better able to teach the material relevant for some of the questions. A way to avoid this issue would be to run the regression using a single student per classroom. This analysis is reported in Appendix S3, and the results indicate that there is no substantial bias by not considering them.

Remark 4: Equations 4 and 5 denote the likelihood conditional on x_{ic}. Using the appropriate covariates in the estimation can help improve the detection of manipulated test scores, whereas adding redundant ones increases the risk of overfitting the model and hinders the detection of manipulated test scores. Hence, the selection of the covariates to use in the estimation is important to limit the number of misclassified tests. To address this point, the covariates were chosen using k-fold cross-validated forward selection. This method selects variables iteratively, and by using cross-validation, only those variables that improve the out-of-sample fit are selected. Alternative methods exist to select covariates (see, e.g., Friedman et al., 2001), but given the large number of observations and limited number of covariates, this method has a relatively limited cost in terms of computational time.[16] The exact algorithm is described in Appendix B.

3.1 Cheating Correction

Using the estimates of the treatment group, it is possible to compute the likelihood of observing the results of a single classroom. In the absence of manipulation, very high test scores would be infrequent, but with manipulation, there would be an excessive number of them. The correction I propose is based on this idea, and it consists of two steps: first,

[16]Subset selection methods such as forward selection are relatively fast to implement when the number of covariates is small relative to the number of observations. For the case in which the number of covariates is relatively large, alternative methods such as LASSO are faster to compute.

the distribution of the likelihood of the test scores of the group with the internal monitor is shifted to match the distribution of the group with the external monitor; and second, I compute the expected mean fair score conditional on the likelihood and the observed test score. The correction equals the difference between the actual score and the expectation.

Formally, let the students' total number of correct answers in classroom c be given by $r_c \equiv (r_{1c}, ..., r_{N_c c})$. The probability of getting at least r_c correct answers at random, denoted by $\mathbb{P}(R \geq r_c)$, is computed using the CBRE estimates for the group with the external monitor:[17]

$$\hat{l}_c = \left[\sum_{b_1 \in \overline{B}_{r_{1c}}} ... \sum_{b_{N_c c} \in \overline{B}_{r_{N_c}}} \int_{[0,1]^{N_c}} \frac{\exp\left(\sum_{i=1}^{N_c} \sum_{q=1}^{Q} b_{iq} \left(x'_{ic}\hat{\beta} + \eta_{ic} + \hat{\xi}_q\right)\right)}{\prod_{i=1}^{N_c} \prod_{q=1}^{Q} \left(1 + \exp\left(x'_{ic}\hat{\beta} + \eta_{ic} + \hat{\xi}_q\right)\right)} dC(u_c; \hat{\rho}) \right]^{\frac{1}{N_c}} \quad (6)$$

where $\overline{B}_{r_{ic}} \equiv \left\{ b_{iq} : \sum_{q=1}^{Q} b_{iq} \geq r_{ic} \right\}$, *i.e.* all the possible combinations of correct answers that would yield a test score of at least r_{ic}. This probability takes into account that some results are less likely to occur in the absence of manipulation, even if the within-class mean test score is the same. For example, assume female students perform worse than male students in mathematics in the presence of an external monitor. Then, for two classrooms with identical test scores, if in the first one female students have higher scores, and in the second one they have lower scores, the probability of observing the result in the first classroom would be smaller than in the second one. Similarly, if students' answers displayed too much or too little correlation relative to what would be expected, the probability of observing those results would be small.

For the first step, denote by $F_{\mathcal{L},EX}(l)$ and $F_{\mathcal{L},IN}(l)$ the cdf of the likelihood for the treatment and control groups, respectively. The corrected likelihood is given by $\check{l}_c \equiv F^{-1}_{\mathcal{L},EX}\left(F_{\mathcal{L},IN}\left(\hat{l}_c\right)\right)$. By construction, the cdf of the corrected likelihood of the classes with an internal monitor equals the cdf of the classes with an external monitor. For the second step, I use the following assumption.

Assumption 1. *Distribution of test scores manipulation*

Let r^*_c denote the observed mean test score of classroom c with an internal monitor. This

[17] Since this probability is mechanically different depending on the class size, to make the comparison fair, I compute the geometric mean of this probability. See Appendix A for the details on the computation of the sum of all possible permutations.

score is decomposed into the sum of the score without manipulation, r_c, and the manipulation, α_c. These two components are mutually independent, and the distribution of the manipulation is given by an exponential(λ) distribution.

With this assumption, it is possible to estimate the corrected test score, which equals the expected fair test score, conditional on the observed test score and the corrected likelihood, *i.e.* $\mathbb{E}\left[r|r^*, \check{l}\right]$.[18] The idea is similar to Wei and Carroll (2009), whose estimator of quantile regression with measurement error is adapted to the current framework:

$$\mathbb{E}\left[r|r^*, \check{l}\right] = \frac{\int_0^{r^*} r f\left(\check{l}|r\right) \lambda \exp\left(-\lambda\left(r^* - r\right)\right) dF\left(r\right)}{\int_0^{r^*} f\left(\check{l}|r\right) \lambda \exp\left(-\lambda\left(r^* - r\right)\right) dF\left(r\right)} \tag{7}$$

where the equality follows by Bayes' theorem. Equation 7 suggests the following sample analogue to estimate the corrected test scores:

$$\tilde{r} \equiv \frac{\frac{1}{\sum_{c=1}^{C_0} \mathbf{1}(r_c \leq r^*)} \sum_{c=1}^{C_0} r_c \hat{f}\left(\check{l}|r_c\right) \hat{\lambda} \exp\left(-\hat{\lambda}\left(r^* - r_c\right)\right)}{\frac{1}{\sum_{c=1}^{C_0} \mathbf{1}(r_c \leq r^*)} \sum_{c=1}^{C_0} \hat{f}\left(\check{l}|r_c\right) \hat{\lambda} \exp\left(-\hat{\lambda}\left(r^* - r_c\right)\right)} \tag{8}$$

where $\hat{f}\left(\check{l}|r\right) = \sum_{k=1}^{K} \frac{\tau_{k+1} - \tau_k}{\hat{Q}_L(\tau_k|r) - \hat{Q}_L(\tau_{k+1}|r)} \mathbf{1}\left(\hat{Q}_L\left(\tau_k|r\right) < \check{l} \leq \hat{Q}_L\left(\tau_{k+1}|r\right)\right)$, $\hat{\lambda}$ is estimated using the method of moments, and $\hat{Q}_L\left(\tau|r\right)$ is estimated using linear quantile regression on a polynomial of r and applying Chernozhukov et al. (2010) rearrangement.

Assumption 1 is not likely to hold in practice if manipulation of the test scores has a strategic component and if some test scores in the control group are not manipulated. If this assumption was relaxed, one would still need to make an assumption on the distribution of the amount of manipulation conditional on the fair test score. Nevertheless, it allows the correction to be expressed in closed form with the desirable property that the smaller the likelihood of the results and the higher the mean test scores, the higher the correction.

This approach efficiently uses the information from the answers to every question, including its difficulty. However, as with any other correction measure, it is subject to misclassification. Relative to existing alternatives, this approach acknowledges the existence of effective teachers and students in classrooms with internal monitors, reducing the correction applied to high-achieving students. Moreover, students' characteristics are only used to calculate $\check{l}$, so if two students in

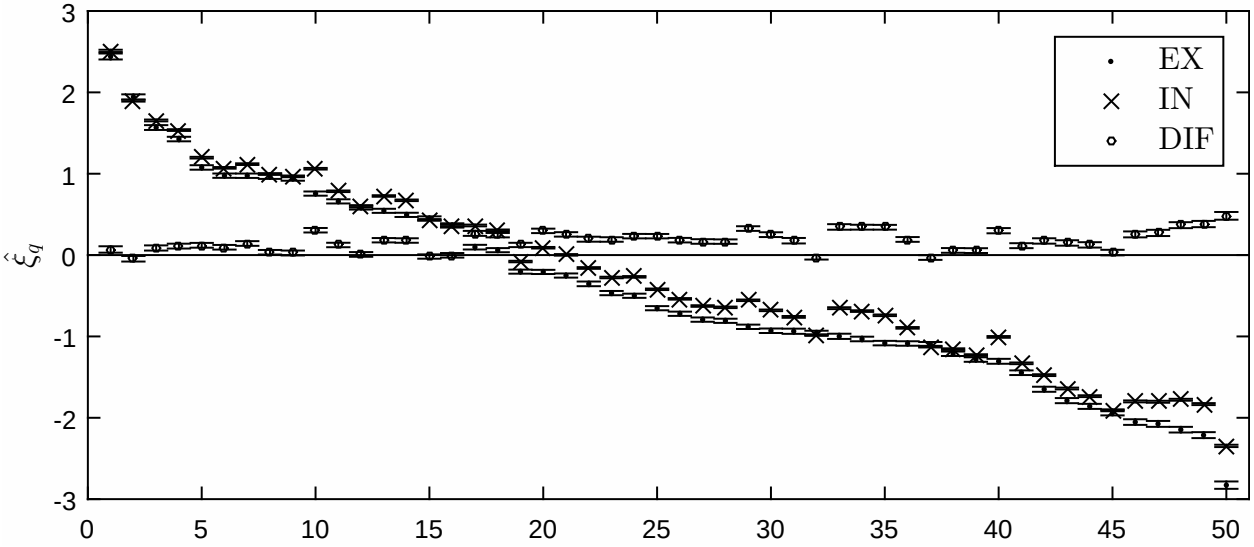
[18] Using a parametric distribution with positive support, such as the exponential distribution, ensures that the correction does not result in an increase of the test scores.

the same classroom got the same scores, they would be corrected by the same amount. Still, this correction could have some unintended side effects. As far as I know, the only paper that studies the effects of correcting test scores on students' psychological well-being is Lucifora and Tonello (2016), who considered the effects of Quintano et al. (2009) correction, but did not find any evidence that it hurt students' psychological well-being.

4 Results

Figure 2 shows the RE logit estimates (Equation 4) of the question effects without covariates. The results show that for 38 out of the 50 questions, the coefficient for the control group is significantly larger than for the treatment group, and for half of the remaining 12, they are not significantly different. Ignoring unobserved heterogeneity results in significantly biased estimates, as Table 10 in Appendix S1 shows.

Figure 2: RE logit estimates, 10th grade mathematics exam



RE logit estimates of the question effects (ξ_q in Equation 4) for the group with an external monitor (EX), the group with an internal monitor (IN), and the difference between them (DIF). They are reported along with the 95% confidence intervals, and sorted by how frequently they were correctly answered by students proctored by an external monitor.

Table 4 shows the Average Partial Effects (APE) of the different covariates and the estimates of the parameters of the distribution of individual effects: (σ_η, ρ). The covariates

and interactions were selected as reported in Section 3.[19] The CBRE coefficients for the group with an external monitor (column 4) show that students from the Center and South & Islands regions respectively scored on average 7% and 14% less correct answers than students from the North. Female students also scored lower than their male counterparts (around 6% less), whereas native Italians outperformed immigrant students (around 4.5% more). The number of classrooms in the school played a minor role, and an extra classroom is correlated with an increase of 0.05% of correct answers. Similarly, students in small classrooms scored worse than those in large classrooms, and increasing class size by one student is correlated with an increase of less than 1.5% correct answers.[20] The estimate of the copula correlation parameter indicates substantial within-classroom correlation in the unobserved individual-class effect.[21]

The amount of manipulation and how much it benefited each demographic group can be measured by looking at the difference between the coefficients of the two groups (column 6). Students from the Center and South & Islands had an average of almost 7% extra correct answers than their northern counterparts. Similarly, female and immigrant students' answers were more manipulated. For each group, 1.7% and 0.3% extra correct answers can be attributed to manipulation, respectively. The manipulation was also slightly larger in schools with more classrooms and in smaller classrooms. Adding an extra classroom or an extra student per classroom increased the amount of manipulation by about 0.2% and 0.4%, respectively. Finally, the copula correlation coefficient was smaller in the control group, indicating that the correlation in the unobserved effects did not increase because of the manipulation.

Many of the results apply to most exams in the sample. Table 5 summarizes the differences in performance between students with an internal and an external monitor for

[19]The APE reflect the overall effect of increasing each variable, which affects also the interactions between that variables and other selected variables. Consequently, even though the number of chosen terms in the specification is large, Table 4 reports the APE with respect to the main variables, the interpretation of which is more transparent.

[20]Small classrooms are defined as those whose size is smaller than the median class size.

[21]ρ is not interpreted as the linear correlation coefficient. Using the relation between the Clayton and Gaussian copulas with Kendall's τ statistic, the linear correlation for this group is approximately 0.77.

Table 4: RE & CBRE logit estimates, 10th grade mathematics exam

	RE			CBRE		
	External	Internal	Difference	External	Internal	Difference
	(1)	(2)	(3)	(4)	(5)	(6)
$\hat{\bar{\xi}}$	-6.73	-7.88	1.15	-4.08	-4.42	0.34
	(1.86)	(1.39)	(2.32)	(1.02)	(1.07)	(1.48)
FEMALE	-6.08	-5.62	-0.46	-5.77	-4.11	-1.66
	(0.18)	(0.07)	(0.20)	(0.06)	(0.03)	(0.07)
CENTER	-7.72	-0.69	-7.02	-7.34	-0.63	-6.71
	(0.40)	(0.31)	(0.51)	(0.22)	(0.27)	(0.35)
SOUTH & ISLANDS	-14.24	-4.55	-9.70	-13.69	-7.03	-6.66
	(0.37)	(0.46)	(0.59)	(0.22)	(0.34)	(0.41)
ITALIAN STUDENT	5.17	6.37	-1.20	4.51	4.16	0.35
	(0.25)	(0.11)	(0.28)	(0.09)	(0.05)	(0.10)
NUMBER OF CLASSES	0.19	0.24	-0.05	0.05	0.26	-0.21
	(0.03)	(0.01)	(0.03)	(0.02)	(0.00)	(0.02)
CLASS SIZE	1.02	0.97	0.06	1.38	1.01	0.37
	(0.02)	(0.01)	(0.02)	(0.01)	(0.00)	(0.01)
$\hat{\sigma}_\eta$	0.85	0.92	-0.08	0.75	0.88	-0.14
	(0.01)	(0.00)	(0.01)	(0.00)	(0.00)	(0.00)
$\hat{\rho}$	-	-	-	2.55	1.56	1.00
				(0.02)	(0.00)	(0.02)

Notes: $\hat{\bar{\xi}}$ denotes the APE of the question effects, $\hat{\sigma}_\eta$ denotes the standard deviation of the individual effects distribution, and $\hat{\rho}$ the parameter of its Clayton copula. Columns 1-3 show the APE of the covariates and the estimates of σ_η with the RE logit estimator (Equation 4); columns 4-6 show the same estimates and those of ρ with the CBRE estimator (Equation 5).

all exams, and it represents the percentage of manipulation in favor of each demographic group. The single most important variable is the dummy for the South & Islands region, which is significantly negative in all exams. Thus, it was in this region that the largest amount of manipulation took place, an average of 5.6% extra correct answers. There was also more manipulation in the Center than in the North in most exams, and the difference was significant in seven of them. On average, 3.5% extra correct answers could be attributed to manipulation in this region.

Test scores of female students were more manipulated than their male counterparts, both in mathematics (1% on average) and Italian exams (0.2% on average). This means that manipulation favored female students more in those exams in which male students consistently outperform them. The existence of persistent differences in academic performance by gender is well documented (Machin and Pekkarinen, 2008; Lavy and Sand, 2015). Consistently with the results of this paper, Lavy (2008) found that male students face discrimination with

Table 5: Summary CBRE logit estimates

	2nd grade		5th grade		6th grade		8th grade		10th grade	
	M	I	M	I	M	I	M	I	M	I
FE	-1.2**	-0.4*	-1.1**	-0.3**	-0.4**	0.0	-0.9**	0.0	-1.7**	-0.3**
CE	-2.9**	-0.1	-3.8**	-1.7**	-2.0*	-1.2**	-0.9	0.2**	-6.7**	-15.7**
SI	-8.7**	-3.1**	-7.1**	-3.6**	-3.4**	-1.0**	-3.3**	-0.7**	-6.7**	-19.0**
IT	1.4**	1.8**	0.0	0.7**	0.7**	1.1**	1.2**	1.5**	0.3**	0.1*
NC	0.4**	0.2**	0.2**	0.1**	0.1	0.2**	0.0	0.0	-0.2**	0.0*
CS	0.5**	0.0	0.2**	0.0	0.0	-0.1**	0.2**	0.0**	0.4**	0.2**

Notes: FE, CE, SI, IT, NC and CS refer to the difference between externally and internally monitored students of the APE for females, Center region, South & Islands region, natives, number of classes in the school, and class size, as reported in column 6 from Tables 4 and 11 to 19, expressed in %. I and M respectively denote the Italian and mathematics exams. *, and ** respectively denote statistical significance at the 95, and 99% confidence level.

respect to females in every subject. In contrast, Diamond and Persson (2016) found that teachers' grading leniency was not different for male and female students.

