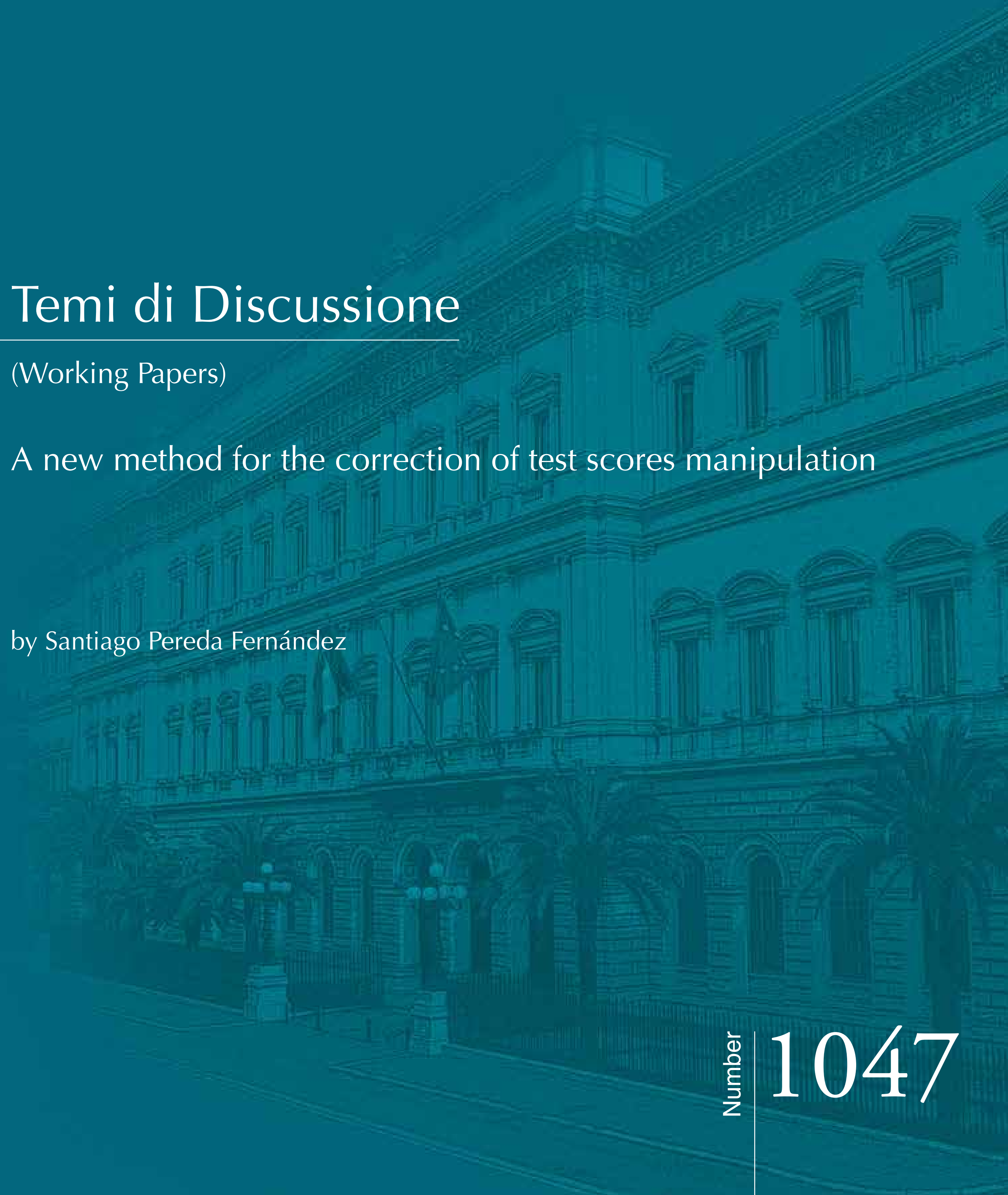
BANCA D'ITALIA
EUROSISTEMA

Temi di Discussione

(Working Papers)

A new method for the correction of test scores manipulation

by Santiago Pereda Fernández



January 2016

Number 1047

BANCA D'ITALIA
EUROSISTEMA

Temi di discussione

(Working papers)

A new method for the correction of test scores manipulation

by Santiago Pereda Fernández

Number 1047 - January 2016

ISSN 1594-7939 (print)
ISSN 2281-3950 (online)

Printed by the Printing and Publishing Division of the Bank of Italy

A NEW METHOD FOR THE CORRECTION OF TEST SCORES MANIPULATION

by Santiago Pereda Fernández*

Abstract

I propose a method to correct for test scores manipulation and apply it to a natural experiment in the Italian education system consisting in the random assignment of external monitors to classrooms. The empirical strategy is based on a likelihood approach, using nonlinear panel data methods to obtain clean estimates of cheating controlling for unobserved heterogeneity. The likelihood of each classroom's scores is later used to correct them for cheating. Cheating is not associated with an increase in the correlation of the answers after we control for mean test scores. The method produces estimates of manipulation more frequent in the South and Islands and among female students and immigrants in Italian tests. A simulation shows how the manipulation reduces the accuracy of an exam in reflecting students' knowledge, and the correction proposed in this paper makes up for about a half of this loss.

JEL Classification: C23, C25, I28.
Keywords: cheating correction, copula, nonlinear panel data, test scores manipulation.

Contents

* Bank of Italy, Structural Economic Analysis Directorate.

1 Introduction[*]

A policy maker interested in evaluating the education system requires a homogeneous measure of academic achievement. Standardized tests permit the comparison of students' knowledge, but even the best possibly designed exam will reflect other factors than students' knowledge in their scores, thus threatening the comparability. Despite this shortcoming, standardized test scores can be and have been used to evaluate teachers,[1] principals,[2] and schools. One major threat to the comparability of these tests is the manipulation of the scores, which alters the students' achievement ranks.[3]

Relying on test scores as the only source of information can hinder the efforts of detecting test scores manipulation, though it is not impossible. Jacob and Levitt (2003) provided an algorithm to detect cheating based on unlikely patterns in the answers of students and sudden increases in students' test scores followed by sudden decreases in the subsequent year. Dee et al. (2011) on the other hand examined the distribution of test scores, finding significant discontinuities at the cutoff scores that determine students' eligibility to graduate. Moreover, the evidence they found suggests that teachers were responsible for increasing students test scores that were below the cutoff. Alternatively, several studies (Figlio and Getzler, 2006; Figlio, 2006; Cullen and Reback, 2006; Hussain, 2015) have shown manipulation in the pool of students who take tests, which despite not being outright cheating, affects the distribution of test scores and the performance rankings.

The Italian education system recently established standardized tests in mathematics and

[*] Banca d'Italia, via Nazionale 91, 00184 Roma, Italy. This paper was previous circulated with the name *Teachers and Cheaters. Just an Anagram?*. I would like to thank Stéphane Bonhomme, Nicola Curci, Domenico Depalo, Raquel Fernández, Iván Fernández-Val, Marco Savegnago, Paolo Sestito, Martino Tasso, Jeffrey Wooldridge, Stefania Zotteri, and seminar participants at Banca d'Italia, EUI, 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] All the literature of teacher value added rests on the assumption that improvements in students' performance can be attributed to their teachers. See, for instance, Hanushek (1971), Rothstein (2010), Bacher-Hicks et al. (2014) or Chetty et al. (2014).

[2] Grissom et al. (2014).

[3] Throughout this paper I refer to test scores manipulation and cheating as any action done by the students or the teachers that results in a variation of the test scores, usually an increase. This could take 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).

Italian language that are compulsory to all students. Students take these exams in their own schools, proctored by a teacher from their school who was not their teacher during the academic year. These teachers are also 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 had an external monitor instead, who was responsible for the same tasks, but had no prior connection to the school. This constitutes a large scale natural experiment to study test scores manipulation in the absence of an external monitor.

Several academic articles have studied the results of this natural experiment. Quintano et al. (2009) proposed a method to detect and correct for test scores manipulation based on a fuzzy clustering approach using different functions of the results at the classroom level. Bertoni et al. (2013) distinguished between the direct effect of having a monitor in the classroom and the indirect effect, *i.e.* having an external monitor in another classroom of the same school. Moreover, they showed markedly different regional patterns of cheating. Lucifora and Tonello (2015) did an excess variance analysis, finding a large social multiplier for cheating resulting from students' interactions. Angrist et al. (2014) employed an IV strategy based on discontinuities in the class size to find the effect of class size on test scores, which vanished after controlling for cheating. Moreover, they argued that the manipulation is because of teachers' shirking. Battistin et al. (2014) provided bounds for the average test scores net of cheating. Finally, Paccagnella and Sestito (2014) compare the regional incidence of cheating to other social capital measures, finding high correlations.

In this paper there are four contributions. First, I propose a new method to detect test scores manipulation based on a likelihood approach, in which the different questions of the test play the role usually assigned to time in a panel data framework. Second, based on the assessment of the manipulation and its joint distribution with the mean class test scores, I propose a method to correct for cheating. Third, I describe new findings on test scores manipulation patterns, with some demographic groups being more favored than others by the manipulation. Fourth, I propose a binary panel data estimator that incorporates the correlation of the individual unobserved heterogeneity when individuals are split into groups of potentially different size.

The empirical strategy applies nonlinear panel data methods to the natural experiment, controlling for the individual and class effects. In the absence of test scores manipulation, and controlling for the aforementioned effects, individual answers are independent across students. Therefore, unlikely outcomes are associated with low values for the likelihood function of the scores of each class. I consider both fixed and random effects estimation methods. The former do not impose any restriction on the individual fixed effects, but they cannot be used to construct a likelihood function to detect cheating. The latter allow to calculate a likelihood function for each classroom, but at the cost of imposing extra distributional assumptions. Based on this estimator, I propose a cheating correction which is a two step procedure that estimates the amount of manipulation based on how likely the results were at the classroom level in a first step, and then reduces the test scores using the distribution of test scores without manipulation.

The method I propose fits the framework of the INVALSI exams, as it is based in the comparison of the likelihood of the results between two groups, one in which it is assumed that the test scores are fair, and another one in which scores could have been manipulated. The reasoning is that, unlikely results can happen, just not too often. If they take place too often in one group relative to the other one, this supports the hypothesis of test scores manipulation. This contrasts with methods like those proposed by Jacob and Levitt (2003), Quintano et al. (2009), or Battistin et al. (2014), which look for suspicious patterns of answers at the classroom level.[4] If we observe two classrooms in which every student got the maximum score, these methods would identify those scores as manipulated, whereas the method I propose would look at the relative frequency of this event in the two groups, as the scores could simply reflect that the exam was easy or the students performed well. Nevertheless, all these methods have something in common: they use the answers to each

[4] It is also worth emphasizing that Jacob and Levitt (2003) uses data from the same students in the preceding and subsequent academic years, making this method more demanding in terms of data than the other methods.

single item to detect cheating, though in a different manner.

The results produce estimates of substantial test scores manipulation. There is a great variability in the amount of cheating across questions, which is only weakly related to the difficulty of each question, and does not affect much the correlation in students' answers once we control for the mean score. The estimates show manipulation in every region, but increasing as we move south. The manipulation patterns do not support the hypothesis that teacher's objective is to reduce the dispersion of students test scores, as the manipulation tends to favor females, who outperform males in Italian, but are outperformed in mathematics. Immigrants tend to benefit from the manipulation in Italian but not so much in mathematics.

2 Italian National Evaluation Test

INVALSI is the Italian institute responsible for the design and administration of annually standardized tests for Italian students. It was created in 1999, and in the academic year 2008/09 these tests acquired nationwide status. All students enrolled in some grades were required to take two tests, one in mathematics and another one in Italian language.[5] Even though 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*), which corresponds to the end of the compulsory secondary education, and the results of the test account for a proportion of their final marks.

The way these exams are proctored provides a natural experiment to assess test scores manipulation. A set of randomly selected classes were assigned an external monitor,[6] who

[5] The grades whose students were tested were not the same every year. For the year 2012/2013, the test was taken by students enrolled in 2nd, 5th, 6th, 8th and 10th grade.