Similarly, test scores of immigrant students were more manipulated in most exams, with an average of 0.9% extra correct answers. This difference was larger in the Italian exams, which could mean that teachers were trying to compensate for the handicap immigrants face by having to learn the local language. These results contrast with Diamond and Persson (2016), who found no discrimination between natives and immigrants, Sprietsma (2013), who found that German teachers discriminate against students with Turkish names, and Hanna and Linden (2012), who found that Indian teachers discriminate against lower caste students.

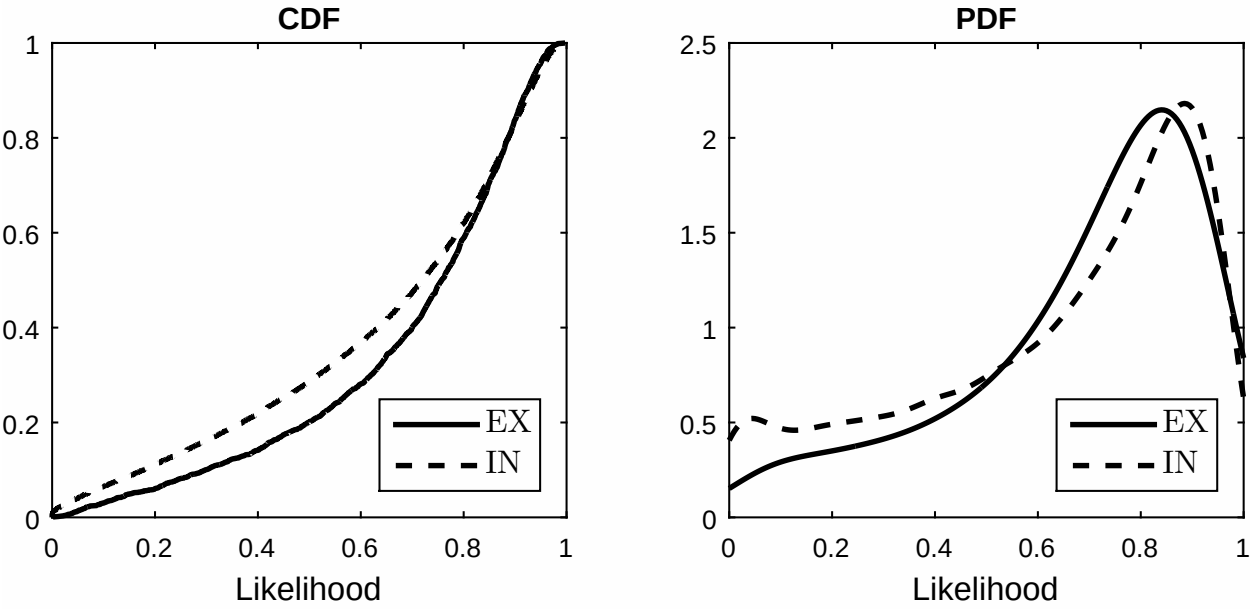
Manipulation was slightly larger in schools with many classrooms, albeit by a small margin, and this difference was significantly positive in six of the exams. Finally, the exams were more manipulated in small classrooms, although this result is not homogeneous, and in the sixth grade Italian exam, the manipulation was larger in large classrooms.

5 Cheating Correction

The distribution of the estimated likelihood from Equation 6, based on the estimates of column (6) from Table 4, is shown in Figure 3. As expected, the two do not coincide: there is approximately the same proportion of classes with likely results, *i.e.* those on the right

tail. However, the left tail of the distribution of the group with internal monitors has more mass probability, which indicates that there is an excessive number of unlikely results relative to the amount there would have been without manipulation.[22]

Figure 3: Distribution of the likelihood, 10th grade mathematics exam



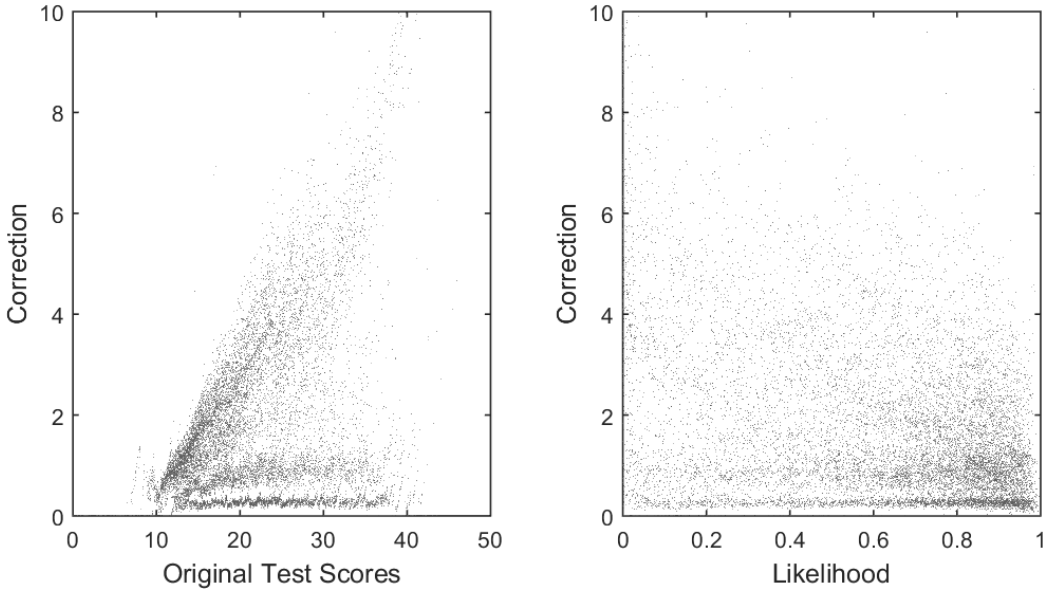
Distribution of the estimated likelihood of the class scores (Equation 6). EX and IN respectively denote the groups with the external and the internal monitor.

Given the large regional differences in test scores manipulation, the correction method proposed in Section 3.1 is applied to each class using only data from that region.[23] Each dot in Figure 4 represents the correction applied to a single class and relates it to their uncorrected test scores and estimated likelihood (Equation 6). The correction is higher for more unlikely, higher test scores. Since the majority of the test scores with unlikely results are located in the South & Islands region, the correction is higher there (Figure 5). Consequently, the class and regional rankings are changed once the correction is applied.

[22]Consistently with the estimation results, the difference between the two distributions is largely explained by the difference in the South & Islands. See Appendix S1.

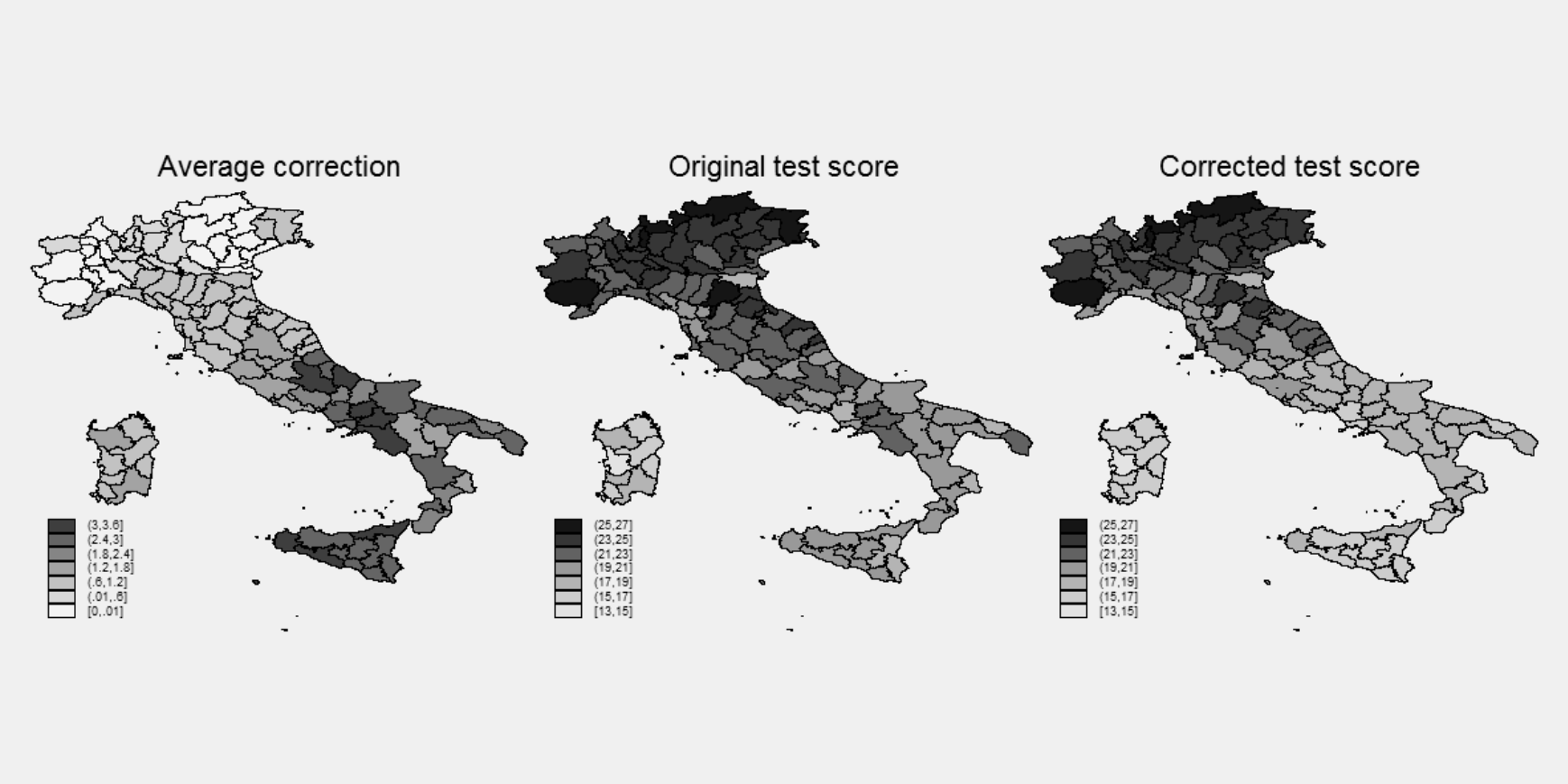
[23]To measure the sensitivity of these estimates, I allow the correction to be applied only to those classes with a difference between the corrected likelihood, $\tilde{\ell}_c$, and its estimated counterpart, $\hat{\ell}_c$, that is greater or equal than a threshold. These results are shown in Figure 10 in Appendix S1.

Figure 4: Correction for cheating, test scores, and likelihood, 10th grade mathematics exam



The left and right graphs respectively show the scatter plot of the mean correction to the classes with an internal monitor, with the class mean test scores and with the estimated likelihood of the test scores of each class (Equation 6).

Figure 5: Correction for cheating, provincial variation, 10th grade mathematics exam



6 Discussion

The results presented thus far do not identify the mechanism behind manipulation. However, INVALSI tests are comprised of two types of questions: multiple choice and open-ended.[24] Multiple choice questions require minimal effort to grade and transcribe, and students would find it easier to copy the answer from one another. Open-ended questions may involve an elaborate answer that takes more time to grade and students may find it harder to copy.

Table 6 shows that open-ended questions were more manipulated than multiple choice. However, the pattern for missing answers was the opposite (Figure 6): for the control group, the proportion of missing answers decreased more for the open-ended questions than for the multiple choice questions. If students had copied each other during the exam, it would have produced the opposite result. Moreover, another reason why students were less likely to be responsible for the manipulation is the fact that there were several versions of each exam with the same questions but in a different order. Hence, this evidence supports the hypothesis that teachers were more responsible than students for the manipulation of the test scores.

Table 6: Multiple choice versus open-ended questions

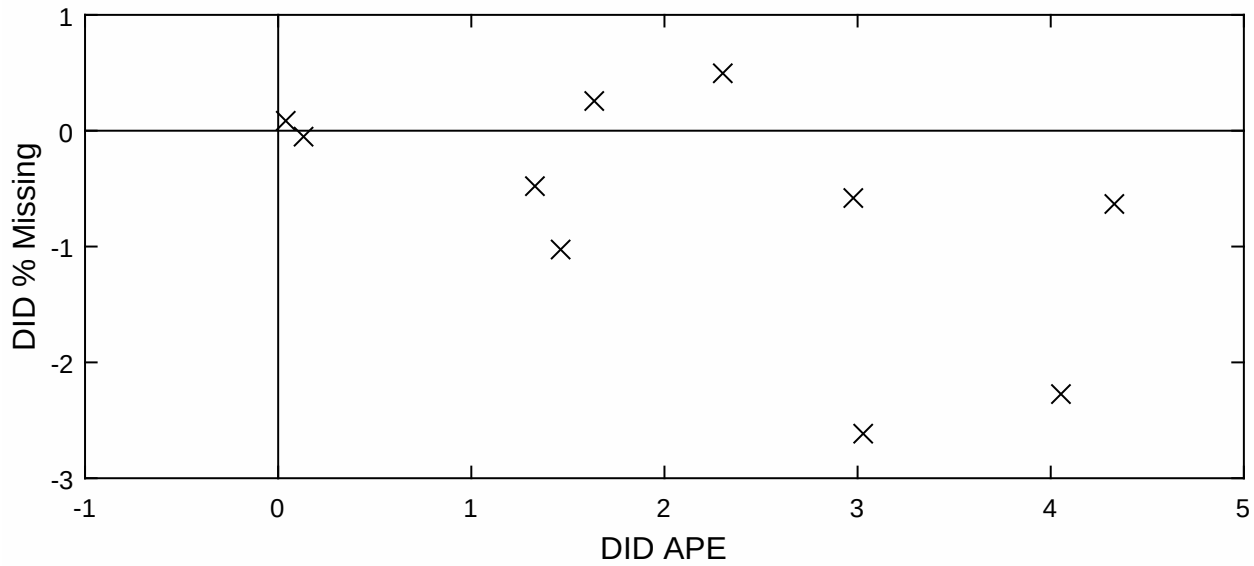
	2nd grade		5th grade		6th grade		8th grade		10th grade	
	M	I	M	I	M	I	M	I	M	I
$\Delta_{Y,MC}$	-6.72	-4.07	-3.72	-2.24	-4.45	0.30	-1.48	-0.83	-2.11	-0.85
$\Delta_{Y,OE}$	-8.19	-8.12	-8.05	-3.57	-5.15	-2.00	-1.61	-0.87	-5.14	-3.83
DID	1.47	4.05	4.33	1.33	0.70	2.30	0.13	0.04	3.03	2.97
$\Delta_{M,MC}$	1.19	1.07	0.22	0.52	-0.11	-0.08	0.24	-0.04	-0.11	-0.29
$\Delta_{M,OE}$	2.21	3.34	0.84	1.00	-0.37	-0.58	0.28	-0.12	2.50	0.29
DID	-1.02	-2.27	-0.63	-0.48	0.26	0.50	-0.05	0.08	-2.61	-0.58

Notes: $\Delta_{APE,MC}$ and $\Delta_{APE,OE}$ respectively denote the mean difference between the treatment and control groups of the mean question APE of the CBRE logit estimates (first row of Table 4) for open ended and multiple choice questions; DID_{APE} denotes the difference between these two; $\Delta_{M,MC}$ and $\Delta_{M,OE}$ respectively denote the mean difference between the treatment and control groups percentage of missing answers for open ended and multiple choice questions; DID_M denotes the difference between these two. All numbers are reported as %.

If teachers graded their own students, it would make sense to increase their test scores to improve the perception of the teachers' ability. However, this is not the case. A possible

[24]The proportion of open-ended questions ranged between 21% and 50%, depending on the exam.

Figure 6: Multiple choice versus open-ended questions



The horizontal axis represents the difference between open ended and multiple choice questions of the difference between the two groups of the mean APE; the vertical axis represents the mean difference between the two types of questions of the difference between the two groups of the proportion of missing questions.

explanation would be grading leniency: even though INVALSI provides a correction grid, open-ended questions may leave more room for interpretation, so some teachers may be less strict in determining when an answer is correct. In contrast, multiple choice questions leave no room for teachers' discretion in grading. However, this does not explain the marked decrease in missing answers when the monitor is internal, nor why the manipulation favored some demographic groups more than others. All of these suggests an active behavior behind the manipulation.