[6] The selection mechanism is the same used by the IEA-TIMSS survey. In a first stage, a fixed number of schools from each region are selected at random. In a second stage, one or more classrooms from each of the selected schools is selected at random. Monitors were teachers and principals who had not worked in the town of the school they were assigned for at least two years before the exam. Some, but not all of these teachers were retired, while others were *precari*, *i.e.* teachers with no tenure position.

would be present during the exam and then transcript the results of each student to a sheet which is finally sent to INVALSI.[7] In the remaining classes a teacher from a different class in the same school proctored the exam and transcribed the results. The treatment group is composed of those students whose exams were proctored by an external monitor, and the remaining students conform the control group. Teachers, unlike external monitors, may have incentives to manipulate test scores: despite the low stakes nature of the exam, and although their salaries are not linked to the exam results, teachers may perceive that their school or themselves can be evaluated based on the results.[8] Moreover, INVALSI reveals to each school their own results, which principals can make public to entice parents to enroll their children in their school. Finally, it is possible that cheating reflects a cultural component, as it occurs more frequently in areas that display low values of several measures of social capital (Paccagnella and Sestito, 2014).

2.1 Data and Descriptive Statistics

Altogether, for the academic year 2012/13, the experiment involved over 2.3 million students, of which over 143,000 were assigned to the treatment group. Table 1 shows the number of students and classes assigned to both the treatment and control groups for each grade. It also shows the mean percentage of correct answers for both the treatment and control groups in all ten exams, which was higher for the control group, *i.e.* the presence of an external monitor reduces the scores. The difference between the two groups varies across grades and is larger for the mathematics exam.

I focus my analysis on the 10th graders mathematics exam[9]. A total of 38,273 students in 2,203 classes were assigned an external monitor, whereas the remaining 382,259 students in 21,599 classrooms conform the control group. The total number of questions in this exam was

[7] Some, but not all questions were multiple choice, so this task cannot be automatically done by a machine.

[8] External monitors were paid between 100 and 200 EUR for the job. Since they can be asked in the subsequent years to monitor more exams, they have incentives to grade fairly.

[9] 10th graders constitute the largest treatment group, and the percentage of manipulation is larger for the mathematics exam. Many results are qualitatively similar for all exams, and differences across exams are also reported in this paper.

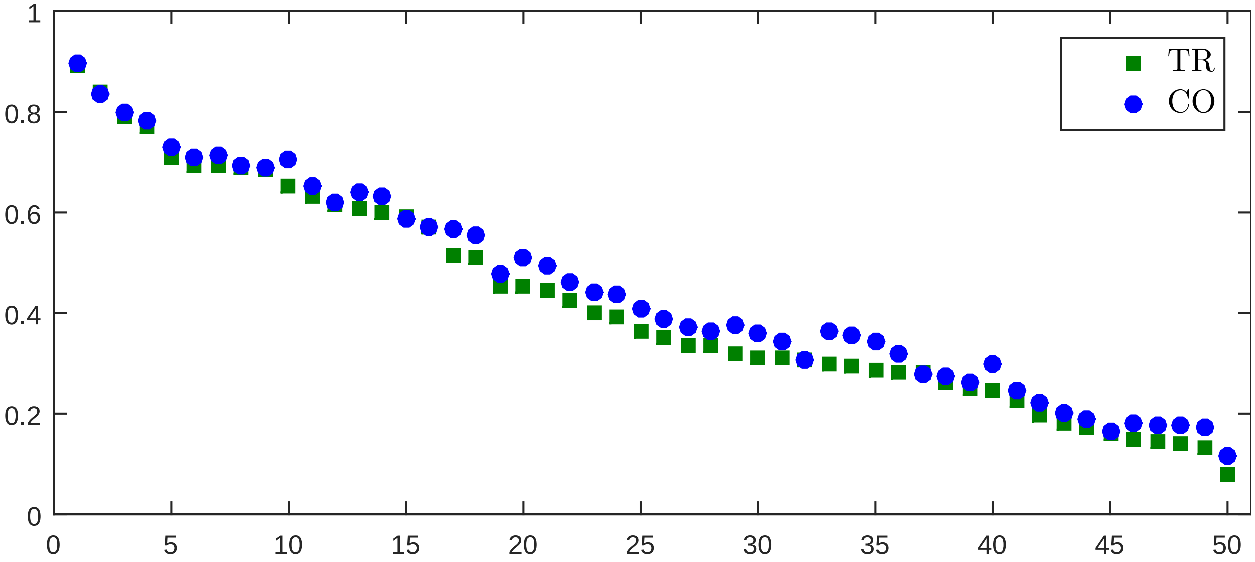
Table 1: Size of the treatment and control groups, academic year 2012/13

	2nd grade		5th grade		6th grade		8th grade		10th grade	
	TR	CO	TR	CO	TR	CO	TR	CO	TR	CO
N	25070	474784	24773	460452	27504	456937	28153	490304	38273	382259
C	1424	27555	1426	27782	1457	24164	1464	25546	2203	21599
S	737	7104	736	7075	732	5797	1416	5910	1094	4290
% Correct (Mathematics)	53.87	61.06	54.79	59.47	44.53	45.27	50.83	52.48	42.09	44.94
% Correct (Italian)	59.90	64.63	74.36	76.79	64.25	64.42	72.44	73.21	64.20	65.57

N, C and S respectively denote the number of students, classrooms and schools, and TR and CO respectively denote the treatment and control groups. A treated school is defined as a school with at least one treated class, and a control school is similarly defined.

50. Figure 1 shows the proportion of students who answered each question correctly. Even though there is a lot of variability across answers, the treatment group scored worse than the control group in all but three of the questions, suggesting that there was some manipulation of the scores. Both difficult and easy questions can have large or small differences between the two groups, although there is a weak correlation between the difficulty of a question and the difference between the two groups.[10]

Figure 1: Proportion of correct answers by question, 10th grade mathematics exam



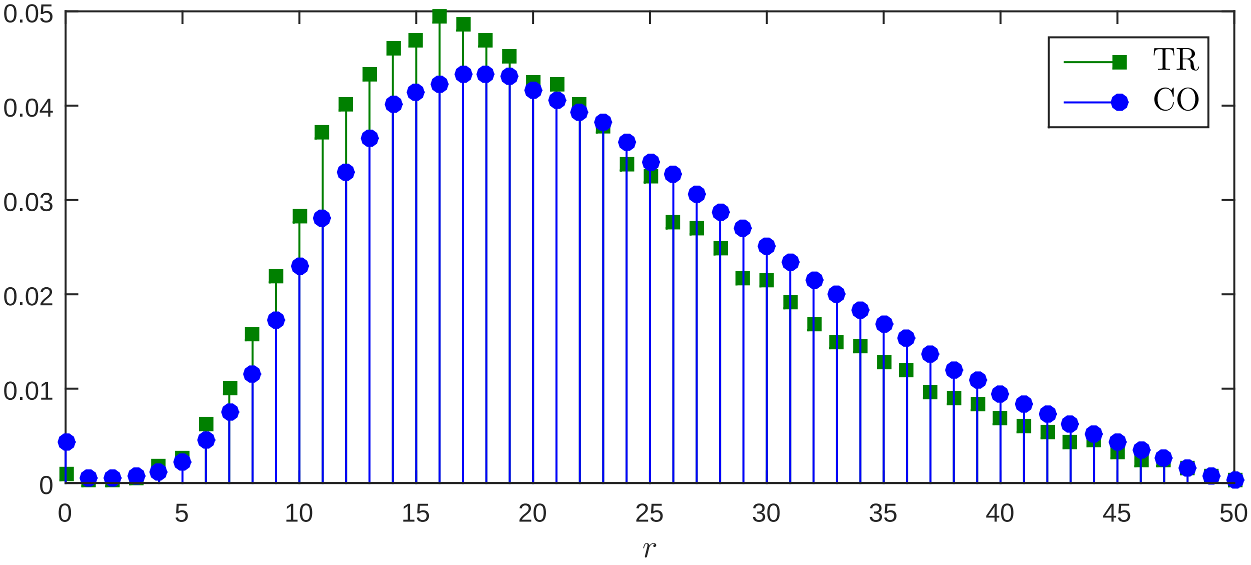
Questions are sorted by how frequently they were correctly answered by students in the treatment group.

The increase in the proportion of students answering each question correctly is reflected in the change of the distribution of the total number of correct answers, which is shown

[10] This result is, however, not consistent across exams, and for some of them there is no correlation.

in figure 2. There is a change in the mass to the right, with the mode increasing from 17 to 18, the mean from 21.05 to 22.47, and the median from 20 to 21. The majority of the change takes place around the center of the distribution, whereas the tails show a change much smaller in magnitude. Unsurprisingly, since this is a low stakes exam, the change is quite smooth.[11]

Figure 2: Distribution of test scores, 10th grade mathematics exam



Quintano et al. (2009) approach to detect cheating is the method currently used by INVALSI to correct for cheating. This method 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. This method has a few issues, as the distributions of these statistics, and even their support, depend on the number of questions and students in the class.[12] Moreover, the distribution of these statistics depend on each other in a nontrivial way.[13]

[11] There is not a sudden decrease of the frequency below a cutoff score, followed by an even larger increase above it, which could have been caused if students' academic outcomes depended on attaining a particular score. See Dee et al. (2011) for such an example.

[12] 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 $\{0, {}^2\!/_9, {}^6\!/_9, {}^8\!/_9\}$.

[13] For example, as the mean approaches either zero or the maximum, the variance converges to zero, but it can take a variety of values in between.

Further, these statistics may not reflect cheating accurately: a high mean could imply good that students' ability is high, and a high correlation could be caused by sorting or peer effects. Finally, if these statistics are not sufficient, some information is missing.

To illustrate these points, consider the mean and the variance of the test scores. There is strong evidence that the variance decreased in the presence of an internal monitor, but this was accompanied by an increase in the mean (Bertoni et al., 2013), making them not directly comparable by construction. Therefore, to evaluate the possibility of using the variance to detect cheating, it is necessary to control for the mean. Consider an alternative statistic: the probability of two students correctly answering the same s questions, conditional on them having correctly answered $\overline{r}$ and $\underline{r}$ answers, respectively. If students copied each other, or if a teacher graded students systematically giving particular grades to some answers, then this statistic, which controls for the mean test scores, should be substantially different for the treatment and control groups. This is estimated by

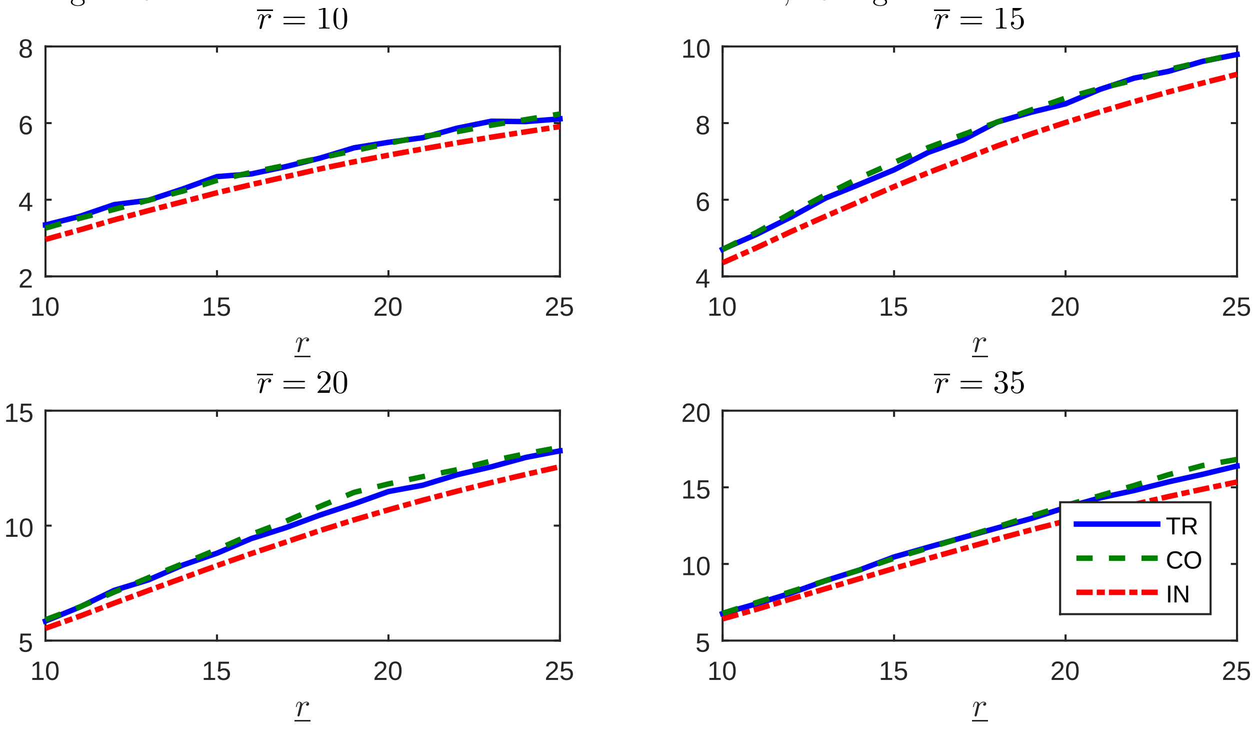
$$\mathbb{P}_n\left(s|\overline{r},\underline{r}\right) \equiv \frac{\sum_{c=1}^{C}\sum_{i=1}^{N_c}\sum_{j\neq i}\mathbf{1}\left(r_i=\overline{r},r_j=\underline{r},\underline{r}_{ij}=s\right)}{\sum_{c=1}^{C}\sum_{i=1}^{N_c}\sum_{j\neq i}\mathbf{1}\left(r_i=\overline{r},r_j=\underline{r}\right)} \tag{1}$$

Based on this statistic, one can estimate the mean number of correct answers in common when two students have $\overline{r}$ and $\underline{r}$ correct answers, *i.e.* $\mathbb{E}_n\left[s|\overline{r},\underline{r}\right]=\sum_{q=0}^{Q}q\mathbb{P}_n\left(q|\overline{r},\underline{r}\right)$. Figure 3 shows the values of this statistic for different values of $(\overline{r},\underline{r})$, both for students in the same classroom in each of the two groups, and for students who are in different classrooms. Unsurprisingly, this conditional mean is uniformly larger for students in the same classroom relative to students in different classrooms. What is more surprising, is that the difference between the treatment and control groups is tiny, and in fact, this mean is larger for the treatment than for the control group for some values of $(\overline{r},\underline{r})$. Hence, it could be argued that once we control for the mean, the variance has little cheating detection power.