Unfortunately, the dataset has no information on teachers that can be used to uncover the mechanism behind the manipulation. Given that the manipulation was larger in those regions that scored lower in the presence of an external monitor, a plausible conjecture would be that manipulation is a means to prevent linking students' performance to teachers pay by making the results not comparable across schools or regions.

Another important matter is the consequences of cheating on the accumulation of human capital. To access university, Italian students need to pass the final high school state exam (*esame di maturità*). If admission to the university depended only on this exam and it

was subject to similar manipulation patterns, it would create a problem of misallocation of human capital: many students whose actual performance should not warrant access to university would have the opportunity to do so, while others whose test scores were not manipulated might not get this opportunity when they should. To mitigate this problem, many universities have their own entry tests for students who want to enroll in certain degrees. Another potential consequence is the allocation of scholarships during the first year at the university, which in Italy is determined by family income. Finally, some public exams for civil servants take into consideration the final score at the end of secondary education to rank all candidates, so it could have some direct effects on employment.

This raises the question of how should the policy maker use the correction. INVALSI implemented a sanctioning program aimed at discouraging cheating by focusing on schools' reputation. Specifically, two thresholds were pre-specified. If the correction proposed by Quintano et al. (2009) was smaller than both, the results would not be corrected, if it was between them, they would be corrected, and if it was larger than both, the results would not be returned. Lucifora and Tonello (2016) studied the consequences of this program and found a small and not statistically significant effect on cheating in the following year's tests. An alternative use of the correction would be to base the assignment of monitors on the amount of manipulation found in previous years: if schools with higher amount of manipulation were more likely to be assigned external monitors, it would reduce the overall incidence of cheating.

7 Comparison with Previous Studies

Bertoni et al. (2013) found that the presence of an external monitor had a negative impact on test scores of students who were proctored by this monitor and also those proctored by an internal monitor in the same school. Using the estimates from Section 4, I compute the expected scores for the three groups: those with an external monitor, those with an internal monitor in a school that had an external monitor in another class, and those in a school with

only internal monitors.[25] The direct effect can be then computed as the difference between the expected scores of the first and the second groups, whereas the indirect effect is the difference between the second and the third groups.

Table 7: Direct and indirect effect of external monitoring

	2nd grade		5th grade		6th grade		8th grade		10th grade	
	M	I	M	I	M	I	M	I	M	I
INDIRECT	4.6	2.6	3.1	1.8	0.6	-0.1	1.2	0.6	1.9	-0.2
DIRECT	3.0	2.4	1.8	0.7	0.5	0.5	0.6	0.3	1.1	1.5
OVERALL	7.7	5.1	4.8	2.5	1.1	0.4	1.8	1.0	3.0	1.3

The results shown in Table 7 largely support the findings in Bertoni et al. (2013), and the external monitors have an effect both on the test scores of students proctored by them, and the other students in the school. Both effects are important, and only in two Italian exams was the indirect effect negligible. However, in contrast with their findings, the indirect effect dominates in these data: the average direct and indirect effects equal 1.3 and 1.6%, respectively, whereas in Bertoni et al. (2013) they amounted to 2.8 and 0.8%. This difference could be attributed to the change in the assignment of external monitors. Before the 2012-13 academic year, principals had some degree of ability to assign the external monitor to the class they preferred, whereas this choice was made by public procedure afterwards. Hence, if principals assigned the external monitors to low-performing classes, that would bias the estimates of the direct and indirect effects.

The results in this paper are only partially the same as those found in Battistin et al. (2016): as shown in Table 5, manipulation was positively correlated with class size in only six of the exams, and in one of them the opposite was true. Moreover, this effect was stronger in the South & Islands region only for 2nd graders.[26] These differences can be partly explained because they focused on 2nd and 5th graders, which displayed the first and third largest amount of manipulation in small classrooms of all exams. However, the change in the assignment of the external monitor could also explain these differences.

[25] Because the characteristics were not the same across the three groups, I compute the expected value for each student using the three sets of estimates, and average them over the whole sample.

[26] Results available upon request.

The method currently used by INVALSI to correct for cheating is based on the approach proposed by Quintano et al. (2009). It is based on a fuzzy clustering approach that depends on four statistics: within class mean test scores, within class standard deviation of test scores, within class average percentage of missing answers, and within class index of answer homogeneity. Thus, if the mean test scores of a classroom are high relative to those found in the treatment group, which are assumed to be free of manipulation, they are more likely to be classified as manipulated. Similarly, the probability of being classified as manipulated increases if their standard deviation, the average percentage of missing answers, or the index of answer homogeneity is low.[27]

While manipulation can be reflected in those four statistics, this method suffers from two problems: comparability and the existence of confounders. To see the first problem, notice that the distributions of these statistics, and even their support, depend on the number of questions and students in the class. Therefore, the same within class standard deviation conveys different information if the classrooms are of different sizes, or if the tests have a different number of questions.[28]

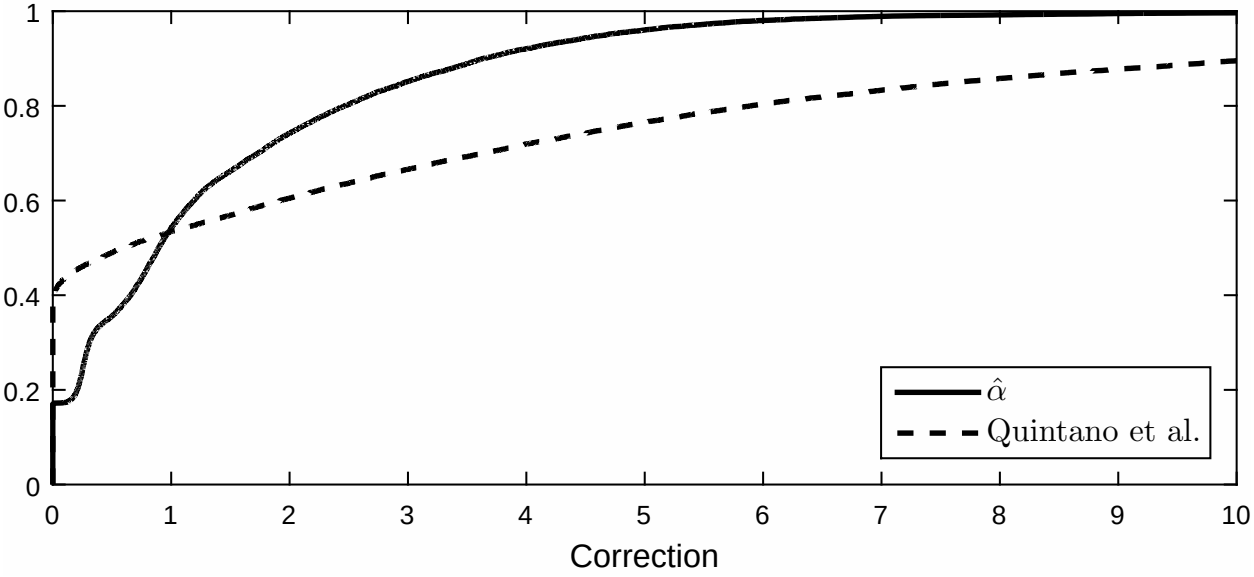
The correction proposed by Quintano et al. (2009) is positively correlated to the one proposed in this paper (the linear correlation coefficient equals 0.52). Figure 7 shows their distribution: the one proposed in this paper only leaves almost 20% of the test scores unchanged, and a correction of less than 3 points (out of a maximum of 50) is applied to nearly 90% of them. On average, the correction equals 1.4 points. In contrast, Quintano et al. (2009) correction does not correct about twice as many test scores, but the average correction for the remaining ones is much larger: more than 10% of the test scores have a correction of at least 10 points, and the average correction equals 4 points.

Finally, Figure 8 compares the mean correction applied in each region by each correction method and relates it to the actual change in mean test scores between the two groups,

[27]The index of answer homogeneity takes a value of zero when every student's answers coincide, and takes higher values, the more heterogeneous they are.

[28]For example, if the number of questions equals 2, and the number of students equals 2, the variance of the test scores can take values $\{0, 1/4, 1\}$, but if the number of students equals 3, then it can take values $\{0, 2/9, 6/9, 8/9\}$.

Figure 7: Distribution of correction for cheating, 10th grade mathematics exam



$\overline{\hat{\alpha}}$ and *Quintano et al.* respectively denote the empirical cdf of the correction methods presented in this paper and the one proposed by Quintano et al. (2009).

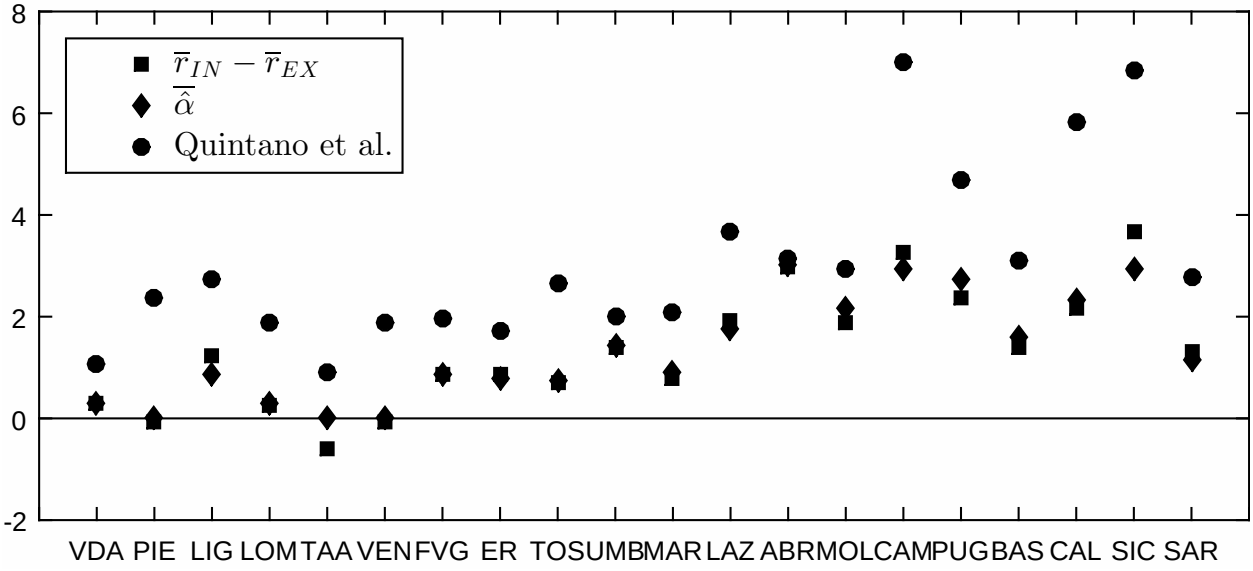
$r_{IN} - r_{EX}$. Both corrections lead to a change in the regional rankings, and regions where the test scores were more manipulated are those in which the correction was the highest. However, they greatly differ in their fit: the correction proposed by Quintano et al. (2009) consistently overestimates the $r_{IN} - r_{EX}$, resulting in a larger reduction of the mean test scores for students with an internal monitor. Conversely, the correction proposed in this paper matches the mean difference between the two groups by region better.

8 Conclusion

In this paper, I propose a novel approach to detect test score manipulation and correct for it, based on the comparison of a group of test scores suspected of having been manipulated with a group of test scores that are assumed to be fair. Taking advantage of a natural experiment in the Italian education system, I apply nonlinear panel data regression methods to describe patterns in test score manipulation, and based on these estimates, I calculate the corrected test scores.

The manipulation was limited in the North of Italy, frequent in the Center and widespread

Figure 8: Correction for cheating, regional variation, 10th grade mathematics exam



For each region, $\overline{r}_{IN} - \overline{r}_{EX}$ denotes the mean difference in test scores between students with an internal and an external monitor, $\overline{\hat{\alpha}}$ denotes the mean correction of the method presented in this paper, and *Quintano et al.* denotes the mean correction of the method proposed by Quintano et al. (2009).

in the South & Islands. Moreover, it tended to favor female and immigrant students. Unobserved heterogeneity accounted for an important share of the total variation, and it exhibited a substantial level of correlation within classrooms, reflecting a combination of teacher effects, sorting of students, and peer effects. These findings are consistent with the conjecture that teachers were responsible for the manipulation. In particular, the difference between open-ended and multiple choice questions between the amount of manipulation and the decrease of missing answers are negatively correlated. However, the exact mechanism behind remains unknown. Future work should investigate why some groups benefit more than others, the geographic patterns and the correlation between missing answers and manipulation.

The correction method I propose allows the results of a classroom to be well or highly correlated because of factors unrelated to manipulation, such as effective teachers or able students. The correction then depends on how likely the observed result are to occur without manipulation, and the higher and more unlikely the results are, the higher the correction. For the majority of the classrooms the correction is quite modest or even zero, and it displays

a large regional variation.

References

Aaronson, D., L. Barrow, and W. Sander (2007). Teachers and student achievement in the chicago public high schools. *Journal of Labor Economics 25*(1), 95–135.

Bacci, S., F. Bartolucci, and M. Gnaldi (2014). A class of multidimensional latent class irt models for ordinal polytomous item responses. *Communications in Statistics-Theory and Methods 43*(4), 787–800.

Battistin, E., J. D. Angrist, and D. Vuri (2016). In a small moment: Class size and moral hazard in the italian mezzogiorno. *American Economic Journal: Applied Economics 9*(4), 216–49.

Battistin, E., M. De Nadai, and D. Vuri (2017). Counting rotten apples: Student achievement and score manipulation in italian elementary schools. *Journal of Econometrics 200*(2), 344–362.

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.

Bonhomme, S. (2012). Functional differencing. *Econometrica 80*(4), 1337–1385.

Chamberlain, G. (1980). Analysis of covariance with qualitative data. *The Review of Economic Studies 47*(1), 225–238.

Chernozhukov, V., I. Fernández-Val, and A. Galichon (2010). Quantile and probability curves without crossing. *Econometrica 78*(3), 1093–1125.

Chernozhukov, V., I. Fernández-Val, J. Hahn, and W. Newey (2013). Average and quantile effects in nonseparable panel models. *Econometrica 81*(2), 535–580.

Chetty, R., J. N. Friedman, and J. E. Rockoff (2014, September). Measuring the impacts of teachers i: Evaluating bias in teacher value-added estimates. *American Economic Review 104*(9), 2633–79.

Cullen, J. B. and R. Reback (2006). *Tinkering toward accolades: School gaming under a performance accountability system*, Volume 14. Emerald Group Publishing Limited.

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 regents examinations. Unpublished working paper.

Diamond, R. and P. Persson (2016). The long-term consequences of teacher discretion in grading of high-stakes tests. Technical report, National Bureau of Economic Research.

Figlio, D. N. (2006). Testing, crime and punishment. *Journal of Public Economics 90*(4), 837–851.

Friedman, J., T. Hastie, and R. Tibshirani (2001). *The elements of statistical learning*, Volume 1. Springer series in statistics New York.

Grissom, J. A., D. Kalogrides, and S. Loeb (2014). Using student test scores to measure principal performance. *Educational Evaluation and Policy Analysis XX*(X), 1–26.

Hanna, R. N. and L. L. Linden (2012). Discrimination in grading. *American Economic Journal: Economic Policy 4*(4), 146–168.

Hanushek, E. (1971). Teacher characteristics and gains in student achievement: Estimation using micro data. *The American Economic Review 61*(2), 280–288.

Hinnerich, B. T., E. Höglin, and M. Johannesson (2011). Are boys discriminated in swedish high schools? *Economics of Education review 30*(4), 682–690.

Hussain, I. (2015). Subjective performance evaluation in the public sector evidence from school inspections. *Journal of Human Resources 50*(1), 189–221.

Jacob, B. A. and S. D. Levitt (2003). Rotten apples: An investigation of the prevalence and predictors of teacher cheating. *The Quarterly Journal of Economics 118*(3), 843–877.

Lavy, V. (2008). Do gender stereotypes reduce girls' or boys' human capital outcomes? evidence from a natural experiment. *Journal of public Economics 92*(10), 2083–2105.

Lavy, V. and E. Sand (2015). On the origins of gender human capital gaps: Short and long term consequences of teachers' stereotypical biases. Technical report, National Bureau of Economic Research.

Levitt, S. D. and M.-J. Lin (2015). Catching cheating students. Technical report, National Bureau of Economic Research.

Lucifora, C. and M. Tonello (2015). Cheating and social interactions: Evidence from a randomized experiment in a national evaluation program. *Journal of Economic Behavior and Organization 115*(C), 45–66.

Lucifora, C. and M. Tonello (2016). Monitoring and sanctioning cheating at school: What works? evidence from a national evaluation program. Technical report, Università Cattolica del Sacro Cuore, Dipartimenti e Istituti di Scienze Economiche (DISCE).

Machin, S. and T. Pekkarinen (2008). Global sex differences in test score variability. *Science 322*(5906), 1331–1332.

Manski, C. F. (1975). Maximum score estimation of the stochastic utility model of choice. *Journal of Econometrics 3*(3), 205–228.

Nelsen, R. B. (2006). *An introduction to copulas*. Springer-Verlag, New York.

Paccagnella, M. and P. Sestito (2014). School cheating and social capital. *Education Economics 22*(4), 367–388.

Pereda-Fernández, S. (2017). Copula-based random effects models for clustered data. Technical report, Bank of Italy Temi di Discussione (Working Paper) No 1092.

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 outliers on assessment test scores. *Statistica Applicata 7*(2), 149–171.

Rockoff, J. E. (2004). The impact of individual teachers on student achievement: Evidence from panel data. *American Economic Review 94*(2), 247–252.

Rothstein, J. (2010). Teacher quality in educational production: Tracking, decay, and student achievement. *The Quarterly Journal of Economics 125*(1), 175–214.

Rothstein, J. (2017). Revisiting the impacts of teachers. *American Economic Review 107*(6), 1656–84.

Sklar, M. (1959). Fonctions de répartition à n dimensions et leurs marges. *Publications de l'Istitut de Statistique de l'Universitè de Paris 8*, 229–231.

Sprietsma, M. (2013). Discrimination in grading: experimental evidence from primary school teachers. *Empirical Economics 45*(1), 523–538.

Wei, Y. and R. J. Carroll (2009). Quantile regression with measurement error. *Journal of the American Statistical Association 104*(487), 1129–1143.

Appendix

A Some linear algebra results

Let z be a vector of dimension T, Z the matrix whose main diagonal are the elements of vector z, and the off diagonal elements all equal zero, ι_T a vector of ones of dimension T, and G be a $T \times T$ matrix whose (i,j) element equals $\mathbf{1}(i<j)$, *i.e.* the elements below the main diagonal equal one, and the remaining elements equal zero. Then, the sum of the permutations of $r \leq T$ distinct elements from z is given by 0 for $r=0$, and $\sum_{k_1=1}^{K-r+1} ... \sum_{k_r=k_{r-1}+1}^{T} \prod_{j=1}^{r} z_{k_j} = \iota_T' (ZG)^{r-1} Z\iota_T$ for $1 \leq r \leq T$. Now consider Equation 5. The probability of observing a particular result, $(b_1, ..., b_{N_c})$, can be written as

$$\mathbb{P}(b) = \int_{[0,1]^{N_c}} \frac{\exp\left(\sum_{i=1}^{N_c}\sum_{q=1}^{Q} b_{iq}(\eta_{ic}+\xi_q)\right)}{\prod_{i=1}^{N_c}\prod_{q=1}^{Q}(1+\exp(\eta_{ic}+\xi_q))} dC(u_c;\rho)$$

To compute $\mathbb{P}(R_1 \geq r_1, ..., R_{N_c} \geq r_{N_c})$, *i.e.* the probability that each student in class c gets at least at many correct answers as they actually got, the preceding trick can be combined with the numerical approximation of the integral with respect to the copula to obtain an estimate of the aforementioned probability, which would be exact if not for the integral. Formally,

$$\begin{aligned}
\mathbb{P}(R_1 \geq r_1, ..., R_{N_c} \geq r_{N_c}) &= \sum_{b_1 \in \overline{B}_{r_1}} ... \sum_{b_{N_c} \in \overline{B}_{r_{N_c}}} \int_{[0,1]^{N_c}} \frac{\exp\left(\sum_{i=1}^{N_c}\sum_{q=1}^{Q} b_{iq}(\eta_{ic}+\xi_q)\right)}{\prod_{i=1}^{N_c}\prod_{q=1}^{Q}(1+\exp(\eta_{ic}+\xi_q))} dC(u_c;\rho) \\
&= \int_{[0,1]^{N_c}} \prod_{i=1}^{N_c} \frac{\sum_{s=r_{ic}}^{Q} \iota_Q' (Z_{ic}G)^{s-1} Z_{ic}\iota_Q}{\prod_{q=1}^{Q}(1+\exp(\eta_{ic}+\xi_q))} dC(u_c;\rho) \\
&\approx \frac{1}{N_1}\sum_{j=1}^{N_1}\prod_{i=1}^{N_c}\left[\frac{1}{N_2}\sum_{h=2}^{N_2} \frac{\sum_{s=r_{ic}}^{Q} \iota_Q' (Z_{icjh}G)^{s-1} Z_{icjh}\iota_Q}{\prod_{q=1}^{Q}(1+\exp(\eta_{jh}+\xi_q))}\right]
\end{aligned}$$

where Z_{ic} and Z_{icjh} are the diagonal matrices whose (q,q) element equal $\exp(\eta_{ic}+\xi_q)$ and $\exp(\eta_{jh}+\xi_q)$, respectively.[29] The approximation in the last row uses the algorithm presented in Pereda-Fernández (2017), which evaluates the integral at a set of points that depend on

[29] Inclusion of covariates is straightforward and is achieved by letting $z_q = \exp\left(\eta + \xi_q + x_{1ic}'\beta + x_{2icq}'\zeta_q\right)$, and substituting the denominator by $\prod_{q=1}^{Q}\exp\left(\eta + \xi_q + x_{1ic}'\beta + x_{2icq}'\zeta_q\right)$.

N_1 and N_2.

B Variable Selection

The following algorithm is used to select the covariates for each exam individually:

1. Estimate the RE estimator with all considered covariates, and select those with a t-statistic of at least 1.96.

2. For the remaining covariates do forward selection with 5-fold cross validation using the following algorithm:

   (a) Split the sample into 5 groups of equal size.

   (b) For each group, compute the estimate using the remaining 4 groups.

   (c) Compute the likelihood of the selected group using this estimate.

   (d) Sum the likelihood from the five groups, obtaining the cross-validated likelihood (CVL).

   (e) Select the variable that increases the CVL the most and repeat until there is no CVL improvement.

3. Repeat steps 1-2 for the control group.

4. The variables selected for either the treatment or the control group conform the vector x_{ict} to be used both for the RE and CBRE estimators.

Standard k-fold cross validation is computationally slow given the large dataset used in this study. Consequently, the proposed algorithm is modified to reduce the required amount of time to select the covariates. In particular, step 1 reduces the number of regressions required in standard forward selection, which begins from the specification without regressors and adds subsequent regressors one at a time. Moreover, in step 3 I randomly select a total number of classrooms in the control group equal to the number of classrooms in the treatment

group. On the other hand, in the regressions presented in the text I interact female dummies with question effects, since the analysis in Appendix S2 suggests that there are non-trivial difference between the two genders across questions. Finally, some variables could be good predictors of students' performance in either the treatment or control group, but not on the other. Pooling both groups together could result in these variables not being selected, especially if their relative sample sizes are markedly different. Thus, step 4 increases the likelihood that this type of variables being selected.

The initial pool of covariates to choose from is the following: regional dummies, female dummy, native Italian dummy, small class size (*i.e.* smaller than the median class size) a quadratic polynomial of class size, number of classes in school, and interactions between regional dummies and class size, and between regional dummies and number of classes in school.

Supplementary Material

S1 Full Results

Table 8: RE logit estimates

	2nd grade		5th grade		6th grade		8th grade		10th grade	
	M	I	M	I	M	I	M	I	M	I
$\hat{\xi}^{EX} > \hat{\xi}^{IN}$	0	0	4	0	29	5	2	9	3	0
$\hat{\xi}^{EX} < \hat{\xi}^{IN}$	32	39	37	82	10	54	36	37	40	88
$\hat{\xi}^{EX} = \hat{\xi}^{IN}$	0	0	6	0	9	12	7	32	7	0

Notes: EX and IN respectively denote the groups with the external and the internal monitor. A coefficient is considered as larger than the other if it is significantly larger at the 95% confidence level, and equal if none is statistically larger than the other.

Table 9: Correlation between RE logit and conditional FE logit estimates

	2nd grade		5th grade		6th grade		8th grade		10th grade	
	M	I	M	I	M	I	M	I	M	I
EX	1.00	0.87	1.00	0.98	1.00	0.97	0.97	1.00	1.00	0.95
IN	0.99	1.00	0.98	0.95	1.00	1.00	1.00	1.00	0.99	1.00
Δ	0.98	0.07	0.33	0.62	0.94	0.38	-0.91	-0.51	0.98	0.03

Notes: EX and IN respectively denote the groups with the external and the internal monitor.

Table 10: Comparison between RE logit and logit estimates

	2nd grade		5th grade		6th grade		8th grade		10th grade	
	M	I	M	I	M	I	M	I	M	I
EX,$\neq$	30	35	40	78	33	70	37	74	43	83
EX,=	2	4	7	4	15	1	8	4	7	5
IN,$\neq$	32	39	46	82	45	71	45	77	47	88
IN,=	0	0	1	0	3	0	0	1	3	0

Notes: EX and IN respectively denote the groups with the external and the internal monitor; = and $\neq$ respectively denote that the coefficients are significantly equal or different at the 95% level of confidence. The quantities represent the number of questions that fit into each category for each exam.

Table 11: RE & CBRE logit estimates, 2nd grade mathematics exam

	RE			CBRE		
	External	Internal	Difference	External	Internal	Difference
	(1)	(2)	(3)	(4)	(5)	(6)
$\hat{\bar{\xi}}$	10.99	19.97	-8.98	9.96	23.32	-13.35
	(2.34)	(0.57)	(2.41)	(2.48)	(0.31)	(2.50)
FEMALE	-2.07	-0.94	-1.13	-1.83	-0.67	-1.16
	(0.27)	(0.06)	(0.28)	(0.10)	(0.03)	(0.11)
CENTER	-1.91	2.63	-4.53	-2.27	0.63	-2.90
	(0.54)	(0.45)	(0.70)	(0.54)	(0.18)	(0.57)
SOUTH & ISLANDS	-4.69	5.14	-9.83	-5.05	3.63	-8.68
	(0.52)	(0.45)	(0.69)	(0.53)	(0.23)	(0.58)
ITALIAN STUDENT	9.33	7.77	1.57	9.19	7.78	1.41
	(0.59)	(0.15)	(0.61)	(0.24)	(0.07)	(0.25)
NUMBER OF CLASSES	0.10	-0.34	0.44	-0.03	-0.40	0.38
	(0.07)	(0.02)	(0.07)	(0.07)	(0.01)	(0.07)
CLASS SIZE	0.30	0.06	0.24	0.59	0.07	0.53
	(0.04)	(0.01)	(0.04)	(0.04)	(0.01)	(0.04)
$\hat{\sigma}_\eta$	1.03	1.16	-0.14	1.05	1.06	-0.01
	(0.01)	(0.00)	(0.01)	(0.02)	(0.00)	(0.02)
$\hat{\rho}$	-	-	-	1.62	0.85	0.77
				(0.11)	(0.00)	(0.11)

Notes: $\hat{\bar{\xi}}$ denotes the APE of the question effects, $\hat{\sigma}_\eta$ denotes the standard deviation of the individual effects distribution, and $\hat{\rho}$ the parameter of its Clayton copula. Columns 1-3 show the APE of the covariates and the estimates of σ_η with the RE logit estimator (Equation 4); columns 4-6 show the same estimates and those of ρ with the CBRE estimator (Equation 5).

Table 12: RE & CBRE logit estimates, 5th grade mathematics exam

	RE			CBRE		
	External	Internal	Difference	External	Internal	Difference
	(1)	(2)	(3)	(4)	(5)	(6)
$\hat{\bar{\xi}}$	-2.41	4.88	-7.30	-12.19	2.28	-14.47
	(2.12)	(0.54)	(2.19)	(1.34)	(0.35)	(1.38)
FEMALE	-3.50	-2.66	-0.84	-3.61	-2.52	-1.09
	(0.25)	(0.06)	(0.26)	(0.07)	(0.02)	(0.07)
CENTER	-2.04	0.04	-2.09	-1.80	1.95	-3.75
	(0.48)	(0.38)	(0.61)	(0.25)	(0.42)	(0.49)
SOUTH & ISLANDS	-5.72	-1.31	-4.41	-6.19	0.90	-7.09
	(0.46)	(0.39)	(0.60)	(0.26)	(0.44)	(0.51)
ITALIAN STUDENT	8.17	8.38	-0.21	8.33	8.38	-0.05
	(0.46)	(0.11)	(0.47)	(0.14)	(0.04)	(0.14)
NUMBER OF CLASSES	0.23	-0.04	0.28	0.19	-0.02	0.22
	(0.06)	(0.01)	(0.06)	(0.03)	(0.01)	(0.04)
CLASS SIZE	0.19	0.12	0.07	0.69	0.47	0.22
	(0.03)	(0.01)	(0.03)	(0.02)	(0.00)	(0.02)
$\hat{\sigma}_\eta$	0.91	0.99	-0.08	0.91	0.92	-0.01
	(0.01)	(0.00)	(0.01)	(0.01)	(0.00)	(0.01)
$\hat{\rho}$	-	-	-	2.14	1.46	0.68
				(0.10)	(0.01)	(0.10)

Notes: $\hat{\bar{\xi}}$ denotes the APE of the question effects, $\hat{\sigma}_\eta$ denotes the standard deviation of the individual effects distribution, and $\hat{\rho}$ the parameter of its Clayton copula. Columns 1-3 show the APE of the covariates and the estimates of σ_η with the RE logit estimator (Equation 4); columns 4-6 show the same estimates and those of ρ with the CBRE estimator (Equation 5).

Table 13: RE & CBRE logit estimates, 6th grade mathematics exam

	RE			CBRE		
	External	Internal	Difference	External	Internal	Difference
	(1)	(2)	(3)	(4)	(5)	(6)
$\hat{\bar{\xi}}$	-17.43	-6.28	-11.15	-24.89	-5.75	-19.14
	(2.26)	(0.77)	(2.39)	(2.99)	(1.02)	(3.16)
FEMALE	-2.94	-2.49	-0.45	-2.89	-2.50	-0.38
	(0.23)	(0.06)	(0.23)	(0.09)	(0.02)	(0.09)
CENTER	-2.72	-2.79	0.07	-4.13	-2.09	-2.04
	(0.41)	(0.33)	(0.53)	(0.75)	(0.71)	(1.04)
SOUTH & ISLANDS	-8.34	-7.11	-1.23	-10.16	-6.81	-3.35
	(0.38)	(0.35)	(0.51)	(0.70)	(0.76)	(1.03)
ITALIAN STUDENT	9.14	8.42	0.72	9.01	8.35	0.66
	(0.33)	(0.09)	(0.34)	(0.14)	(0.04)	(0.15)
NUMBER OF CLASSES	0.08	0.07	0.01	0.12	0.02	0.10
	(0.04)	(0.01)	(0.04)	(0.08)	(0.01)	(0.08)
CLASS SIZE	0.25	0.32	-0.08	0.57	0.53	0.03
	(0.03)	(0.01)	(0.03)	(0.05)	(0.01)	(0.05)
$\hat{\sigma}_\eta$	0.79	0.78	0.01	0.82	0.78	0.04
	(0.01)	(0.00)	(0.01)	(0.01)	(0.00)	(0.01)
$\hat{\rho}$	-	-	-	1.38	1.34	0.03
				(0.06)	(0.02)	(0.06)

Notes: $\hat{\bar{\xi}}$ denotes the APE of the question effects, $\hat{\sigma}_\eta$ denotes the standard deviation of the individual effects distribution, and $\hat{\rho}$ the parameter of its Clayton copula. Columns 1-3 show the APE of the covariates and the estimates of σ_η with the RE logit estimator (Equation 4); columns 4-6 show the same estimates and those of ρ with the CBRE estimator (Equation 5).