3 Empirical Strategy

There are several factors other than cheating that can affect the student test scores, which can be split into three main categories: individual characteristics, class characteristics and

Figure 3: Mean number of correct answers in common, 10th grade mathematics exam



TR, CO and IN respectively denote the mean number of correct answers in common of two students with $\overline{r}$ and $\underline{r}$ correct answers when they are in the same class in the treatment group, in the same class in the control group, or in different classes in either group.

question characteristics. Each factor operates in a different manner: individual characteristics affect a particular student but not his classmates;[14] class characteristics are more broadly defined, since they include anything that can affect several or all students in the same classroom;[15] question characteristics refer to the factors that, while being the same to all students, they vary across questions.[16]

Hence, a school in which students are segregated based on their past performance, even in the absence of cheating, one would have a high degree of correlation in the answers of students in the same classroom. Similarly, if a question is extremely difficult, the majority of the students will answer incorrectly that question, creating a high correlation in the answers for that particular question, not just within each classroom, but also between classrooms. These factors also affect the class mean test score, the within variance of correct answers, the frequency of each question being correctly answered, etc. The cheating detection method I propose is based on the differences between the treatment and control groups after we have controlled for the aforementioned effects in the absence of cheating, thus allowing for the possibility of good results being a product of good performance, and not necessarily cheating.

Empirical studies on education often rely on tests scores as a measure of students' attainments, and are frequently standardized to have zero mean and unit standard deviation.[17] This global measure may not be however the most appropriate in order to detect cheating. Having information on the results of every single test item is much more informative, because it allows to look at a richer correlation structure in the answers. As Jacob and Levitt (2003) showed, if teachers inflate students grades in a systematic way, having the results of every test item allows to detect cheating with a higher degree of confidence than if the data at disposal comprises only the total score of the test.

From an econometric perspective, using the normalized test scores implies using a cross section with a group correlation structure. However, if we do not normalize them, we have

[14] For example: intelligence, motivation, hours devoted to study, or individual attendance.

[15] For example: effectiveness of the teacher, quality of the peer group composition (*i.e.* peer effects), physical characteristics of the classroom, or sorting of students (both at the classroom and school levels)

[16] For example: difficulty of each question, or the amount of space to answer the question

[17] The logic behind this is to express the estimated coefficients in a comparable unit across studies.

a panel data in which instead of seeing the performance of students over time, we see their performance over questions. Moreover, since all students answered every question of an exam at the same time, all the individual and class characteristics are constant across questions, *i.e.* $x_{icq} = x_{ic} \forall q = 1, ..., Q$. Given that the questions in the INVALSI tests are corrected on a right/wrong basis (*i.e.*, no partial grading of any question), the data takes the form of a binary panel. The empirical strategy of the paper is therefore to specify a likelihood function that combines the individual, class, and question effects. In the absence of test scores manipulation (treatment group), since all the interaction among students and with the teachers took place before the exam (and the questions of the exam were unknown to both students and teachers until the day of the exam), once we control for the individual and class effects, the answers between students are independent. If some manipulation took place, then this should be reflected in the estimates of the control group.

3.1 Conditional Likelihood (Fixed Effects) Approach

Let y_{icq} denote the dummy variable that takes value one if student i in classroom c answered correctly question q in the exam. This variable depends on a latent variable, y^*_{icq}, which in turn depends on three unobserved variables: an individual-class fixed effect, η_{ic}, a question fixed effect, ξ_q, and a specific individual-class-question *iid* shock, ε_{icq}.[18] Let the relation between these variables be given by

$$\begin{aligned} y_{icq} &= \mathbf{1}\left(y^*_{icq} \geq 0\right) \\ y^*_{icq} &= \eta_{ic} + \xi_q + \varepsilon_{icq} \end{aligned} \tag{2}$$

where the number of questions (Q) is fixed, but the number of students (N) and classroom (C) is large. It is therefore not possible to obtain consistent estimates of the individual-class fixed effects, since both the number of questions and the number of students per class are finite, but it is possible for the question fixed effects. Under the assumption that ε_{icq} is logistically distributed, one can follow Chamberlain (1980) to overcome the incidental

[18] The parameter ξ_q may also capture the location of the question in the exam. However, several versions of the exam were provided, with the only difference among them being the ordering of the questions. Unfortunately, the version assigned to each student is not recorded in the dataset, and hence it is not possible to estimate if questions asked at the end of the exam are not answered correctly more often.

parameter problem, obtaining estimates of the question fixed effects.[19]

Because of multicollinearity, it is necessary to exclude one of the question fixed effects. Then, the interpretation of the remaining $Q-1$ fixed effects is the difficulty of question q relative to the excluded question. In other words, we 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.* $B_r \equiv \left\{ b : \sum_{q=1}^{Q} b_q = r \right\}$.[20] Under the assumption of no cheating, once the individual-class effects are accounted for, the answers of two students are independent. Hence, the log-likelihood function is given by

$$\mathcal{L}(\xi) = \sum_{c=1}^{C} \sum_{i=1}^{N_c} \log \left[\mathbb{P}(y_{ic} | r_{ic}) \right] \ = \ \sum_{c=1}^{C} \sum_{i=1}^{N_c} y'_{ic}\xi - \sum_{c=1}^{C} \sum_{i=1}^{N_c} \log \left[\sum_{b \in B_{r_{ic}}} \exp(b'\xi) \right] \qquad (3)$$

Equation 3 is a conditional likelihood, so it cannot be directly used to detect cheating, as the distribution of the total number of correct answers, r_{ic}, is different for the treatment and control groups. Moreover, because all covariates are constant across questions, it is not possible to see which characteristics are predictive of test scores manipulation. A random effects approach can overcome these two issues at the cost of imposing extra parametric restrictions.

3.2 Random Effects Approach

Taking equation 2 as our baseline, a random effects estimator models the distribution of the individual-class effects, leaving the question effects as the parameters to be estimated. Denote by F_ε the cdf of ε_{icq}, by $y_c \equiv (y_{1c1}, ..., y_{1cQ}, ..., y_{N_c cQ})$ the vector with the results of all students in classroom c. Moreover, assume that the distribution of the unobservables are given by $\varepsilon_{icq} \sim \mathit{Logistic}(0,1)$, and $\eta_{ic} \sim \mathcal{N}\left(0, \sigma_\eta^2\right)$. In the absence of correlation across

[19] 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.

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

students in the same classroom, this model is a random effects logit with normally distributed random effects, whose likelihood is given by

$$\mathcal{L}\left(\theta\right) = \sum_{c=1}^{C}\sum_{i=1}^{N_c}\log\left(\int_0^1 \frac{\exp\left(\sum_{q=1}^{Q} y_{icq}\left(\eta_{ic}+\xi_q\right)\right)}{\prod_{q=1}^{Q}\left(1+\exp\left(\eta_{ic}+\xi_q\right)\right)}d\Phi\left(\frac{\eta_{ic}}{\sigma_\eta}\right)\right) \tag{4}$$

The independence of the individual-class effects is highly unrealistic, as sorting of students, peer effects, or sharing the same teacher can create a correlation in these effects. One possibility to model this correlation would be to impose a parametric joint distribution on the individual effects. An alternative would be to separately model the marginal distribution of each individual effect and their correlation structure using a copula.[21] An advantage of this approach is that the copula depends on the ranks of the individual effects, $u_{ic} \equiv F_\eta\left(\eta_{ic}\right)$, which do not depend on the parameters of the marginal distribution of η_{ic}. Moreover, copulas can conveniently handle differences in the dimensionality that arise from having classes of different size. Denote by η_c and u_c the vectors of dimension N_c of individual effects and their ranks in class c. If the cdf of the copula is given by $C\left(u_c;\rho\right)$, the likelihood function is written as[22]

$$\mathcal{L}\left(\theta\right) = \sum_{c=1}^{C}\log\left(\int_0^1 \frac{\exp\left(\sum_{i=1}^{N_c}\sum_{q=1}^{Q} y_{icq}\left(\eta_{ic}+\xi_q\right)\right)}{\prod_{i=1}^{N_c}\prod_{q=1}^{Q}\left(1+\exp\left(\eta_{ic}+\xi_q\right)\right)}dC\left(u_c;\rho\right)\right) \tag{5}$$

Finally, note that the random effects estimator permits the inclusion of question-invariant covariates, as well as interactions between the questions and the covariates.[23]

3.3 Assessment of Cheating

When trying to detect cheating there are two candidates for the unit of analysis, the classroom and the individual students. As I argued above, and also motivated by the findings in this paper and in Angrist et al. (2014), teachers may play an important role in cheating, and therefore I consider the classroom as the unit of analysis. Assume that the students $i = 1, ..., N_c$ in classroom c obtained a total number of correct answers of $r_{1c}, ..., r_{N_c c}$, and

[21] As proved by Sklar (1959), any multivariate cdf can be written as a copula whose arguments are the marginal distributions, *i.e.* $\mathbb{P}\left(X_1 \leq x_1, ..., X_d \leq x_d\right) = C\left(F_1\left(x_1\right), ..., F_1\left(x_1\right)\right)$.

[22] The details of the estimation are presented in appendix B.

[23] See appendix A for generalizations to other distributions for the unobservables and the inclusion of covariates.

collapse them into vector r_c. Using equation 5, the probability of each student obtaining at least as many correct answers as they actually got is given by[24]

$$\mathbb{P}\left(R \geq r_c\right) \;=\; \sum_{b_1 \in \overline{B}_{r_{1c}}} \cdots \sum_{b_{N_c} \in \overline{B}_{r_{N_c c}}} \int_{[0,1]} \frac{\exp\left(\sum_{i=1}^{N_c} \sum_{q=1}^{Q} b_{iq}\left(\eta_{ic}+\xi_q\right)\right)}{\prod_{i=1}^{N_c} \prod_{q=1}^{Q}\left(1+\exp\left(\eta_{ic}+\xi_q\right)\right)} dC\left(u_c;\rho\right) \tag{6}$$

where $\overline{B}_r \equiv \left\{b : \sum_{q=1}^{Q} b_q \geq r\right\}$. This likelihood presents a problem of comparability across classes, since class size is not constant.[25] To avoid this problem, I use the estimated geometric mean of the probability to detect the likelihood of cheating[26]

$$\hat{l}_c \;=\; \left[\sum_{b_1 \in \overline{B}_{r_{1c}}} \cdots \sum_{b_{N_c c} \in \overline{B}_{r_{N_c}}} \int_{[0,1]^{N_c}} \frac{\exp\left(\sum_{i=1}^{N_c} \sum_{q=1}^{Q} b_{iq}\left(\hat{\eta}_{ic}+\hat{\xi}_q\right)\right)}{\prod_{i=1}^{N_c} \prod_{q=1}^{Q}\left(1+\exp\left(\hat{\eta}_{ic}+\hat{\xi}_q\right)\right)} dC\left(u_c;\hat{\rho}\right)\right]^{\frac{1}{N_c}} \tag{7}$$

3.4 Cheating Correction

The correction for manipulation proposed in this paper is split into three steps: the first one involves changing the estimated likelihood of each classroom in the control group in such a way that the resulting distribution matches the distribution of the treatment group; the second step uses this likelihood and the observed test scores to calculate the correction at the class level; the third step individualizes the correction by equalizing the within class variances of the two groups. Denote by $F_{\mathcal{L},j}\left(l\right)$ the cdf of the likelihood for the treatment ($j=1$) and control ($j=0$) groups. The corrected likelihood, $\check{l}_c$, is given by

$$\check{l}_c \equiv F_{\mathcal{L},TR}^{-1}\left(F_{\mathcal{L},CO}\left(\hat{l}_c\right)\right) \tag{8}$$

In words, the cdf of the corrected likelihood of the classes in the control group equals the cdf of the treatment group by construction. Graphically, it involves a nonlinear horizontal shift of the cdf for the control group. The second step corrects the test scores based on the corrected likelihood, for which we make the following assumption:

Assumption 1. *Distribution of the test scores manipulation*

Let $\overline{r}_c^*$ denote the observed mean test score of classroom c in the control group. This score

[24] In this case, the number of permutations is given by $\prod_{i=1}^{N_c} \sum_{s=r_{ic}}^{Q} \binom{Q}{s}$.

[25] To see this more clearly, as class size increases, the likelihood goes to zero, as equation 6 becomes an infinite products of terms bounded between 0 and 1.

[26] See appendix C for the details on the computation of the sum of all possible permutations.

is decomposed as the sum of the score without manipulation, $\overline{r}_c$, and the manipulation, η_c. These two components are mutually independent and the distribution of the manipulation is given by an exponential(λ) distribution.

This assumption allows us to estimate $\mathbb{E}\left[r|r^*,\check{l}\right]$, which is the corrected test score.[27] The idea is similar to Wei and Carroll (2009), whose estimator of quantile regression with measurement error is adapted to the current framework:

$$\mathbb{E}\left[r|r^*,\check{l}\right] \;=\; \frac{\int_0^{r^*} rf\left(\check{l}|r\right)\lambda\exp\left(-\lambda\left(r^*-r\right)\right)dF\left(r\right)}{\int_0^{r^*} f\left(\check{l}|r\right)\lambda\exp\left(-\lambda\left(r^*-r\right)\right)dF\left(r\right)} \tag{9}$$

where the equality follows by Bayes' theorem and the independence between test scores and manipulation stated in assumption 1. Equation 9 suggests the following sample analogue to estimate the corrected test scores:

$$\tilde{r} \;\equiv\; \frac{\frac{1}{\sum_{c=1}^{C_0}\mathbf{1}(r_c\leq r^*)}\sum_{c=1}^{C_0} r_c\hat{f}\left(\check{l}|r_c\right)\hat{\lambda}\exp\left(-\hat{\lambda}\left(r^*-r_c\right)\right)}{\frac{1}{\sum_{c=1}^{C_0}\mathbf{1}(r_c\leq r^*)}\sum_{c=1}^{C_0} \hat{f}\left(\check{l}|r_c\right)\hat{\lambda}\exp\left(-\hat{\lambda}\left(r^*-r_c\right)\right)} \tag{10}$$

where $\hat{f}\left(\check{l}|r\right) = \sum_{k=1}^{K}\frac{\tau_{k+1}-\tau_k}{\hat{Q}_L(\tau_k|r)-\hat{Q}_L(\tau_{k+1}|r)}\mathbf{1}\left(\hat{Q}_L\left(\tau_k|r\right)<\check{l}\leq\hat{Q}_L\left(\tau_{k+1}|r\right)\right)$, and $\hat{Q}_L\left(\tau|r\right)$ is estimated by using linear quantile regression on a polynomial of r and applying Chernozhukov et al. (2010) rearrangement, and $\hat{\lambda}$ is estimated using the method of moments.[28] Finally, to extend the correction at the individual level, I denote the ratio of the within class variances of the treatment and control groups as $\hat{\Sigma}_W$, and then the corrected test score for student i in classroom c is given by

$$\tilde{r}^*_{ic} = \tilde{r}_{ic} + \left(\hat{\Sigma}_W - 1\right)\left(r_{ic}-\overline{r}_c\right)\mathbf{1}\left(r_c\neq\tilde{r}_c\right) \tag{11}$$

In words, when the class was applied a positive correction in the second step, if the within variance is smaller with manipulation, then those students with a higher than average test score are corrected less than the average of the class, and those with a score below the average are corrected more than the average. If the ratio of variances equals one, then the correction remains the same for all students in the classroom.

[27] 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.

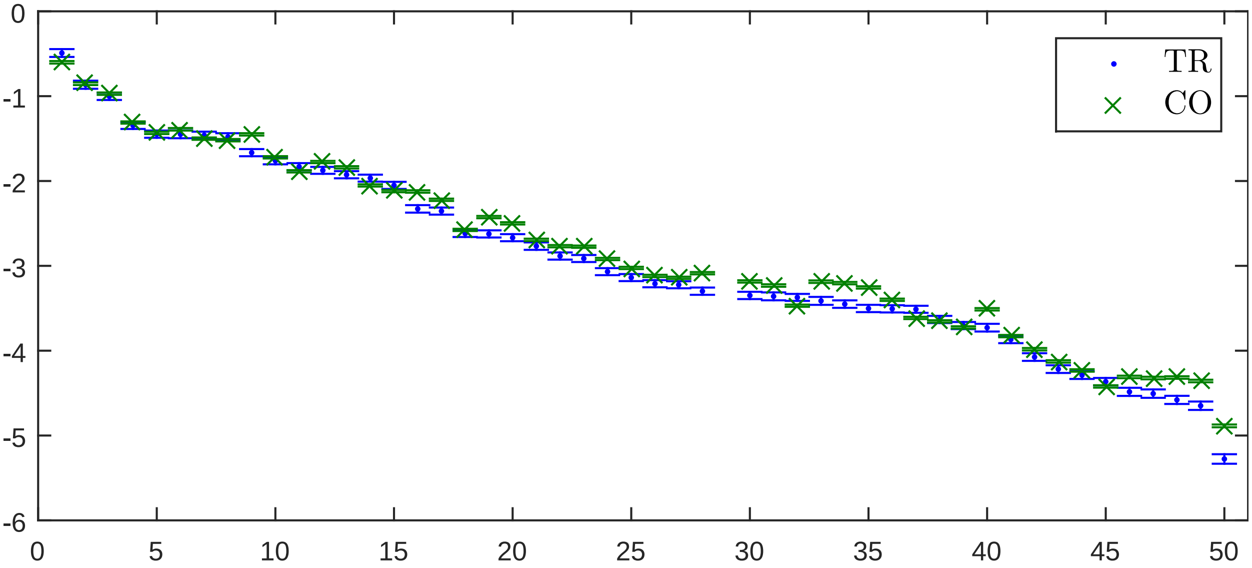
[28] Although the current framework does not prevents us from using maximum likelihood to estimate this parameter, the method of moments is more convenient, as it does not require to plug in the estimate of the marginal distribution of r.

4 Results

4.1 Conditional Fixed Effect Logit Estimates

Figure 4 shows the conditional fixed effect logit estimates of ξ for the mathematics exam of 10th graders.[29] 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.

Figure 4: Conditional FE logit estimates, 10th grade mathematics exam



TR and CO respectively denote the coefficients for the treatment and control groups. They are reported with the 95% confidence intervals, and are sorted by the proportion of students who answered them correctly in the treatment group.

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, if we compare the male and female estimates for each groups, and they show that even in the absence of manipulation, there are remarkable gender differences in performance, with

[29] 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.

male students performing relatively better than females for 17 questions, and the other way around for 16 questions.[30] 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. This result is robust to all exams, but not to all possible categories.[31] 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,[32] and if we consider the three macro regions of Italy, we observe large differences between the estimates for the control groups in the North and South & Islands regions.[33]

4.2 Random Effects Logit Estimates

Figure 5 shows the random effects logit estimates of ξ when the copula is independent and no covariates are included. 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 6 out of the remaining 12, they are not significantly different. Moreover, although the coefficients are not directly comparable to the estimates shown in figure 4, 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 does not play a big role in determining the value of the coefficients.[34]

Table 2 shows the coefficients of the remaining parameters for several specifications.[35]

[30] Previous studies (Machin and Pekkarinen, 2008; Lavy and Sand, 2015) have shown differences in absolute performance between male and female students, but these results indicate that these differences are also very heterogeneous for different concepts.

[31] See tables 10 to 12 in appendix D.

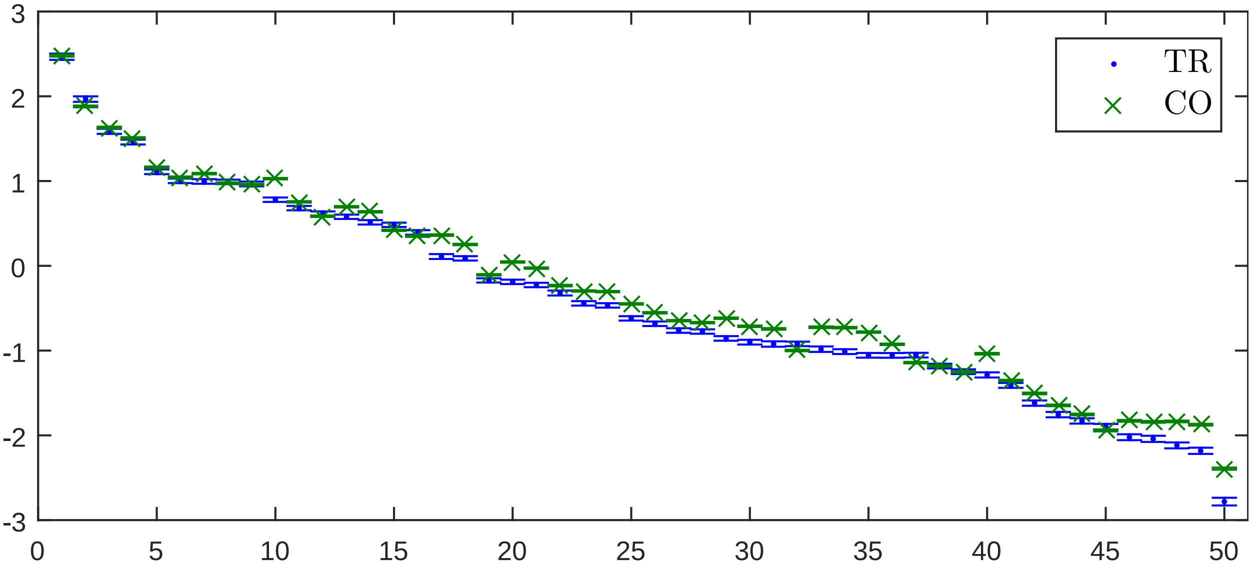
[32] I consider two groups: those with a class size equal to or larger than the median for each grade (LARGE), and those with a smaller one (SMALL).

[33] I consider 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).

[34] On the other hand, ignoring the unobserved heterogeneity results in significantly biased estimates, as shown in table 15 in appendix D.

[35] Further specifications including a polynomial of class size, or an interaction between small class and regional dummies yield similar results and are omitted. These specifications and the estimates of the other 9 exams are available upon request.

Figure 5: RE logit estimates, 10th grade mathematics exam



TR and CO respectively denote the coefficients for the treatment and control groups. They are reported with the 95% confidence intervals, and are sorted by the proportion of students who answered them correctly in the treatment group.

Consistent with previous findings, the coefficients for females, Center, and South & Islands are negative both for the treatment and control groups, but the difference between these two indicates that the manipulation favored those groups. Note that after controlling for the regional dummies and the female-question interactions, the difference between the mean of the question effects of the treatment and control groups vanishes. Natives performance is also higher, and the manipulation of test scores slightly favors them in this exam, though not significantly. The performance of students in smaller classrooms is statistically higher to those in large classrooms, but the difference between the treatment and control groups is again insignificant. Notice also that the estimate of the standard deviation of the individual random effects is large in all specifications, highlighting the importance of considering the student effects both for the estimation of the parameters and the correction of the test scores. The last two rows show the estimates of the random effects logit model with the copula, which we assume is a $Clayton\left(\rho\right)$ copula. ρ is not interpreted as the linear correlation parameter, and using the relation between the Clayton and Gaussian copulas with Kendall's τ statistic.[36]

[36] For the Gaussian copula, $\rho = \sin\left(\frac{\pi}{2}\tau\right)$, and for the Clayton copula, $\rho = \frac{\tau}{2+\tau}$.

the linear correlation for the treatment and control groups is approximately 0.54 and 0.49, indicating that there is a small, yet significant excess of correlation for the treatment group relative to the control group.