Table 14: RE & CBRE logit estimates, 8th grade mathematics exam

	RE			CBRE		
	External	Internal	Difference	External	Internal	Difference
	(1)	(2)	(3)	(4)	(5)	(6)
$\hat{\bar{\xi}}$	-4.45	-5.22	0.77	-6.29	-7.99	1.69
	(2.12)	(0.63)	(2.21)	(1.91)	(0.54)	(1.98)
FEMALE	-4.23	-3.38	-0.85	-3.98	-3.12	-0.86
	(0.24)	(0.07)	(0.25)	(0.09)	(0.02)	(0.09)
CENTER	-0.68	1.26	-1.94	-1.14	-0.29	-0.86
	(0.35)	(0.26)	(0.44)	(0.40)	(0.27)	(0.49)
SOUTH & ISLANDS	-1.96	0.61	-2.57	-2.81	0.52	-3.33
	(0.22)	(0.14)	(0.26)	(0.20)	(0.14)	(0.25)
ITALIAN STUDENT	9.96	8.80	1.17	9.48	8.25	1.23
	(0.39)	(0.11)	(0.40)	(0.15)	(0.04)	(0.15)
NUMBER OF CLASSES	0.17	0.12	0.06	0.01	0.04	-0.03
	(0.04)	(0.01)	(0.04)	(0.04)	(0.01)	(0.04)
CLASS SIZE	0.37	0.38	-0.01	0.98	0.79	0.20
	(0.03)	(0.01)	(0.03)	(0.04)	(0.01)	(0.04)
$\hat{\sigma}_\eta$	0.85	0.86	-0.01	0.79	0.80	-0.01
	(0.01)	(0.00)	(0.01)	(0.00)	(0.00)	(0.00)
$\hat{\rho}$	-	-	-	1.37	1.42	-0.06
				(0.02)	(0.01)	(0.03)

Notes: $\hat{\bar{\xi}}$ denotes the APE of the question effects, $\hat{\sigma}_\eta$ denotes the standard deviation of the individual effects distribution, and $\hat{\rho}$ the parameter of its Clayton copula. Columns 1-3 show the APE of the covariates and the estimates of σ_η with the RE logit estimator (Equation 4); columns 4-6 show the same estimates and those of ρ with the CBRE estimator (Equation 5).

Table 15: RE & CBRE logit estimates, 2nd grade Italian exam

	RE			CBRE		
	External	Internal	Difference	External	Internal	Difference
	(1)	(2)	(3)	(4)	(5)	(6)
$\hat{\bar{\xi}}$	3.86	13.93	-10.07	4.83	15.06	-10.23
	(1.78)	(0.53)	(1.86)	(1.29)	(0.33)	(1.33)
FEMALE	1.42	1.84	-0.41	1.38	1.74	-0.35
	(0.24)	(0.06)	(0.25)	(0.17)	(0.04)	(0.18)
CENTER	-0.15	0.09	-0.23	-1.02	-0.89	-0.13
	(0.39)	(0.24)	(0.46)	(0.43)	(0.10)	(0.44)
SOUTH & ISLANDS	-1.39	1.84	-3.23	-1.73	1.38	-3.10
	(0.33)	(0.23)	(0.40)	(0.33)	(0.11)	(0.35)
ITALIAN STUDENT	9.13	7.03	2.10	8.65	6.89	1.76
	(0.51)	(0.13)	(0.52)	(0.37)	(0.07)	(0.38)
NUMBER OF CLASSES	0.00	-0.26	0.26	-0.03	-0.27	0.24
	(0.06)	(0.01)	(0.06)	(0.06)	(0.01)	(0.06)
CLASS SIZE	0.17	0.10	0.07	0.13	0.10	0.03
	(0.03)	(0.01)	(0.03)	(0.04)	(0.00)	(0.04)
$\hat{\sigma}_\eta$	0.79	0.88	-0.09	0.76	0.85	-0.08
	(0.01)	(0.00)	(0.01)	(0.00)	(0.00)	(0.00)
$\hat{\rho}$	-	-	-	0.19	0.51	-0.32
				(0.01)	(0.00)	(0.01)

Notes: $\hat{\bar{\xi}}$ denotes the APE of the question effects, $\hat{\sigma}_\eta$ denotes the standard deviation of the individual effects distribution, and $\hat{\rho}$ the parameter of its Clayton copula. Columns 1-3 show the APE of the covariates and the estimates of σ_η with the RE logit estimator (Equation 4); columns 4-6 show the same estimates and those of ρ with the CBRE estimator (Equation 5).

Table 16: RE & CBRE logit estimates, 5th grade Italian exam

	RE			CBRE		
	External	Internal	Difference	External	Internal	Difference
	(1)	(2)	(3)	(4)	(5)	(6)
$\hat{\bar{\xi}}$	6.99	11.14	-4.15	4.43	9.55	-5.12
	(1.30)	(0.38)	(1.35)	(0.54)	(0.17)	(0.57)
FEMALE	2.57	2.89	-0.32	2.57	2.90	-0.33
	(0.14)	(0.03)	(0.14)	(0.03)	(0.01)	(0.03)
CENTER	1.15	2.73	-1.59	1.21	2.86	-1.66
	(0.27)	(0.17)	(0.32)	(0.09)	(0.08)	(0.12)
SOUTH & ISLANDS	-2.92	0.53	-3.45	-2.89	0.66	-3.55
	(0.27)	(0.19)	(0.33)	(0.09)	(0.08)	(0.12)
ITALIAN STUDENT	9.77	9.10	0.67	9.86	9.12	0.74
	(0.27)	(0.06)	(0.27)	(0.05)	(0.01)	(0.05)
NUMBER OF CLASSES	0.02	-0.12	0.14	0.02	-0.13	0.15
	(0.04)	(0.01)	(0.04)	(0.01)	(0.00)	(0.01)
CLASS SIZE	0.14	0.14	-0.01	0.17	0.18	-0.01
	(0.02)	(0.00)	(0.02)	(0.01)	(0.00)	(0.01)
$\hat{\sigma}_\eta$	0.93	0.95	-0.02	0.91	0.93	-0.03
	(0.01)	(0.00)	(0.01)	(0.01)	(0.00)	(0.01)
$\hat{\rho}$	-	-	-	3.95	2.83	1.12
				(0.03)	(0.01)	(0.03)

Notes: $\hat{\bar{\xi}}$ denotes the APE of the question effects, $\hat{\sigma}_\eta$ denotes the standard deviation of the individual effects distribution, and $\hat{\rho}$ the parameter of its Clayton copula. Columns 1-3 show the APE of the covariates and the estimates of σ_η with the RE logit estimator (Equation 4); columns 4-6 show the same estimates and those of ρ with the CBRE estimator (Equation 5).

Table 17: RE & CBRE logit estimates, 6th grade Italian exam

	RE			CBRE		
	External	Internal	Difference	External	Internal	Difference
	(1)	(2)	(3)	(4)	(5)	(6)
$\hat{\bar{\xi}}$	-9.76	2.69	-12.45	-13.94	0.75	-14.69
	(1.84)	(0.72)	(1.97)	(0.70)	(0.30)	(0.76)
FEMALE	3.64	3.67	-0.02	3.68	3.67	0.01
	(0.18)	(0.05)	(0.19)	(0.04)	(0.01)	(0.04)
CENTER	-0.98	0.24	-1.22	-0.82	0.34	-1.15
	(0.37)	(0.30)	(0.48)	(0.13)	(0.12)	(0.18)
SOUTH & ISLANDS	-5.17	-4.03	-1.13	-4.97	-3.99	-0.98
	(0.37)	(0.32)	(0.49)	(0.13)	(0.13)	(0.18)
ITALIAN STUDENT	13.40	12.36	1.04	13.56	12.42	1.14
	(0.29)	(0.08)	(0.30)	(0.06)	(0.02)	(0.06)
NUMBER OF CLASSES	0.21	0.04	0.16	0.22	0.04	0.18
	(0.03)	(0.01)	(0.04)	(0.01)	(0.00)	(0.01)
CLASS SIZE	0.25	0.36	-0.11	0.34	0.41	-0.07
	(0.02)	(0.01)	(0.03)	(0.01)	(0.00)	(0.01)
$\hat{\sigma}_\eta$	0.87	0.87	0.00	0.84	0.85	-0.01
	(0.01)	(0.00)	(0.01)	(0.01)	(0.00)	(0.01)
$\hat{\rho}$	-	-	-	3.64	2.77	0.86
				(0.12)	(0.01)	(0.12)

Notes: $\hat{\bar{\xi}}$ denotes the APE of the question effects, $\hat{\sigma}_\eta$ denotes the standard deviation of the individual effects distribution, and $\hat{\rho}$ the parameter of its Clayton copula. Columns 1-3 show the APE of the covariates and the estimates of σ_η with the RE logit estimator (Equation 4); columns 4-6 show the same estimates and those of ρ with the CBRE estimator (Equation 5).

Table 18: RE & CBRE logit estimates, 8th grade Italian exam

	RE			CBRE		
	External	Internal	Difference	External	Internal	Difference
	(1)	(2)	(3)	(4)	(5)	(6)
$\hat{\bar{\xi}}$	5.87	7.16	-1.28	3.57	5.06	-1.49
	(1.71)	(0.48)	(1.77)	(0.82)	(0.23)	(0.85)
FEMALE	3.16	3.12	0.04	3.15	3.11	0.05
	(0.15)	(0.04)	(0.16)	(0.03)	(0.01)	(0.03)
CENTER	0.02	-0.14	0.16	0.04	-0.14	0.18
	(0.14)	(0.09)	(0.17)	(0.05)	(0.04)	(0.06)
SOUTH & ISLANDS	-1.70	-1.14	-0.57	-1.71	-1.06	-0.65
	(0.14)	(0.05)	(0.15)	(0.05)	(0.02)	(0.06)
ITALIAN STUDENT	11.54	10.11	1.43	11.60	10.09	1.51
	(0.26)	(0.07)	(0.27)	(0.05)	(0.02)	(0.06)
NUMBER OF CLASSES	0.11	0.10	0.00	0.10	0.11	0.00
	(0.03)	(0.01)	(0.03)	(0.01)	(0.00)	(0.01)
CLASS SIZE	0.36	0.33	0.04	0.41	0.37	0.03
	(0.02)	(0.00)	(0.02)	(0.01)	(0.00)	(0.01)
$\hat{\sigma}_\eta$	0.86	0.87	-0.01	0.84	0.85	-0.01
	(0.01)	(0.00)	(0.01)	(0.01)	(0.00)	(0.01)
$\hat{\rho}$	-	-	-	3.85	2.77	1.08
				(0.09)	(0.01)	(0.09)

Notes: $\hat{\bar{\xi}}$ denotes the APE of the question effects, $\hat{\sigma}_\eta$ denotes the standard deviation of the individual effects distribution, and $\hat{\rho}$ the parameter of its Clayton copula. Columns 1-3 show the APE of the covariates and the estimates of σ_η with the RE logit estimator (Equation 4); columns 4-6 show the same estimates and those of ρ with the CBRE estimator (Equation 5).

Table 19: RE & CBRE logit estimates, 10th grade Italian exam

	RE			CBRE		
	External	Internal	Difference	External	Internal	Difference
	(1)	(2)	(3)	(4)	(5)	(6)
$\hat{\bar{\xi}}$	6.07	-6.91	12.98	4.40	-7.95	12.35
	(1.56)	(0.98)	(1.84)	(0.76)	(0.37)	(0.85)
FEMALE	2.30	2.88	-0.57	2.22	2.56	-0.34
	(0.14)	(0.05)	(0.15)	(0.04)	(0.02)	(0.04)
CENTER	-5.00	9.63	-14.63	-4.95	10.73	-15.68
	(0.37)	(0.28)	(0.46)	(0.18)	(0.20)	(0.27)
SOUTH & ISLANDS	-10.35	8.09	-18.44	-10.44	8.55	-18.99
	(0.38)	(0.33)	(0.50)	(0.19)	(0.30)	(0.35)
ITALIAN STUDENT	8.12	8.51	-0.39	8.04	7.93	0.11
	(0.21)	(0.08)	(0.22)	(0.05)	(0.02)	(0.06)
NUMBER OF CLASSES	0.16	0.16	0.00	0.17	0.14	0.03
	(0.02)	(0.01)	(0.02)	(0.01)	(0.00)	(0.01)
CLASS SIZE	0.98	0.90	0.08	1.02	0.86	0.16
	(0.01)	(0.00)	(0.01)	(0.01)	(0.00)	(0.01)
$\hat{\sigma}_\eta$	0.76	0.83	-0.07	0.74	0.80	-0.06
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
$\hat{\rho}$	-	-	-	3.77	2.02	1.76
				(0.02)	(0.00)	(0.02)

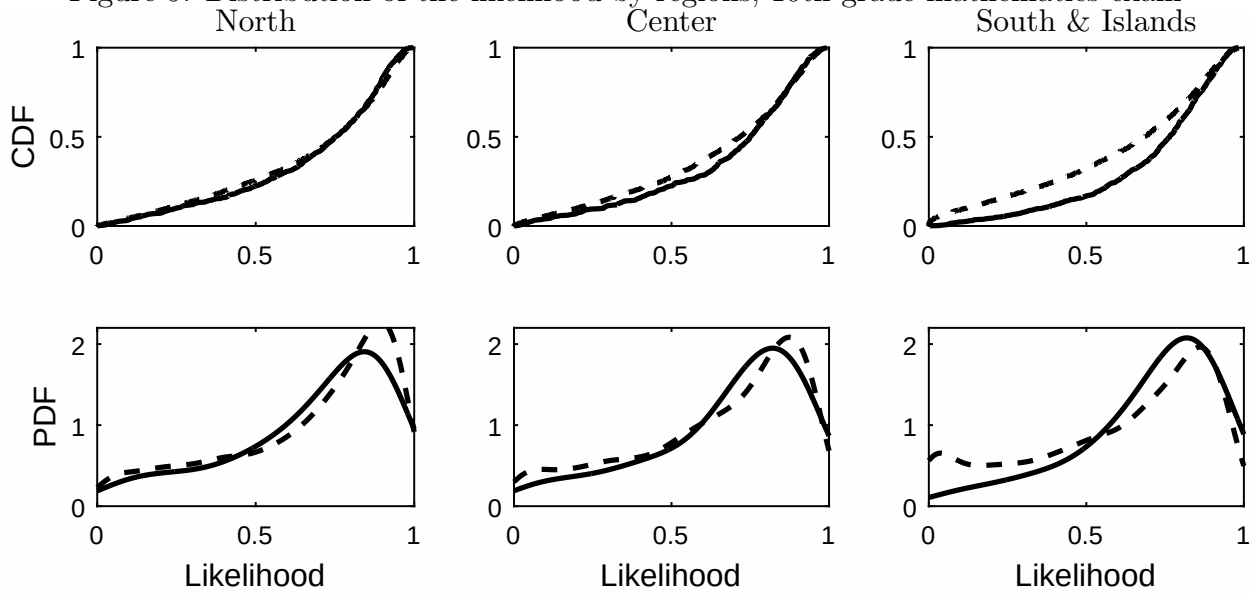
Notes: $\hat{\bar{\xi}}$ denotes the APE of the question effects, $\hat{\sigma}_\eta$ denotes the standard deviation of the individual effects distribution, and $\hat{\rho}$ the parameter of its Clayton copula. Columns 1-3 show the APE of the covariates and the estimates of σ_η with the RE logit estimator (Equation 4); columns 4-6 show the same estimates and those of ρ with the CBRE estimator (Equation 5).

Table 20: Linear correlation equivalent of the copula estimates, all exams

	2nd grade		5th grade		6th grade		8th grade		10th grade	
	M	I	M	I	M	I	M	I	M	I
EX	0.65	0.14	0.73	0.86	0.60	0.85	0.60	0.86	0.77	0.86
IN	0.45	0.31	0.62	0.80	0.59	0.79	0.61	0.79	0.63	0.71

Notes: EX and IN respectively denote the groups with the external and the internal monitor. The coefficients equal the linear correlation of a Gaussian copula that yields the same value of the Kendall's τ statistic as the estimates of the Clayton copula parameter.

Figure 9: Distribution of the likelihood by regions, 10th grade mathematics exam



Distribution of the estimated likelihood of the class scores (Equation 6). EX and IN respectively denote the groups with the external and the internal monitor.