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

		$\hat{\bar{\xi}}$	FE	CE	SI	IT	$SMALL$	$\hat{\sigma}_\eta$	$\hat{\rho}$
Treatment	(1)	-0.40***	-	-	-	-	-	0.95***	-
		(0.01)						(0.01)	
	(2)	-0.27***	-0.42***	-	-	-	-	0.96***	-
		(0.01)	(0.01)					(0.01)	
	(3)	0.06***	-0.25***	-0.29***	-0.69***	-	-	0.90***	-
		(0.01)	(0.01)	(0.01)	(0.01)			(0.01)	
	(4)	-0.01	-0.40***	-0.28***	-0.67***	0.36***	0.06***	0.87***	-
		(0.02)	(0.01)	(0.01)	(0.01)	(0.02)	(0.00)	(0.01)	
	(5)	-0.39***	-	-	-	-	-	0.87***	2.93***
		(0.00)						(0.00)	(0.02)
	(6)	0.07***	-0.37***	-0.29***	-0.71***	0.31***	-0.51***	0.81***	2.42***
		(0.01)	(0.00)	(0.00)	(0.01)	(0.01)	(0.00)	(0.00)	(0.01)
Control	(1)	-0.29***	-	-	-	-	-	1.01***	-
		(0.00)						(0.00)	
	(2)	-0.12***	-0.32***	-	-	-	-	0.99***	-
		(0.00)	(0.00)					(0.00)	
	(3)	0.07***	-0.31***	-0.21***	-0.40***	-	-	0.98***	-
		(0.00)	(0.00)	(0.00)	(0.00)			(0.00)	
	(4)	-0.05***	-0.33***	-0.19***	-0.42***	0.40***	0.06***	0.95***	-
		(0.01)	(0.00)	(0.00)	(0.00)	(0.01)	(0.00)	(0.00)	
	(5)	-0.18***	-	-	-	-	-	0.92***	2.34***
		(0.00)						(0.00)	(0.00)
	(6)	0.10***	-0.24***	-0.17***	-0.43***	0.26***	-0.52***	0.90***	1.93***
		(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)

FE represents the average of the interaction between a dummy for female students and the question dummies, CE, SI, IT, and SMALL are dummies for Center region, South & Islands region, natives, and small class. *, **, and *** denote statistically significant at the 90%, 95%, and 99% level, respectively, standard errors in parentheses.

Most of the results, but not all of them apply to all exams in the sample. Table 3 summarizes the differences in performance between the treatment and control groups for all exams: a negative sign indicates that the manipulation in the control group favored more those students that belong to that category, whereas a positive sign indicates it was detrimental to them. The single most important variable is the dummy for the South & Islands region, which is significantly negative in all exams. There is also more manipulation in the Center than in the North in eight of the exams, and the only one in which it is the other way around, the coefficient is very close to zero. Test scores of female students were also more manipulated than their male counterparts in every exam, and this difference tends to be smaller than the geographic differences, though not always. Hence, even if one argued that the manipulation favors them in the mathematics exam because they have a worse average performance than male students, the reverse should happen in the Italian exams, where female students consistently have a better average performance. Test scores of immigrant students are more manipulated in seven of the exams. Interestingly, five of these were the Italian exams, which could mean that teachers are trying to compensate for the handicap immigrants face by having to learn the local language. Finally, if anything, manipulation is negatively correlated with class size, as in six of the exams the manipulation was larger in classrooms of size smaller than the median, and the reverse was true in two exams.

Table 3: Summary RE logit estimates

	2nd grade		5th grade		6th grade		8th grade		10th grade	
	M	I	M	I	M	I	M	I	M	I
FE	-,***	-,***	-,***	-,***	-,***	-,***	-,***	-,***	-,***	-,***
CE	-,***	-,***	-,***	-,***	0	-,***	-,**	+,***	-,***	-,***
SI	-,***	-,***	-,***	-,***	-,***	-,***	-,***	-,***	-,***	-,***
IT	0	+,***	0	+,***	+,***	+,***	0	+,***	+,***	+,***
$SMALL$	-,***	-,***	-,***	-,***	+,***	-,***	+,***	-,***	0	-,***

FE, CE, SI, IT, and SMALL are dummies for females, Center region, South & Islands region, natives, and small class. A minus sign denotes that the different between the coefficients was significantly smaller for the treatment group, a positive sign denotes it was significantly larger, and a 0 denotes they were significantly equal. *, **, and *** denote statistically significant at the 90%, 95%, and 99% level, respectively.

Questions in INVALSI exams can be split into two main categories: multiple choice questions, and open ended questions, which can help us uncover the source of manipulation. Multiple choice questions require minimal effort to grade and transcript, and at the same time students would find it easier to copy from a fellow classmate. On the other hand, open ended questions can involve an elaborate answer which takes more time to grade and students may find it harder to copy. Even though INVALSI provides a grid for correction of the answers, open ended questions can be interpretable, giving the monitor more discretion in judging whether the answer is right or wrong.

Table 4 shows that both types of questions suffer from manipulation, and with the exception of both exams in 8th grade, the difference between the control and treatment groups' estimates is larger for open ended questions.[37] However, the patterns in missing answers is actually the opposite: for the control group, the proportion of missing answers decreases more for the open ended questions than for the multiple choice questions. If we assumed that students copied each other during the exam, it would lead to the opposite result, since it is easier to copy a multiple choice question than an open ended one. This evidence supports the hypothesis that teachers are more responsible than students in the manipulation of the test scores.

Table 4: Multiple choice vs open ended questions

	2nd grade		5th grade		6th grade		8th grade		10th grade	
	M	I	M	I	M	I	M	I	M	I
$\Delta_{Y,MC}$	-0.31	-0.07	-0.36	-0.08	-0.10	-0.05	-0.20	-0.06	0.08	-0.17
$\Delta_{Y,OE}$	-0.41	-0.19	-0.59	-0.10	-0.29	-0.03	-0.31	-0.25	-0.04	-0.26
DID_Y	0.10	0.12	0.23	0.03	0.20	-0.01	0.10	0.18	0.11	0.09
$\Delta_{M,MC}$	1.18	-0.07	1.04	0.21	0.21	-0.03	0.51	-0.12	-0.10	-0.37
$\Delta_{M,OE}$	2.18	-0.56	3.30	0.28	0.84	-0.05	1.00	2.17	-0.35	-0.04
DID_M	-1.01	0.49	-2.26	-0.08	-0.63	0.02	-0.49	-2.29	0.24	-0.32

$\Delta_{Y,MC}$ and $\Delta_{Y,OE}$ respectively denote the mean difference between the treatment and control groups of the RE logit estimates for multiple choice and open ended questions; DID_Y 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 multiple choice and open ended questions; DID_M denotes the difference between these two.

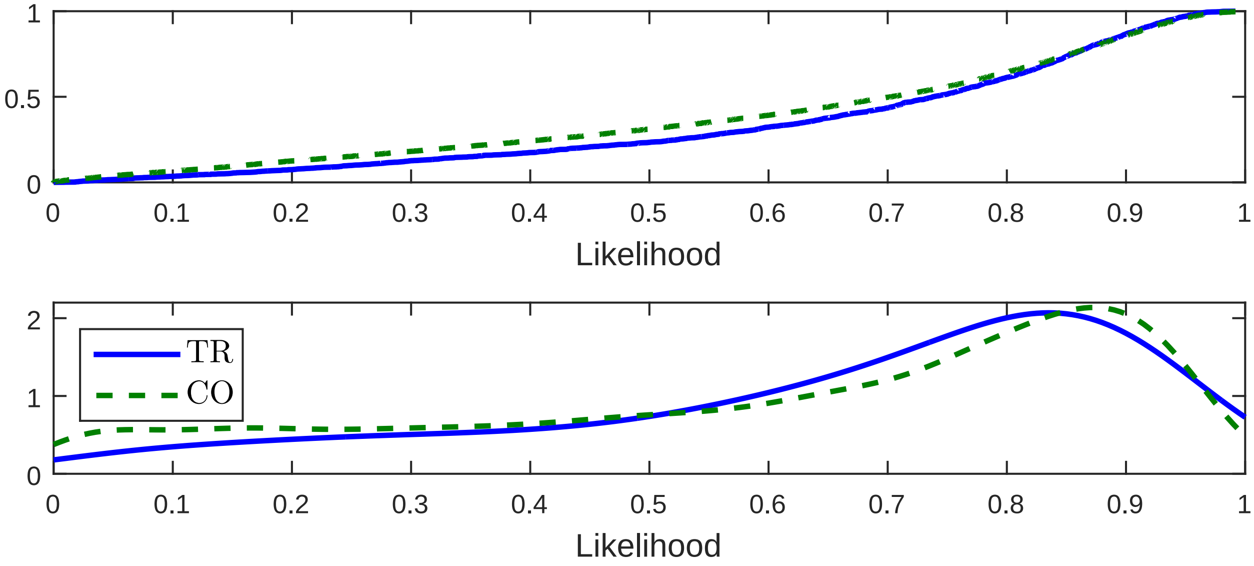
5 Assessment of Cheating

Before correcting the exams for cheating, compare the likelihood of the results for the treatment and control groups. Figure 6 shows both the empirical cdf and a kernel density estimate of the pdf of the likelihood, based on the estimates of the fourth specification of the random effects logit estimator. Unsurprisingly, the two distributions do not coincide. The right tail, representing those classes less likely to have cheated, is approximately the same for both distributions. On the other hand, the left tail of the control group distribution has more mass probability, indicating the number of classes suspicious of test scores manipulation is larger. Consistently with the estimation results, figure 7 shows that the difference between

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

the two distributions differs by macro region, being small in the North and Center, but substantially larger in the South & Islands.

Figure 6: Distribution of the likelihood, 10th grade mathematics exam



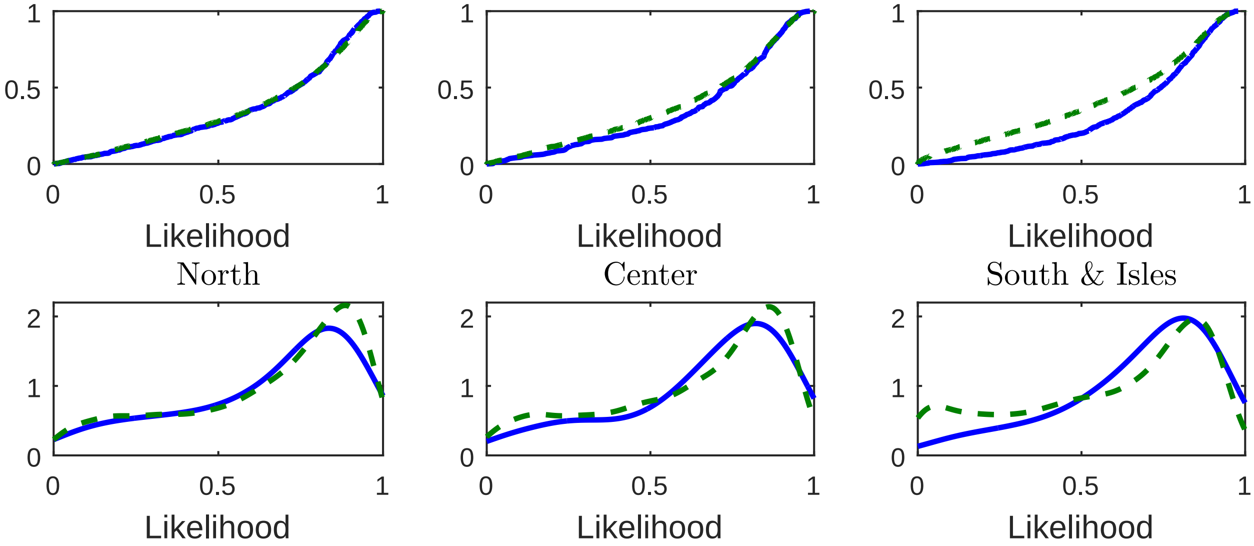
TR and CO respectively denote the estimated cdf and pdf of the estimated likelihood of the test scores of each class (equation 7).

5.1 Cheating Correction

Given the large regional differences regarding manipulation, the correction method proposed in section 3.4 is applied to each class using only data from their region. Figure 8 relates the correction applied to each class to their actual test scores and their likelihood, showing that a higher correction is applied to higher, less likely test scores. The correction proposed in this paper does not match the one proposed by Quintano et al. (2009), but the two of them are positively correlated.[38] There is, however, a large difference between the two of them, as shown in figure 9: the one proposed in this paper only leaves unchanged almost 20% of the test scores in the control group, and the remaining ones are applied a correction that is smaller than 3 points (out of a maximum of 50) for nearly 90% of them. In contrast, Quintano et al. (2009) correction does not correct about twice as many test scores, but the

[38] The linear correlation coefficient equals 0.54.

Figure 7: Distribution of the likelihood by regions, 10th grade mathematics exam



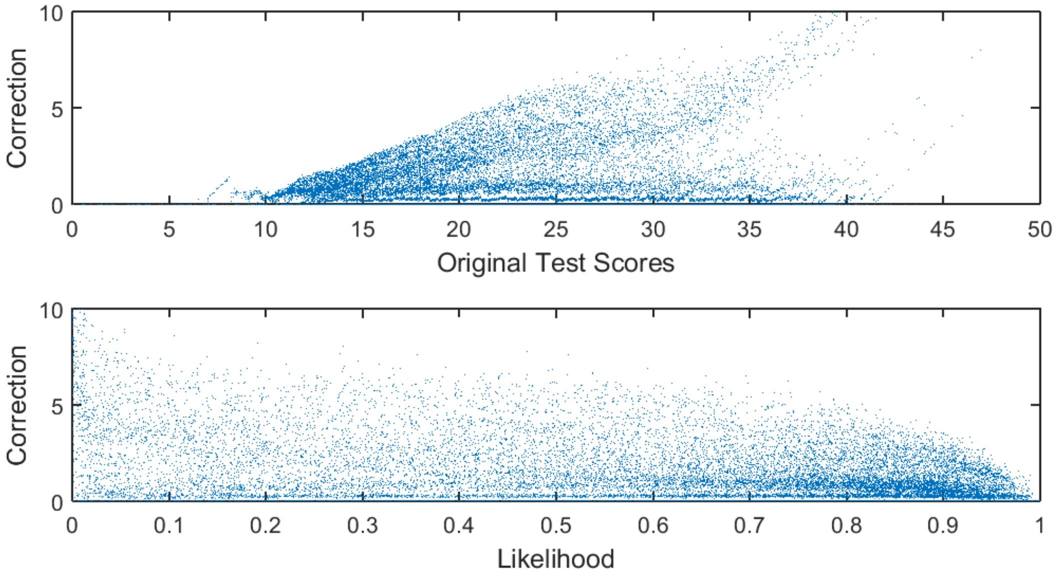
TR and CO respectively denote the estimated cdf and pdf for each of the three macro regions of the estimated likelihood of the test scores of each class (equation 7).

average correction for the remaining ones is much larger, and at least 10% of the test scores are corrected by more than 10 points.

Another way to compare both corrections is to look the mean correction applied in each region to the actual changes in mean test scores between the treatment and control groups. Figure 9 shows that 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: Quintano et al. (2009) correction consistently overestimates the average correction, resulting in a larger reduction of the mean test scores for students in the control group. The correction proposed in this paper matches better the mean difference between the treatment and control groups by region, in particular for those regions in the north and center of Italy, but it underestimates the correction in some of the southern ones. Finally, the maps in figure 11 show that the rankings of each region are changed after applying the correction.

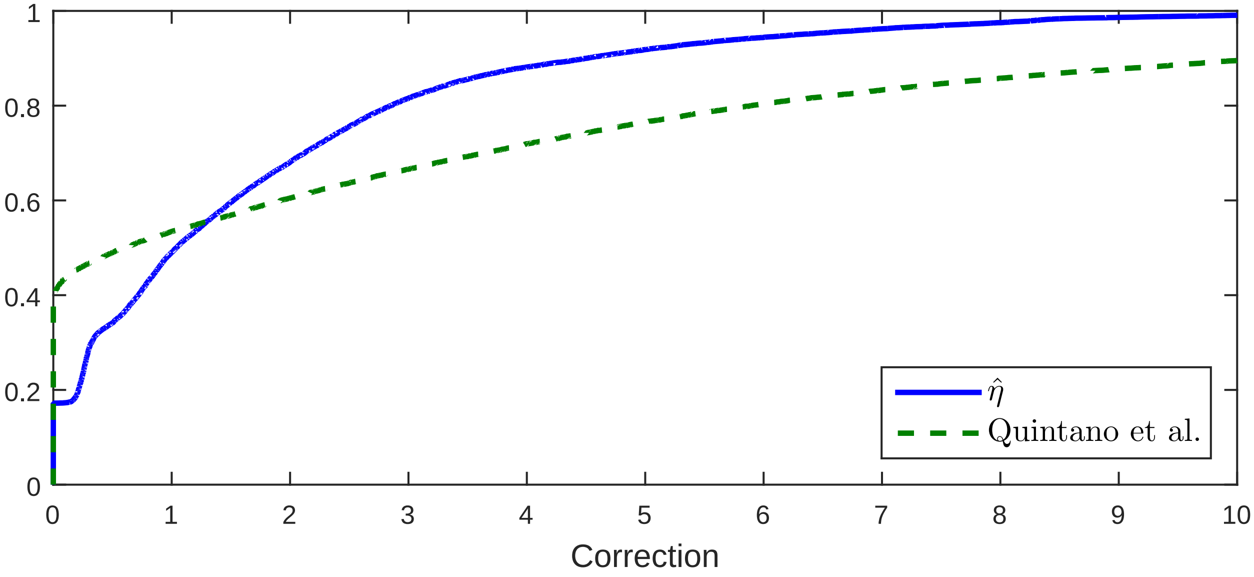
Finally I compare the mean within class variance of test scores before and after the correction is applied to the test scores in the control group, with the variability in the treatment group. The first three rows of table 5 show that the correction proposed in this

Figure 8: Correction for cheating, test scores, and likelihood, 10th grade mathematics exam



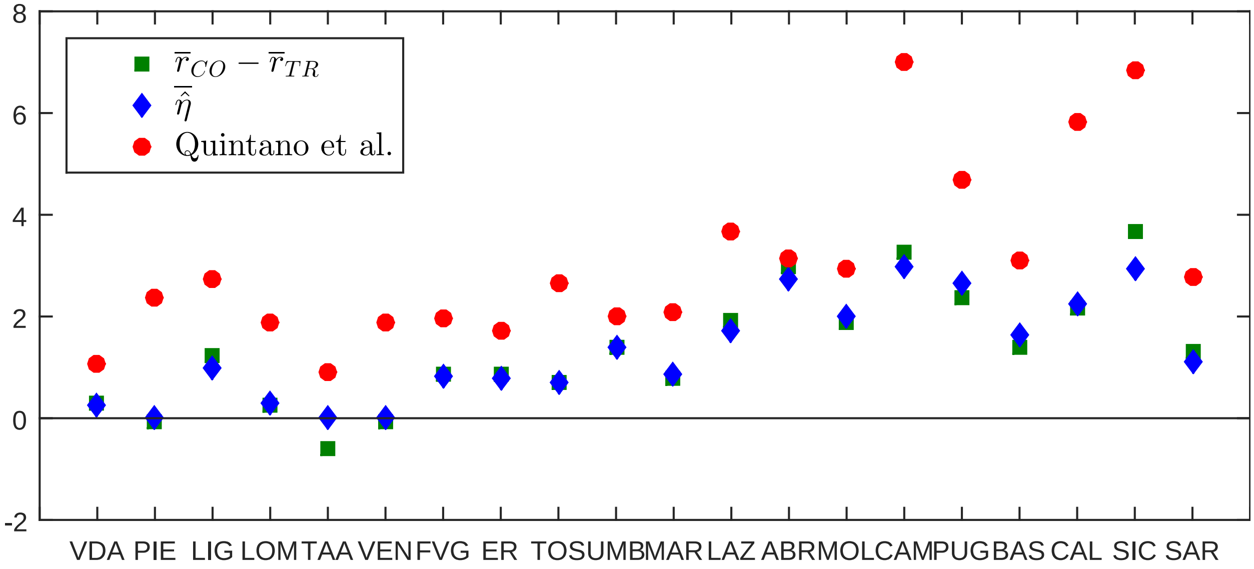
The upper and lower figures respectively show the scatter plot of the mean correction to the classes in the control group with the estimated likelihood of the test scores of each class (equation 7) and with the class mean test scores.

Figure 9: Distribution of correction for cheating, 10th grade mathematics exam



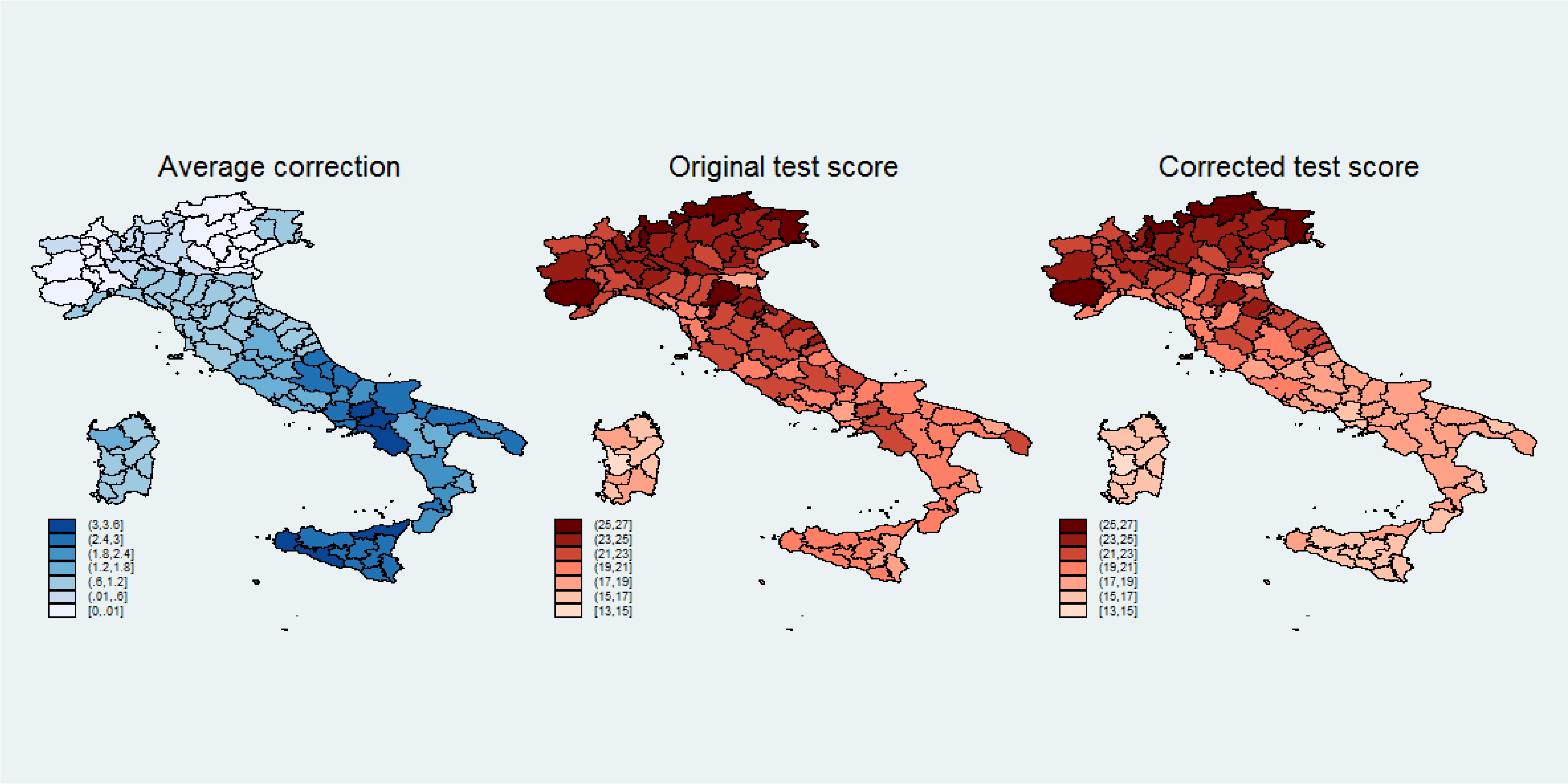
$\overline{\hat{\eta}}$ and *Quintano et al.* respectively denote the empirical cdf of the correction of the methods presented in this paper and the one proposes by Quintano et al. (2009).