Figure 10: Sensitivity of the corrected test scores to the minimum probability of manipulation, 10th grade mathematics exam

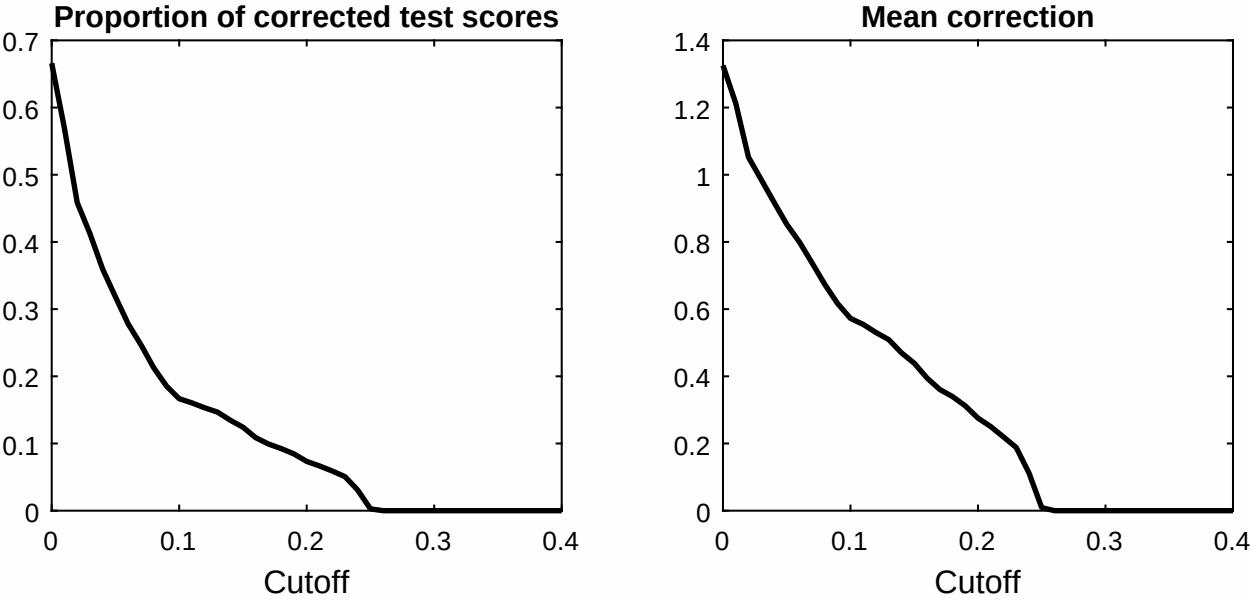


Figure 11: Correction for cheating, provincial variation, 2nd grade mathematics exam

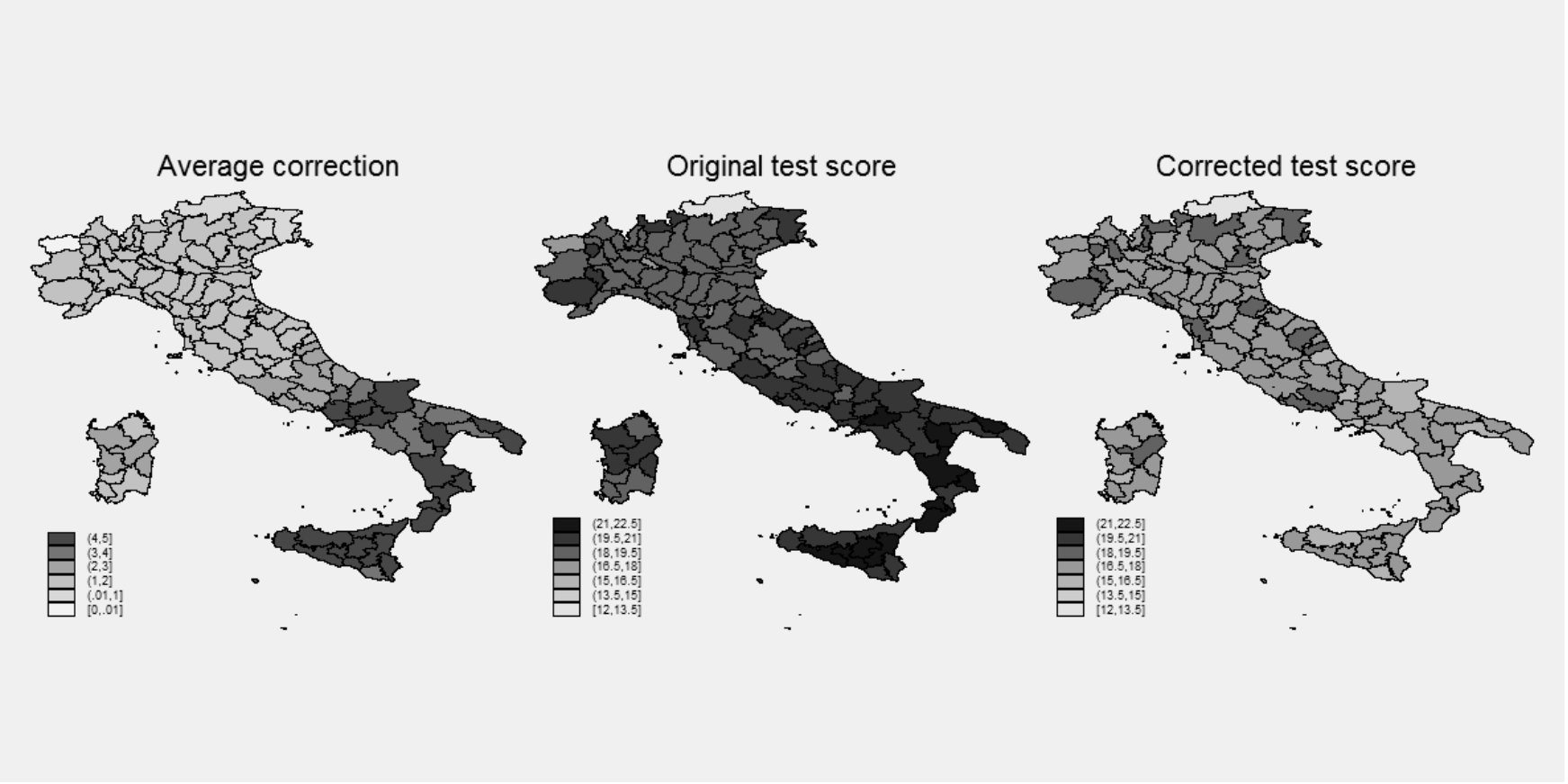


Figure 12: Correction for cheating, provincial variation, 5th grade mathematics exam

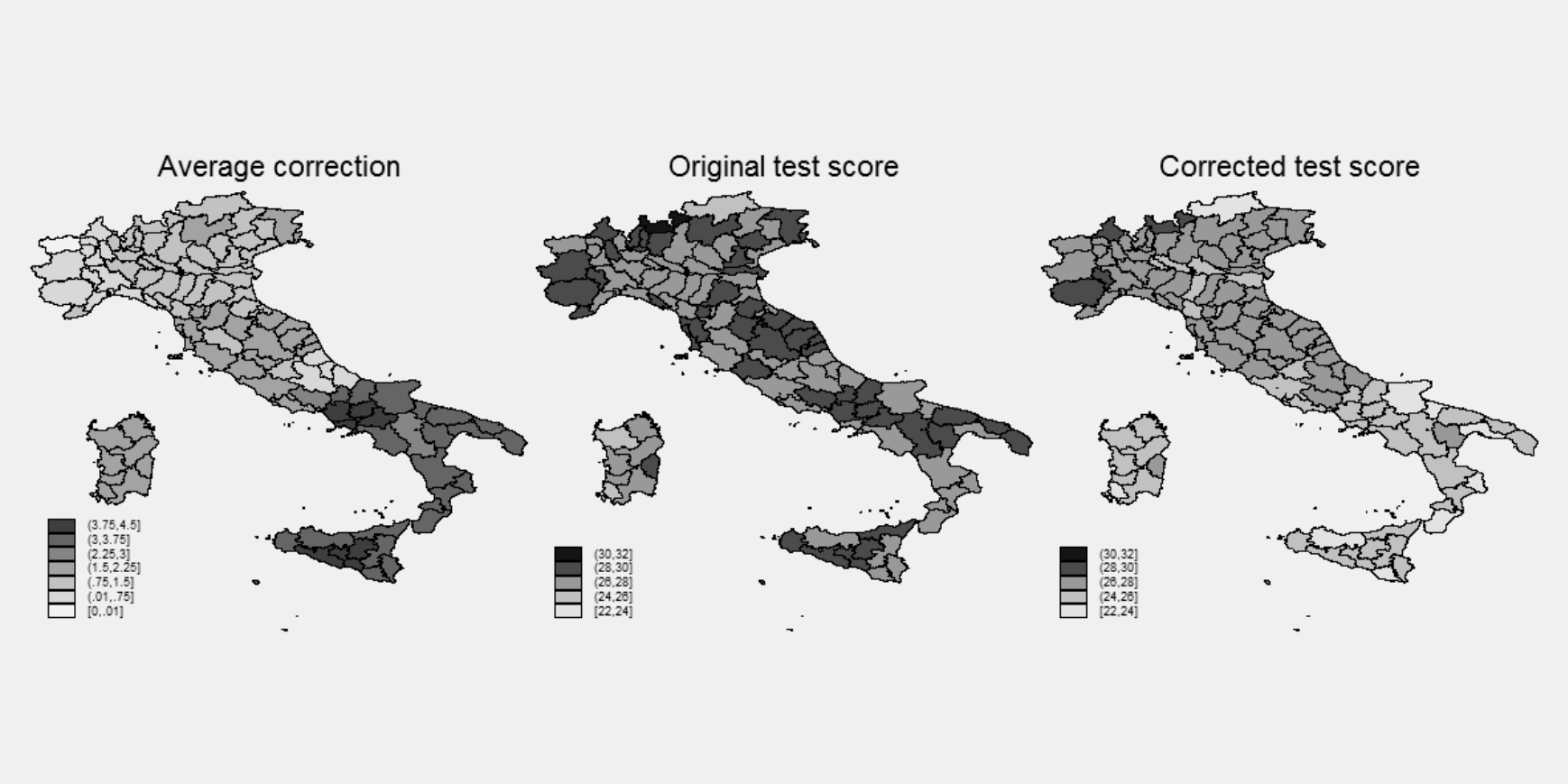


Figure 13: Correction for cheating, provincial variation, 6th grade mathematics exam

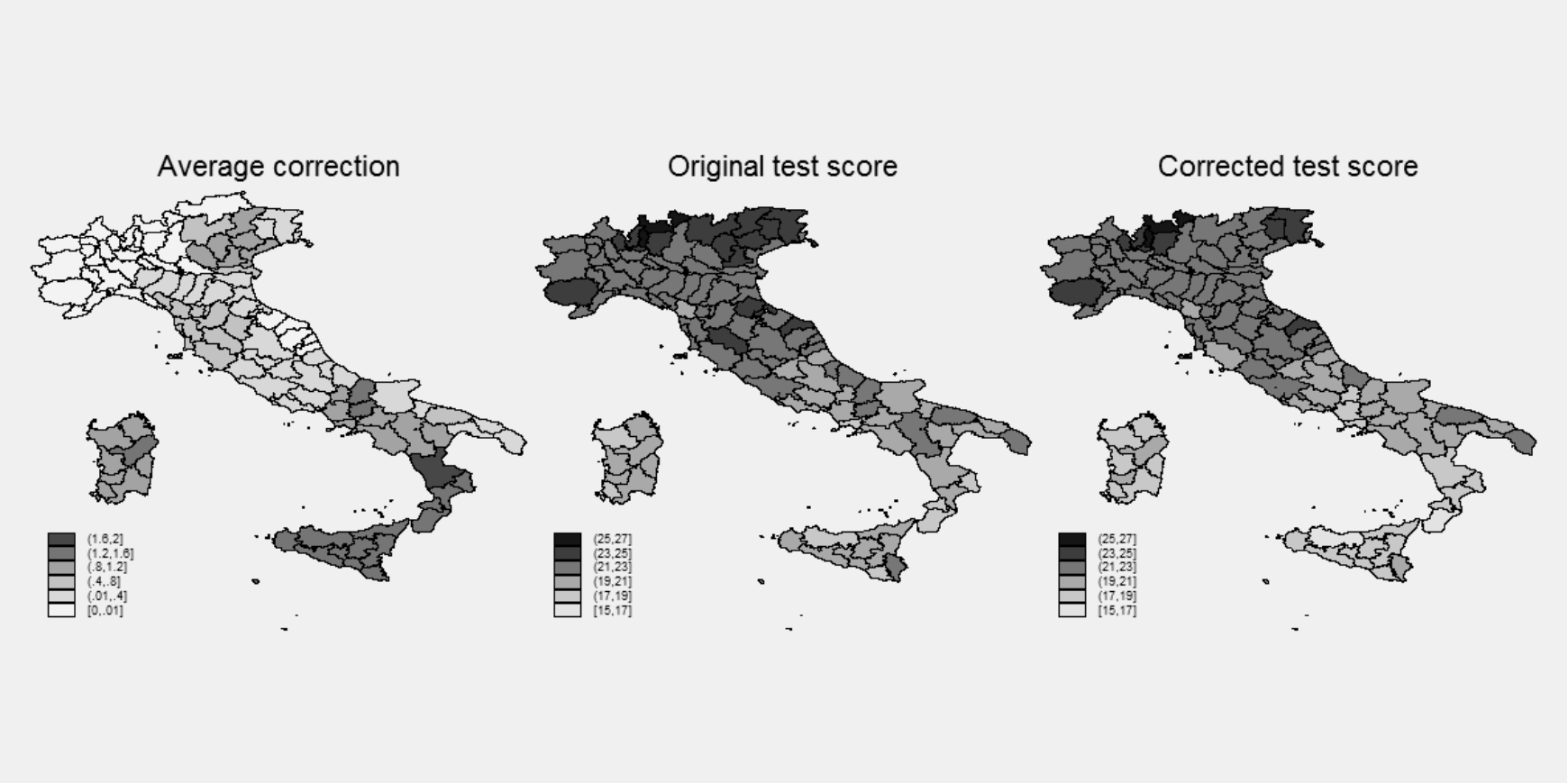


Figure 14: Correction for cheating, provincial variation, 8th grade mathematics exam

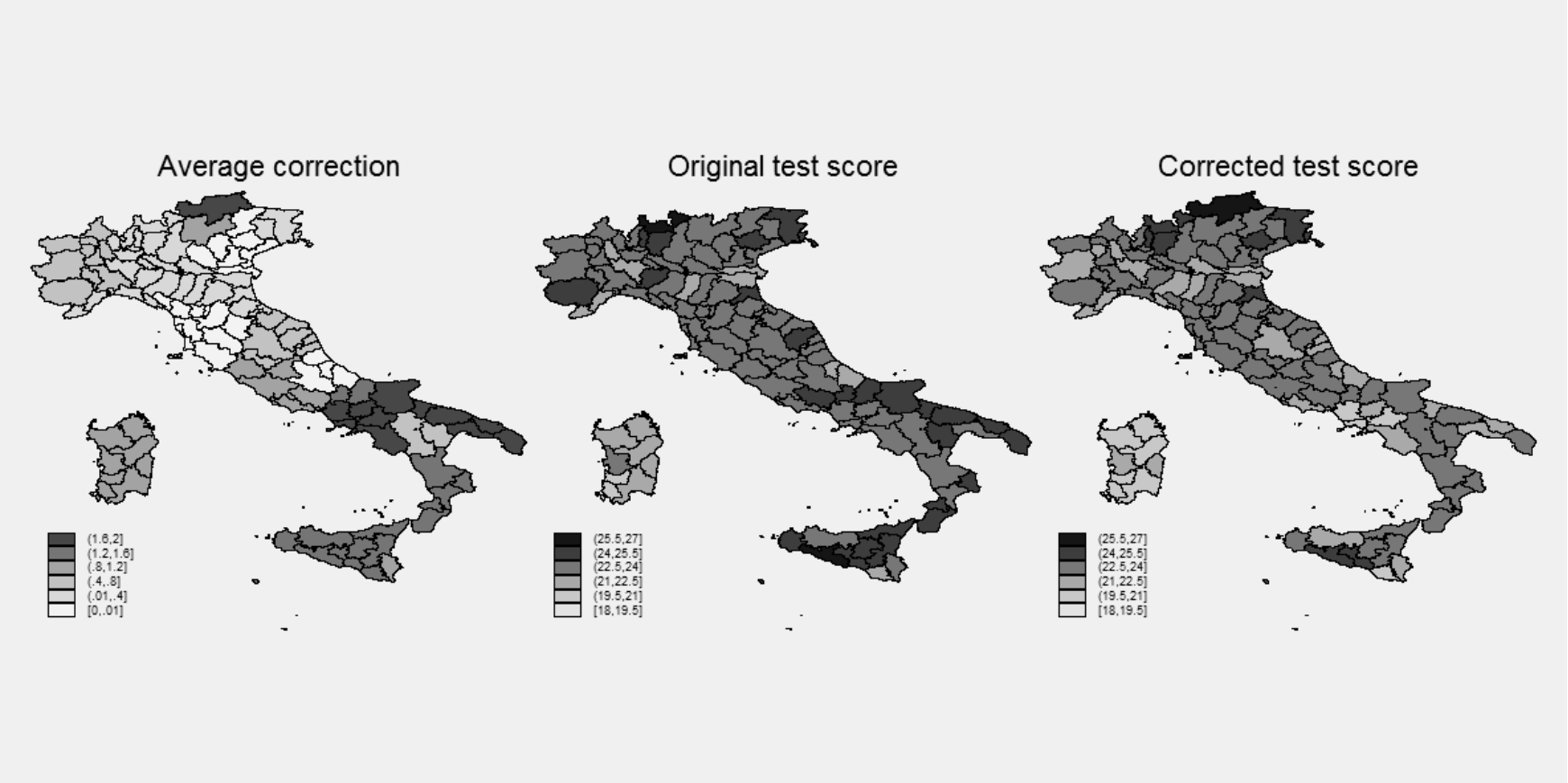


Figure 15: Correction for cheating, provincial variation, 2nd grade Italian exam

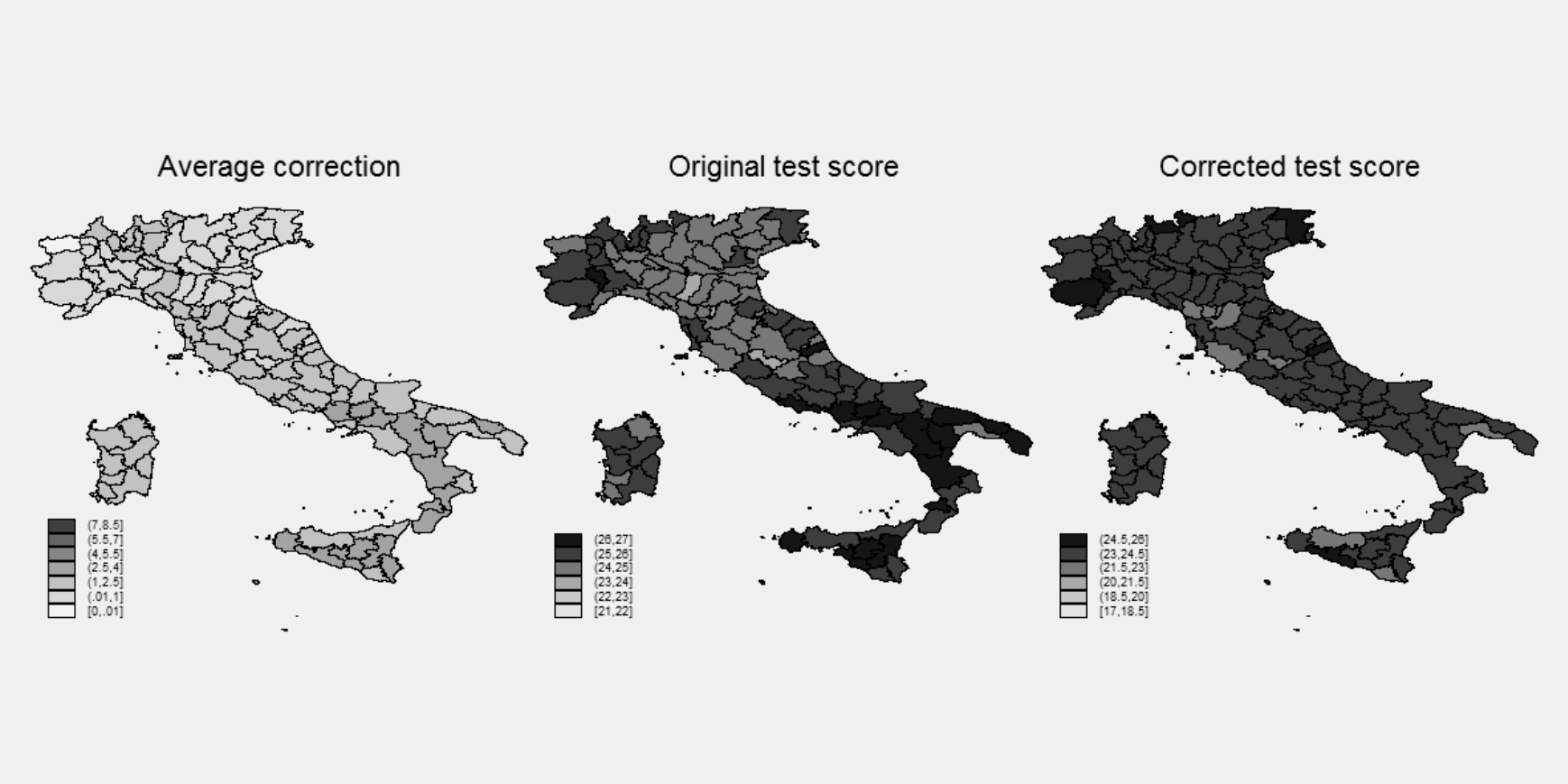


Figure 16: Correction for cheating, provincial variation, 5th grade Italian exam

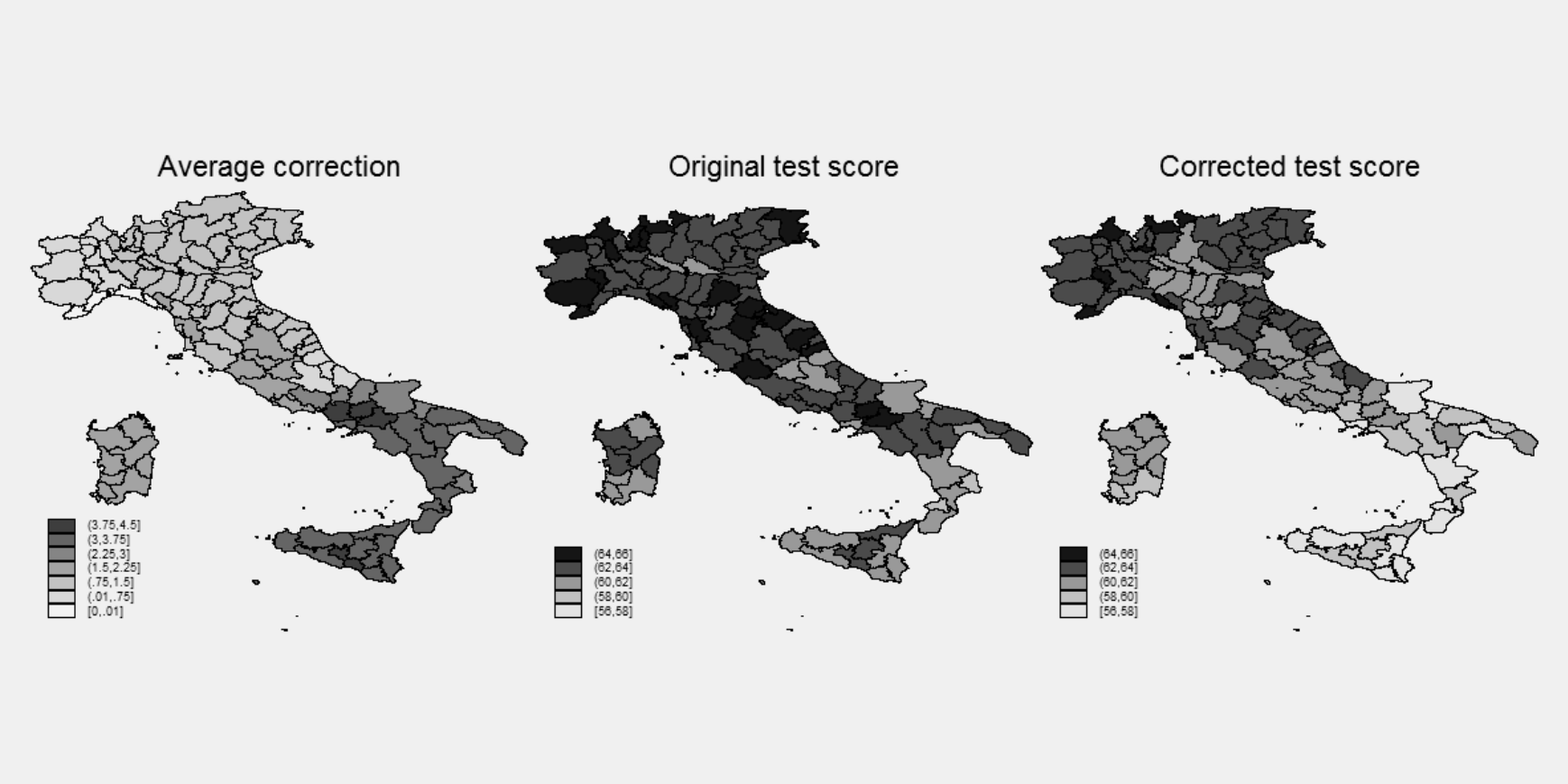


Figure 17: Correction for cheating, provincial variation, 6th grade Italian exam

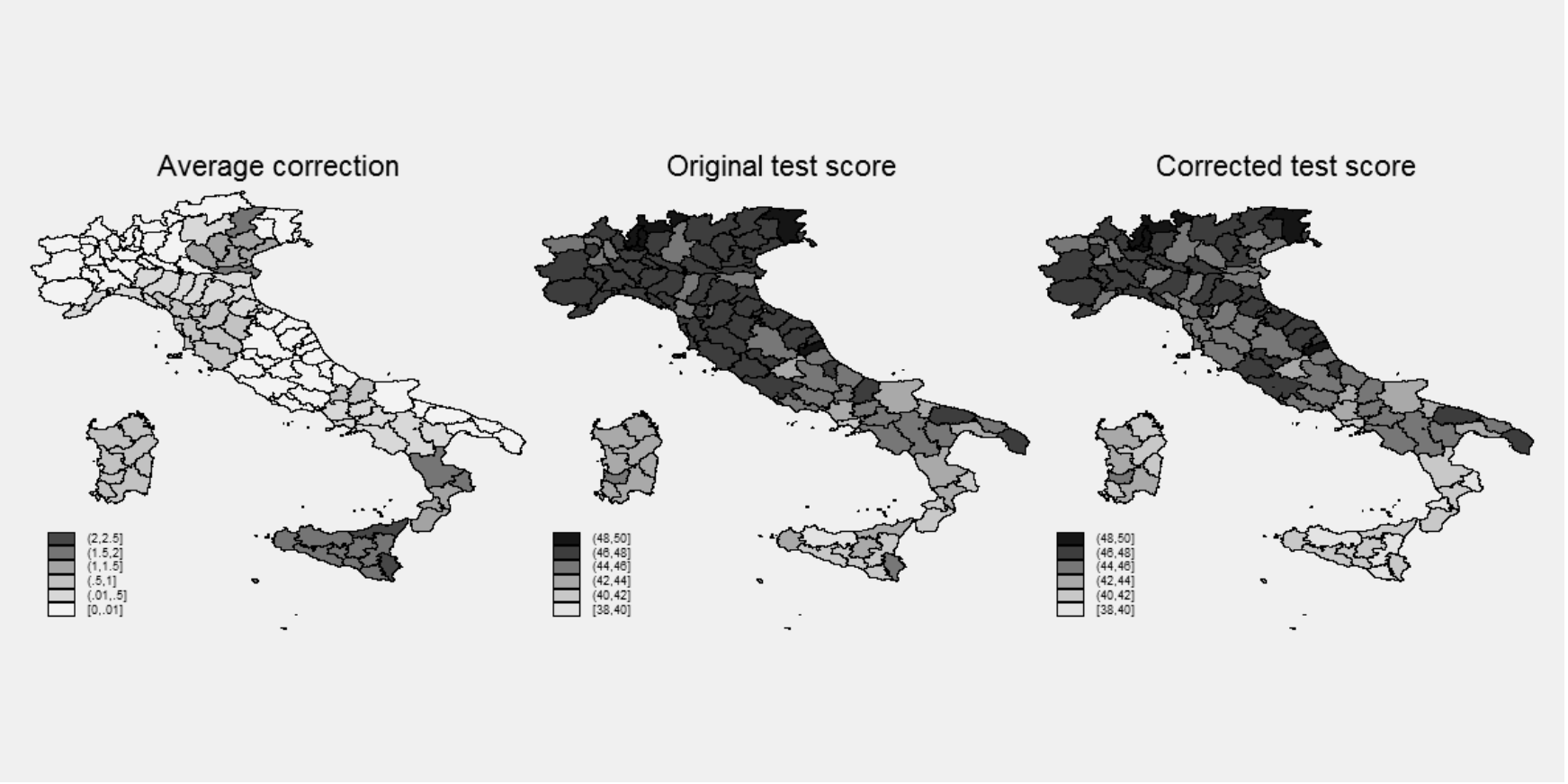


Figure 18: Correction for cheating, provincial variation, 8th grade Italian exam

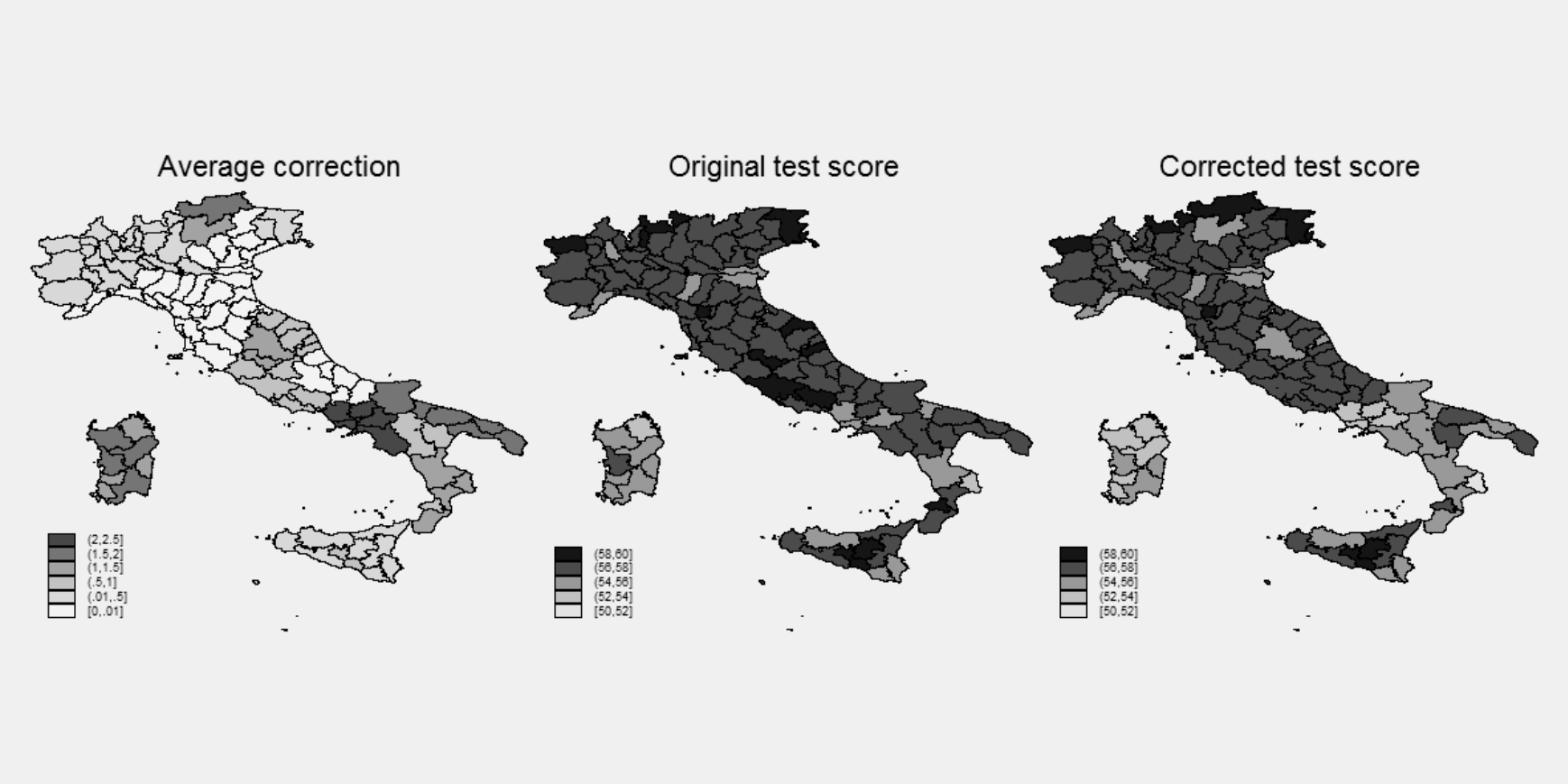
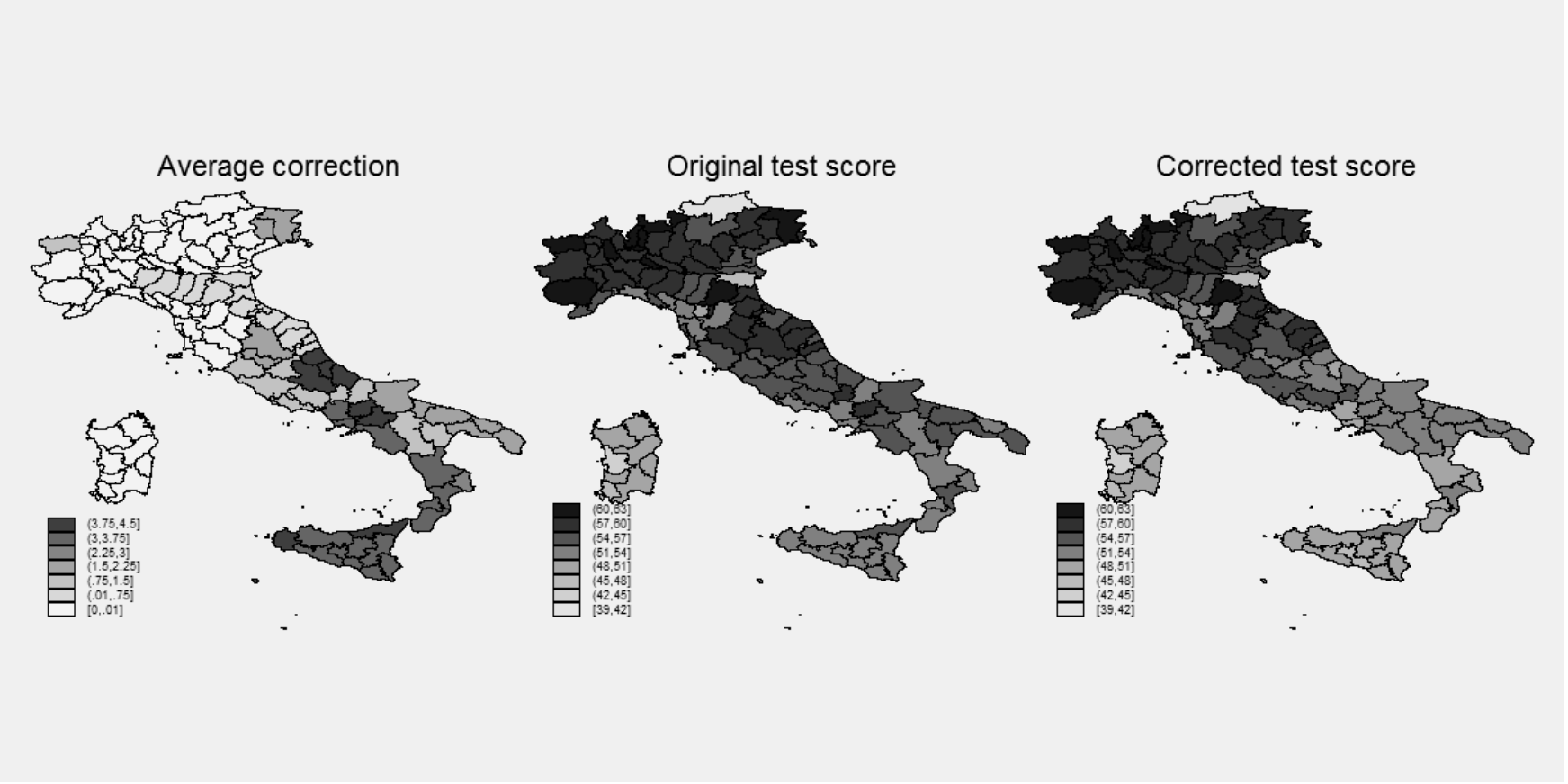