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



For each region, $\overline{r}_{CO} - \overline{r}_{TR}$ denotes the mean difference in test scores between the treatment and control groups, $\overline{\hat{\eta}}$ denotes the mean correction of the method presented in this paper, and *Quintano et al.* denotes the mean correction of the method propose by Quintano et al. (2009).

Figure 11: Correction for cheating, provincial variation, 10th grade mathematics exam



paper increases the within class variance of test scores of the control group beyond that of the treatment group, although it is still closer to that value than before the correction. On the other hand, if we observe the mean within class correlation between the INVALSI test scores and the first semester exams taken at each school, the correction is particularly good at bringing this correlation for the control group closer to that of the treatment group in the Italian exams, but not so much for the mathematics exams, remaining largely unchanged.

Table 5:

	2nd grade		5th grade		6th grade		8th grade		10th grade	
	M	I	M	I	M	I	M	I	M	I
$\overline{Var}_W^T$	36.3	54.7	35.4	152.9	111.1	64.1	56.2	39.1	122.0	106.4
$\overline{Var}_W^{\tilde{C}}$	40.8	56.0	37.0	170.5	114.2	70.7	57.8	43.8	121.9	107.6
$\overline{Var}_W^C$	31.0	51.9	33.0	130.2	106.9	56.3	53.6	33.5	122.0	104.1
$Corr_T$	0.63	0.70	0.41	0.67	0.67	0.67	0.65	0.58	0.70	0.44
$Corr_{\tilde{C}}$	0.61	0.69	0.39	0.65	0.67	0.66	0.65	0.56	0.70	0.41
$Corr_C$	0.61	0.69	0.39	0.63	0.65	0.66	0.65	0.54	0.68	0.37

$\overline{Var}_W^T$, $\overline{Var}_W^{\tilde{C}}$, and $\overline{Var}_W^C$ respectively denote the mean within class variance of the INVALSI test scores for the treatment group, for the corrected test scores of the control group, and for the raw test scores of the control group. $Corr_T$, $Corr_{\tilde{C}}$, and $Corr_C$ respectively denote the mean within class correlation of the INVALSI test scores and the first semester test scores for the treatment group, for the corrected test scores of the control group, and for the raw test scores of the control group.

6 Welfare Analysis

Test scores manipulation is undesirable not just because it hinders the evaluation of the education system, but also because it can lead to distortions in students human capital accumulation, as it can lead to a mismatch between the optimal level of investment and the actual one. In particular, some students would be harmed and would underinvest in human capital, and the reverse could happen to other students. A way to assess the inefficiencies caused by the manipulation is to compare the counterfactual results of the exams with and without manipulation. No exam does a perfect job at identifying the actual level of knowledge of a student, so it is necessary the isolate the effect of the manipulation from that of the exam. To do so, I use the estimates from specification 6 for the treatment and control groups. Then, the final step is to check if the correction proposed in this paper helps to mitigate this problem.

To estimate the welfare loss of the manipulation, I simulate the results of the each exam for each student with the estimates from the treatment and control groups, which I denote by r^0_{ic} and r^1_{ic}, respectively. Then, I apply the correction method to the results with manipulation, denoted by $\tilde{r}^1_{ic}$. Then, I perform two types of exercises: one that compares the accuracy of the correction in terms of the final score, and another in which I set a fictitious passing grade and compute the number of students who got an unfair pass with and without correction. For the first exercise, I report the following statistics: the mean absolute deviation of the test score with manipulation ($I_1 \equiv \frac{1}{QN}\sum_{i=1}^{N} \|r^0_{ic} - r^1_{ic}\|$), and the same after the correction is applied ($I_2 \equiv \frac{1}{QN}\sum_{i=1}^{N} \|r^0_{ic} - \tilde{r}^1_{ic}\|$). For the second exercise I report the proportion of students who answered less than a half of the questions correctly without manipulation, but at least a half with manipulation ($I_3 \equiv \frac{1}{N}\sum_{i=1}^{N} \mathbf{1}\left(r^0_{ic} < \frac{Q}{2}, r^1_{ic} \geq \frac{Q}{2}\right)$), the proportion of these students who scored less than a half after the correction is applied ($I_4 \equiv \frac{1}{N}\sum_{i=1}^{N} \mathbf{1}\left(r^0_{ic} < \frac{Q}{2}, r^1_{ic} \geq \frac{Q}{2}, \tilde{r}^1_{ic} < \frac{Q}{2}\right)$), and the proportion of students who answered more than a half of the answers correctly without manipulation, but less when the correction is applied ($I_5 \equiv \frac{1}{N}\sum_{i=1}^{N} \mathbf{1}\left(r^0_{ic} \geq \frac{Q}{2}, \tilde{r}^1_{ic} < \frac{Q}{2}\right)$), *i.e.* the false positives.

Table 6: Impact of cheating and its correction on welfare

	2nd grade		5th grade		6th grade		8th grade		10th grade	
	M	I	M	I	M	I	M	I	M	I
I_1	13.3	7.7	8.3	5.3	1.3	5.9	3.3	1.4	3.9	7.3
I_2	10.7	5.4	7.8	3.0	1.3	5.9	3.2	1.3	3.5	7.3
I_3	21.0	10.3	14.9	6.2	0.5	2.3	2.6	1.3	6.3	3.4
I_4	12.1	5.0	6.2	1.7	0.5	2.3	2.2	0.5	3.0	2.9
I_5	4.4	2.8	6.1	1.4	0.0	0.0	0.1	0.7	1.0	1.0

All quantities are expressed in percentage points.

Table 6 shows the results of the simulation for all the exams. The manipulation has a heterogeneous effect on each exam, and indeed the mean absolute deviation of the test scores with manipulation can get up to 13%, with the proposed correction reducing this quantity by up to one third. Similarly, in terms of the fictitious passing grade, there is also great variability, making up to 21% of the students unfairly score above the passing grade. The correction also has a variable amelioration effect, which on average is about one half of the students who passed unfairly, at the cost of failing about 1.5% of the students who should have passed the exam because without manipulation.

7 Conclusion

Test scores manipulation can be harmful for a variety of reasons, ranging from the inability to accurately evaluate the education system, to suboptimal investment in education decisions by students. In this paper I propose a novel approach to detect test scores manipulation and correct for it, which is based on the comparison with a group of students whose test scores were not manipulated. Taking advantage of a natural experiment in the Italian education system, I apply a variety of nonlinear panel data regression methods to describe patterns in test scores manipulation, and based on these estimates, I apply the correction method to the test scores.

There is not an excess of correlation in students' answers when the monitor is a teacher from the own school, which rules out the hypotheses that the manipulation is explained by students copying each other or teachers telling them straight answers or grading in a systematic way; moreover, the larger the amount of estimated manipulation there is in open ended questions relative to multiple choice questions, the smaller the difference in missing questions between the two of them.[39] The estimated manipulation appears to be frequent in the South, in large classrooms, it tends to favor female and immigrant students. Unobserved heterogeneity accounts for an important share of the total variation, and it exhibits a substantial level of correlation within classrooms, reflecting a combination of teacher effects, sorting of students, and peer effects.

The correction method I propose punishes more those results that are more unlikely and higher test scores, and shows a large regional variation. When compared to Quintano

[39] These findings again are the opposite of what would happen if students copied each other, since multiple choice questions are easier to copy and answer than multiple choice ones. This does not rule out the possibilities that the manipulation is due to teachers' shirking, as suggested by Angrist et al. (2014), or by teachers having an active role.

et al. (2009) correction method, it fits better the mean change in test scores between the treatment and control groups by regions. Moreover, the simulation exercise suggests that applying the proposed correction method reduces the loss in accuracy of students' knowledge due to manipulation by between one third and three quarters.

References

Angrist, J. D., E. Battistin, and D. Vuri (2014). In a small moment: Class size and moral hazard in the mezzogiorno. Technical report, National Bureau of Economic Research.

Bacher-Hicks, A., T. J. Kane, and D. O. Staiger (2014). Validating teacher effect estimates using changes in teacher assignments in los angeles. Technical report, National Bureau of Economic Research.

Battistin, E., M. De Nadai, and D. Vuri (2014). Counting rotten apples: Student achievement and score manipulation in italian elementary schools. Technical report, IZA.

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.

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.

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

Figlio, D. N. and L. S. Getzler (2006). Accountability, ability and disability: Gaming the system? *Advances in applied microeconomics 14*, 35–49.

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.

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

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. 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.

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

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

Marshall, A. W. and I. Olkin (1988). Families of multivariate distributions. *Journal of the American Statistical Association 83*(403), 834–841.

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

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.

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

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.

Train, K. E. (2009). *Discrete choice methods with simulation.* Cambridge university press.

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

Appendix

A Random Effects Models with Copula Dependence

Let the distributions of the unobservables be given by $\varepsilon_{icq} \sim F_{\varepsilon}$, $\eta_{ic} \sim F_{\eta}\left(\sigma_{\eta}\right)$, and $u_c \equiv F_{\eta}^{-1} \sim C\left(\rho\right)$. Moreover, let x_{1ic} be a K_1 dimensional vector of question invariant covariates, and x_{2icq} a K_2 dimensional vector of question dependent covariates. Conditional on η_{ic}, the probability of each of the students in classroom c of obtaining y_{icq} for each question equals

$$\begin{aligned}\mathbb{P}\left(y_c|\eta_c, x_{1c}, x_{2c}\right) &= \prod_{i=1}^{N_c}\prod_{q=1}^{Q}\left(1 - F_{\varepsilon}\left(-\eta_{ic} - \xi_q - x_{1ic}'\beta - x_{2icq}'\zeta_q\right)\right)^{y_{icq}} \\ &\times \; F_{\varepsilon}\left(-\eta_{ic} - \xi_q - x_{1ic}'\beta - x_{2icq}'\zeta_q\right)^{1-y_{icq}}\end{aligned}$$

In the absence of correlation across students in the same classroom, *i.e.* the copula is independent, and so the likelihood function is written as

$$\begin{aligned}\mathcal{L}\left(\theta\right) &= \sum_{c=1}^{C}\sum_{i=1}^{N_c}\log\left(\int_0^1\prod_{q=1}^{Q}\left(1 - F_{\varepsilon}\left(-\eta_{ic} - \xi_q - x_{1ic}'\beta - x_{2icq}'\zeta_q\right)\right)^{y_{icq}}\right. \\ &\times \; \left. F_{\varepsilon}\left(-\eta_{ic} - \xi_q - x_{1ic}'\beta - x_{2icq}'\zeta_q\right)^{1-y_{icq}} dF_{\eta}\left(\eta_{ic}\right)\right)\end{aligned}$$

More generally, if the vector of individual random effects of students in classroom c is correlated according to the copula distribution, then the likelihood function is written as

$$\begin{aligned}\mathcal{L}\left(\theta\right) &= \sum_{c=1}^{C}\log\left(\int_{[0,1]^{N_c}}\prod_{i=1}^{N_c}\prod_{q=1}^{Q}\left(1 - F_{\varepsilon}\left(-\eta_{ic} - \xi_q - x_{1ic}'\beta - x_{2icq}'\zeta_q\right)\right)^{y_{icq}}\right. \\ &\times \; \left. F_{\varepsilon}\left(-\eta_{ic} - \xi_q - x_{1ic}'\beta - x_{2icq}'\zeta_q\right)^{1-y_{icq}} dC\left(u_c;\rho\right)\right)\end{aligned}$$

B Estimation of the Random Effects Logit with Copula Dependence

Consider the likelihood function given by equation 5 with the aforementioned parametric assumptions for the unobservables.[40] Moreover, define $\tilde{\eta}_{ic} \equiv \sigma_{\eta}^{-1}\eta_{ic}$, *i.e.* the standardized

[40] Notice that in the main text the covariates were omitted for notational clarity.

individual effect. The likelihood can be rewritten as

$$\begin{aligned}\mathcal{L}(\theta) &= \sum_{c=1}^{C} \log \left(\int_{[0,1]^{N_c}} \frac{\exp\left(\sum_{i=1}^{N_c} \sum_{q=1}^{Q} y_{icq} \left(\eta_{ic} + \xi_q + x'_{1ic}\beta + x'_{2icq}\zeta_q\right)\right)}{\prod_{i=1}^{N_c} \prod_{q=1}^{Q} \left(1 + \exp\left(\eta_{ic} + \xi_q + x'_{1ic}\beta + x'_{2icq}\zeta_q\right)\right)} dC(u_c;\rho) \right) \\ &\equiv \sum_{c=1}^{C} \log \left(\int_{[0,1]^{N_c}} H_c(\theta)\, dC(\Phi(\tilde{\eta}_c);\rho) \right) \equiv \sum_{c=1}^{C} l_c(\theta)\end{aligned}$$