Figure 19: Correction for cheating, provincial variation, 10th grade Italian exam



S2 Conditional Fixed Effects Approach

Consider the model given by 3. If ε_{icq} is logistically distributed, one can follow Chamberlain (1980) to overcome the incidental parameter problem, obtaining estimates of the question fixed effects.[30]. Notice however, that because of multicollinearity, it is necessary to exclude one of the question effects for each group. Then, the interpretation of the remaining $Q-1$ question effects is the difficulty of question q relative to the excluded question. In other words, I normalize the excluded question, $\tilde{q}$, to have $\xi_{\tilde{q}}=0$.

Let B_r be defined as the set of permutations of y such that the total number of correct answers is r, *i.e.* $B_r \equiv \left\{b : \sum_{q=1}^{Q} b_q = r\right\}$.[31] Under the assumption of no cheating, once the student-class effects are accounted for, the answers of two students are independent. Hence, the log-likelihood function is given by

$$\mathcal{L}(\xi) = \sum_{c=1}^{C}\sum_{i=1}^{N_c} \log\left[\mathbb{P}\left(y_{ic}|r_{ic}\right)\right] = \sum_{c=1}^{C}\sum_{i=1}^{N_c} y_{ic}'\xi - \sum_{c=1}^{C}\sum_{i=1}^{N_c} \log\left[\sum_{b\in B_{r_{ic}}} \exp\left(b'\xi\right)\right] \tag{9}$$

S2.1 Results

Figure 20 shows the estimates of ξ for the mathematics exam of 10th graders.[32] Similarly to Figure 1, there is a weak pattern, as more difficult questions tend to have slightly larger differences between the treatment and control groups estimates. Further, the estimates of ξ_q are significantly different for the treatment and the control groups for 34 out of 49 questions, of which 29 show that the coefficient for the treatment group is significantly smaller. Moreover, although the coefficients are not directly comparable to the estimates shown in Figure 2, the relation between the two of them is almost linear, with a correlation coefficient of approximately one for this exam, suggesting that the parametric assumption

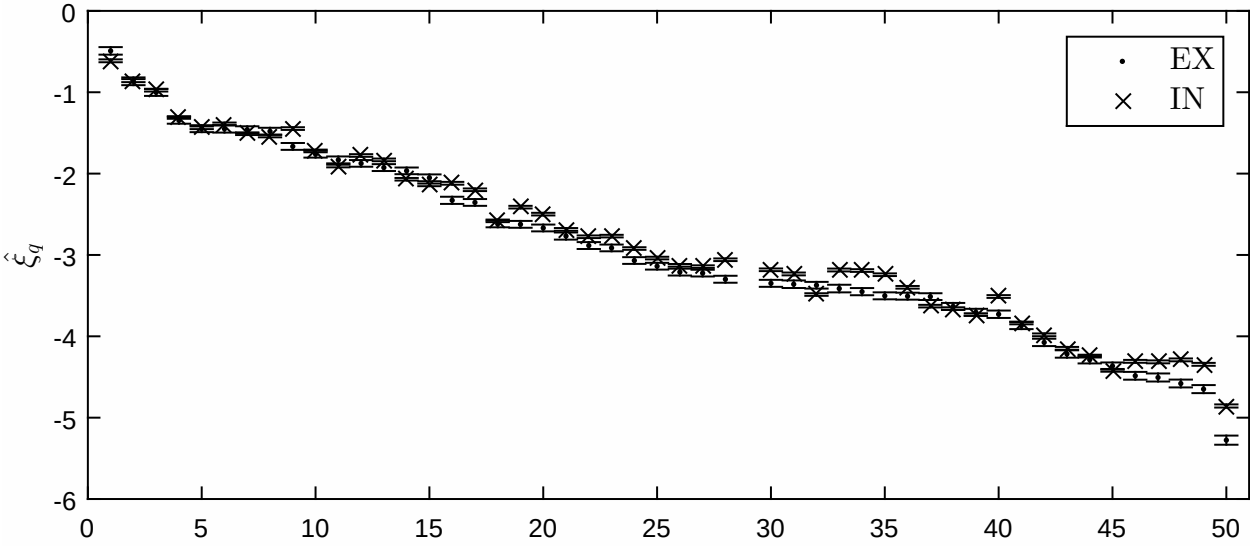
[30] As usual in this kind of setups, the identification relies on a parametric assumption of an unobservable variable that is not verifiable. As recently showed by Bonhomme (2012), it is possible to estimate the question fixed effects even if the parametric distribution of ε_{icq} is not logistic. However, given the large size of the data set, both in terms of number of students and of number of questions in an exam, assuming a distribution other than the logistic is computationally impractical.

[31] The total number of permutations equals $\binom{Q}{r}$.

[32] Since I had to exclude one of the questions to avoid multicollinearity, and in order to make them as interpretable as possible, I excluded the question that was more frequently correctly answered.

does not play a big role in determining the value of the coefficients. These results are robust to most exams, as shown in Table 21.

Figure 20: Conditional FE logit estimates, 10th grade mathematics exam



FE logit estimates of the question effects (ξ_q in Equation 9) for the group with an external monitor (EX) and the group with an internal monitor (IN). They are reported along with the 95% confidence intervals, and sorted by how frequently they were correctly answered by students proctored by an external monitor.

Table 21: Conditional FE logit estimates

	2nd grade		5th grade		6th grade		8th grade		10th grade	
	M	I	M	I	M	I	M	I	M	I
$\hat{\xi}^{EX} > \hat{\xi}^{IN}$	4	3	1	2	0	2	2	1	6	4
$\hat{\xi}^{EX} < \hat{\xi}^{IN}$	6	18	32	18	39	8	6	37	29	13
$\hat{\xi}^{EX} = \hat{\xi}^{IN}$	21	17	13	61	8	60	36	39	14	70

Notes: EX and IN respectively denote the groups with the external and the internal monitor. I and M respectively denote the Italian and mathematics exams. A coefficient is considered as larger than the other if it is significantly larger at the 95% confidence level, and equal if none is statistically larger than the other.

Another alternative is to consider the estimation of the same coefficients for different demographic groups, such as gender. The comparison between the treatment and control groups for each of the genders is very similar to that of the whole population. However, even in the absence of manipulation, there are remarkable gender differences in performance (first three rows in Table 22), with male students performing relatively better than females in 17 questions, and the other way around in 16 questions. For the control group these differences are increased (26 and 19, respectively), which could reflect both the manipulation of the test

scores and the increase in the precision of the estimates derived from the increased sample size.

Table 22: Conditional FE logit estimates by gender

	2nd grade		5th grade		6th grade		8th grade		10th grade	
	M	I	M	I	M	I	M	I	M	I
$\hat{\xi}^{EX,MA} > \hat{\xi}^{EX,FE}$	5	25	24	16	14	12	18	39	17	13
$\hat{\xi}^{EX,MA} < \hat{\xi}^{EX,FE}$	12	1	0	1	17	5	5	0	16	55
$\hat{\xi}^{EX,MA} = \hat{\xi}^{EX,FE}$	14	12	22	64	16	53	21	38	16	19
$\hat{\xi}^{IN,MA} > \hat{\xi}^{IN,FE}$	6	38	26	50	19	32	28	40	26	18
$\hat{\xi}^{IN,MA} < \hat{\xi}^{IN,FE}$	16	0	8	12	25	20	11	18	17	62
$\hat{\xi}^{IN,MA} = \hat{\xi}^{IN,FE}$	9	0	12	19	3	18	5	19	6	7

Notes: EX and IN respectively denote the groups with the external and the internal monitor, whereas MA and FE denote male and female students. I and M respectively denote the Italian and mathematics exams. A coefficient is considered as larger than the other if it is significantly larger at the 95% confidence level, and equal if none is statistically larger than the other.

This result is robust to all exams, but not to all possible categories, as shown in Tables 23 and 24. In particular, splitting the sample by class size leads to almost no differences in the estimates in the treatment group, but significant differences in the control group for most exams. Focusing on the three macro regions of Italy, there are large differences between the estimates for the control groups in the North and South & Islands regions.

Table 23: Conditional FE logit estimates by class size

	2nd grade		5th grade		6th grade		8th grade		10th grade	
	M	I	M	I	M	I	M	I	M	I
$\hat{\xi}^{EX,SM} > \hat{\xi}^{EX,LA}$	0	3	15	0	2	0	4	0	15	11
$\hat{\xi}^{EX,SM} < \hat{\xi}^{EX,LA}$	5	0	0	0	0	0	2	0	6	32
$\hat{\xi}^{EX,SM} = \hat{\xi}^{EX,LA}$	26	35	41	81	45	70	38	77	28	44
$\hat{\xi}^{IN,SM} > \hat{\xi}^{IN,LA}$	14	26	22	64	20	5	31	16	28	21
$\hat{\xi}^{IN,SM} < \hat{\xi}^{IN,LA}$	4	2	1	0	9	22	5	3	6	48
$\hat{\xi}^{IN,SM} = \hat{\xi}^{IN,LA}$	13	10	23	17	18	43	8	58	15	18

Notes: EX and IN respectively denote the groups with the external and the internal monitor, whereas SM and LA denote that the students were in classrooms of size smaller or equal to the median, and larger. I and M respectively denote the Italian and mathematics exams. A coefficient is considered as larger than the other if it is significantly larger at the 95% confidence level, and equal if none is statistically larger than the other.

Table 24: Conditional FE logit estimates by region

	2nd grade		5th grade		6th grade		8th grade		10th grade	
	M	I	M	I	M	I	M	I	M	I
$\hat{\xi}^{EX,NO} > \hat{\xi}^{EX,SI}$	1	1	1	0	1	0	1	0	0	0
$\hat{\xi}^{EX,NO} < \hat{\xi}^{EX,SI}$	1	3	0	16	6	2	14	10	14	30
$\hat{\xi}^{EX,NO} = \hat{\xi}^{EX,SI}$	29	34	44	65	40	68	29	67	35	57
$\hat{\xi}^{IN,NO} > \hat{\xi}^{IN,SI}$	7	8	1	0	1	5	0	0	3	2
$\hat{\xi}^{IN,NO} < \hat{\xi}^{IN,SI}$	16	12	35	76	38	20	43	65	39	71
$\hat{\xi}^{IN,NO} = \hat{\xi}^{IN,SI}$	8	18	10	5	8	45	1	12	7	14

Notes: EX and IN respectively denote the groups with the external and the internal monitor, whereas NO and SI denote that the students were from the North and South & Islands regions. I and M respectively denote the Italian and mathematics exams. A coefficient is considered as larger than the other if it is significantly larger at the 95% confidence level, and equal if none is statistically larger than the other.

S3 Heterogeneous Question Fixed Effects

Equation 4 is based on the assumption that once the combined student-class effects are controlled for, there is no correlation in students' answers, *i.e.* the question effects are homogeneous across all students. This assumption could be violated if teachers are more skilled to teach some particular topics than others, which would create correlation in the question effects among students within classrooms, even without manipulation. A way to overcome this would be to use the observations of a randomly chosen student from each classroom. Since the correlation is caused by the teacher, then students from different classrooms would be affected by a set of independent effects. Moreover, I avoid making any distributional assumption of the individual effects, for which I use the conditional fixed effects logit estimator, whose likelihood function is given by

$$\mathcal{L}(\beta) = \sum_{c=1}^{C} \log\left(\mathbb{P}\left(y_{i(c)c} | r_{i(c)c}\right)\right) = \sum_{c=1}^{C} y_{i(c)c}\xi - \sum_{c=1}^{C} \log\left[\sum_{b \in B_{r_{i(c)c}}} \exp(b\xi)\right] \tag{10}$$

where $i(c)$ denotes a randomly chosen student from class c. This strategy, unlike the precedent, does not provide a unique estimator, since there are as many as permutations of students: $\prod_{c=1}^{C} N_c$. Given the large number of possible estimates, I randomly select one student from each classroom $M = 1000$ times and then report the median estimate across repetitions. Regarding the confidence intervals, I use the 2.5 and 97.5 percentiles. The results are shown in Table 25. For the treatment group, they are roughly the same as the

ones obtained by using all the students in each classroom, and only for one of the questions in the 10th grade Italian exam the estimates are significantly different. For the control group this is not always the case, and in two of the exams (8th and 10th grade Italian exams) the coefficients for the majority of the questions were significantly different. This reflects the manipulation, as well as the larger sample size for the control group, which tightens the confidence intervals. However, since the correction is based on the estimates with an external monitor, the possibility of having heterogeneous question fixed effects would have a modest impact on its reliability.

Table 25: Conditional FE logit estimates, one student per classroom

	2nd grade		5th grade		6th grade		8th grade		10th grade	
	M	I	M	I	M	I	M	I	M	I
$EX, \neq$	0	0	0	0	0	0	0	0	0	1
$EX, =$	31	46	47	44	49	38	81	70	77	86
$IN, \neq$	7	7	0	6	12	12	2	37	1	70
$IN, =$	24	39	47	38	37	26	79	33	76	17

Notes: EX and IN respectively denote the groups with the external and the internal monitor, $=$ and $\neq$ respectively denote the number of coefficients whose 95% confidence intervals overlapped or did not overlap.