Estimation of θ by maximum likelihood is computationally intensive and it requires the combination of Monte Carlo simulation and an approximation of the derivative of the likelihood function with respect to the correlation parameter. The Monte Carlo simulation is standard and is used to approximate the different integrals that come up in the derivatives of the likelihood function. The derivatives with respect to the question effects and the standard deviation of the individual effects do not present any further inconvenience, but the derivative with respect to the correlation parameter is more complicated, as it requires working with the copula *pdf*, which is not analytical or impractical to work with with some of the most common families of copulas. Define $z_{icq} \equiv \exp\left(\eta_{ic} + \xi_q + x'_{1ic}\beta + x'_{2icq}\zeta_q\right)$ and let $c(\Phi(\tilde{\eta}_c);\rho)$ denote the copula pdf. Then, the derivatives of the likelihood function are given by

$$\frac{\partial \mathcal{L}(\theta)}{\partial \xi_p} = \sum_{c=1}^{C} \sum_{i=1}^{N_c} \left(y_{icp} - \frac{\int_0^1 H_c(\theta) \frac{z_{icp}}{1+z_{icp}} dC(u_c;\rho)}{\int_0^1 H_c(\theta)\, dC(u_c;\rho)} \right) \tag{12}$$

$$\frac{\partial \mathcal{L}(\theta)}{\partial \zeta_p} = \sum_{c=1}^{C} \sum_{i=1}^{N_c} \left(y_{icp} - \frac{\int_0^1 H_c(\theta) \frac{z_{icp}}{1+z_{icp}} dC(u_c;\rho)}{\int_0^1 H_c(\theta)\, dC(u_c;\rho)} \right) x'_{2icq} \tag{13}$$

$$\frac{\partial \mathcal{L}(\theta)}{\partial \beta} = \sum_{c=1}^{C} \sum_{i=1}^{N_c} \sum_{q=1}^{Q} \left(y_{icq} - \frac{\int_0^1 H_c(\theta) \frac{z_{icq}}{1+z_{icq}} dC(u_c;\rho)}{\int_0^1 H_c(\theta)\, dC(u_c;\rho)} \right) x'_{1ic} \tag{14}$$

$$\frac{\partial \mathcal{L}(\theta)}{\partial \sigma_\eta} = \sum_{c=1}^{C} \sum_{i=1}^{N_c} \sum_{q=1}^{Q} \frac{\int_0^1 H_c \left(y_{icq} - \frac{z_{icq}}{1+z_{icq}} \right) \tilde{\eta}_{ic}(\theta)\, dC(u_c;\rho)}{\int_0^1 H_c(\theta)\, dC(u_c;\rho)} \tag{15}$$

$$\frac{\partial \mathcal{L}(\theta)}{\partial \rho} = \sum_{c=1}^{C} \frac{\int_0^1 \cdots \int_0^1 H_c(\theta)\, {}^{\partial c(\Phi(\tilde{\eta}_c);\rho)}/_{\partial\rho} \prod_{j=1}^{N_c} d\Phi(\tilde{\eta}_{jc})}{\int_0^1 H_c(\theta)\, dC(\Phi(\tilde{\eta}_c);\rho)} \tag{16}$$

Equations 12 to 16 cannot be computed analytically, and because of the dimensionality of the integrals, methods like Monte Carlo tend to have a poor performance and are slow. However, if the copula is Archimedean, it is possible to approximate these integrals overcoming the curse of dimensionality. By Corollary 2.2 in Marshall and Olkin (1988), an Archimedean copula can be written as

$$C\left(u\right)=\int_{[0,1]^N}\exp\left(-\sum_{i=1}^{N}\theta_i\phi_i^{-1}\left(u_i\right)\right)dG\left(\theta\right) \tag{17}$$

where G is the cdf of θ, and ϕ_i is the Laplace transform of the marginal distributions of G. Some of the most common copulas have θ unidimensional and $\phi_i=\phi\forall i$. For example, if the copula is a $Clayton\left(\rho\right)$, then G is the cdf of a $\Gamma\left(\rho,1\right)$, and ϕ, also known as the generator of the copula, equals $\phi\left(s;\rho\right)=\left(1+s\right)^{-\rho}$. Now consider the following integral:

$$\begin{aligned}\mathcal{L} &= \int_{[0,1]^N}\prod_{i=1}^{N}\Lambda_i\left(u_i\right)dC\left(u\right)\\ &= \int_0^1\prod_{i=1}^{N}\left[\int_0^1\Lambda_i\left(u_i\right)dF^{\theta}\left(u_i\right)\right]dG\left(\theta\right)\end{aligned} \tag{18}$$

where $F^{\theta}\left(u_i\right)=\exp\left(-\theta\phi^{-1}\left(u_i\right)\right)$. Hence, the originally N dimensional integral equals the integral of the product of N independent integrals, reducing the dimensionality from N to 2. Hence, the approximation of the integral can be done as follows:

1. Compute a grid of values of θ, given by $\theta_j=G^{-1}\left(\frac{j}{N_1+1}\right)$, $\forall j=1,...,N_1$.

2. Compute a grid of values of u $\forall j$, given by $u_{jh}=\phi\left(-\frac{1}{\theta_j}\log\left(\frac{h}{N_2+1}\right)\right)$, $\forall h=1,...,N_2$.

3. Approximate the integral by $\mathcal{L}\approx\frac{1}{N_1}\sum_{j=1}^{N_1}\prod_{i=1}^{N}\left[\frac{1}{N_2}\sum_{h=1}^{N_2}\Lambda_i\left(u_{jh}\right)\right]$.

Let $\Lambda_i\left(u_i\right)=\frac{\exp\left(\sum_{q=1}^{Q}y_{icq}\left(F_\eta^{-1}(u_i)+\xi_q+x'_{1ic}\beta+x'_{2icq}\zeta_q\right)\right)}{\prod_{q=1}^{Q}\left(1+\exp\left(F_\eta^{-1}(u_i)+\xi_q+x'_{1ic}\beta+x'_{2icq}\zeta_q\right)\right)}$, and sum over all classes to approximate the likelihood function. The approximation of the Jacobian for all parameters except for ρ is done in a similar way. For equation 16, approximate the derivative numerically by $\left(\mathcal{L}\left(\theta^+\right)-\mathcal{L}\left(\theta\right)\right)/\epsilon$, where θ^+ is evaluated at $\rho+\epsilon$ instead of ρ. To obtain the estimates I used the BHHH algorithm[41]. Estimation of the Hessian is straightforward, and is given by

$$\hat{H}\left(\hat{\theta}_{MLE}\right)=\frac{1}{C}\sum_{c=1}^{C}\frac{\partial\hat{l}_c\left(\hat{\theta}_{MLE}\right)}{\partial\theta}\frac{\partial\hat{l}_c\left(\hat{\theta}_{MLE}\right)}{\partial\theta'} \tag{19}$$

C Some linear algebra results

Let z be a vector of dimension T, Z be 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

[41] For further details on numerical optimization, see for instance Train (2009).

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 $\sum_{k_1=1}^{K-r+1} \ldots \sum_{k_r=k_{r-1}+1}^{T} \prod_{j=1}^{r} z_{k_j} = \iota_T' (ZG)^{r-1} Z \iota_T$ for $1 \leq r \leq T$. Now consider equation 5. If the distribution of ε_{icq} is logistic, then the probability of a particular result, $(b_1, \ldots, b_{N_c})$, 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 equation 6, *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 approximation of the aforementioned probability, which would be exact if not for the integral. Mathematically,

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

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.[42] η_{jh} is defined as in appendix A.

D Full Results

[42] 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 $\prod_{q=1}^{Q} \exp\left(\eta + \xi_q + x_{1ic}'\beta + x_{2icq}'\zeta_q\right)$.

Table 7: FE Logit estimates

	2nd grade		5th grade		6th grade		8th grade		10th grade	
	M	I	M	I	M	I	M	I	M	I
$\hat{\xi}^{TR} > \hat{\xi}^{CO}$	4	3	1	2	0	2	2	1	5	4
$\hat{\xi}^{TR} < \hat{\xi}^{CO}$	6	18	33	17	39	8	9	36	29	12
$\hat{\xi}^{TR} = \hat{\xi}^{CO}$	21	17	12	62	8	60	33	40	15	71

TR and CO respectively denote the treatment and control groups. 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 8: Correlation between FE Logit estimates and raw test scores

	2nd grade		5th grade		6th grade		8th grade		10th grade	
	M	I	M	I	M	I	M	I	M	I
TR	0.99	0.98	1.00	0.99	1.00	0.98	1.00	0.99	0.99	0.99
CO	0.99	0.94	0.99	0.98	1.00	0.97	0.99	0.98	1.00	0.97

TR and CO respectively denote the treatment and control groups.

Table 9: Correlation between FE Logit estimates and Logit estimates

	2nd grade		5th grade		6th grade		8th grade		10th grade	
	M	I	M	I	M	I	M	I	M	I
TR	0.99	1.00	0.99	0.98	1.00	0.98	0.97	0.99	1.00	0.93
CO	0.99	0.99	0.99	1.00	1.00	0.92	0.91	0.99	0.99	0.89

TR and CO respectively denote the treatment and control groups.

Table 10: 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
$\hat{\xi}^{TR,MA} > \hat{\xi}^{TR,FE}$	5	25	24	16	14	12	18	39	17	13
$\hat{\xi}^{TR,MA} < \hat{\xi}^{TR,FE}$	12	1	0	1	17	5	5	0	16	55
$\hat{\xi}^{TR,MA} = \hat{\xi}^{TR,FE}$	14	12	22	64	16	53	21	38	16	19
$\hat{\xi}^{CO,MA} > \hat{\xi}^{CO,FE}$	6	36	28	51	19	34	29	39	26	19
$\hat{\xi}^{CO,MA} < \hat{\xi}^{CO,FE}$	16	1	8	12	25	20	11	21	19	63
$\hat{\xi}^{CO,MA} = \hat{\xi}^{CO,FE}$	9	1	10	18	3	16	4	17	4	5

TR and CO respectively denote the treatment and control groups, whereas MA and FE denote male and female students. 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 11: 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
$\hat{\xi}^{TR,SM} > \hat{\xi}^{TR,LA}$	0	3	15	0	2	0	4	0	15	11
$\hat{\xi}^{TR,SM} < \hat{\xi}^{TR,LA}$	5	0	0	0	0	0	2	0	6	32
$\hat{\xi}^{TR,SM} = \hat{\xi}^{TR,LA}$	26	35	41	81	45	70	38	77	28	44
$\hat{\xi}^{CO,SM} > \hat{\xi}^{CO,LA}$	15	26	24	67	20	5	25	18	28	22
$\hat{\xi}^{CO,SM} < \hat{\xi}^{CO,LA}$	4	3	1	0	9	26	4	5	7	52
$\hat{\xi}^{CO,SM} = \hat{\xi}^{CO,LA}$	12	9	21	14	18	39	15	54	14	13

TR and CO respectively denote the treatment and control groups, whereas SM and LA denote that the students were in classrooms of size smaller or equal to the median, and larger. 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 12: 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
$\hat{\xi}^{TR,NO} > \hat{\xi}^{TR,SI}$	1	1	1	0	1	0	1	0	0	0
$\hat{\xi}^{TR,NO} < \hat{\xi}^{TR,SI}$	1	3	0	16	6	2	14	10	14	30
$\hat{\xi}^{TR,NO} = \hat{\xi}^{TR,SI}$	29	34	44	65	40	68	29	67	35	57
$\hat{\xi}^{CO,NO} > \hat{\xi}^{CO,SI}$	8	8	1	0	1	5	0	1	1	2
$\hat{\xi}^{CO,NO} < \hat{\xi}^{CO,SI}$	15	12	34	77	39	22	43	67	40	68
$\hat{\xi}^{CO,NO} = \hat{\xi}^{CO,SI}$	8	18	11	4	7	43	1	9	8	17

TR and CO respectively denote the treatment and control groups, whereas NO and SI denote that the students were from the North and South & Islands regions. 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 13: RE Logit estimates

	2nd grade		5th grade		6th grade		8th grade		10th grade	
	M	I	M	I	M	I	M	I	M	I
$\hat{\xi}^{TR} > \hat{\xi}^{CO}$	0	0	4	0	28	5	2	10	6	0
$\hat{\xi}^{TR} < \hat{\xi}^{CO}$	32	39	37	82	10	54	36	40	38	87
$\hat{\xi}^{TR} = \hat{\xi}^{CO}$	0	0	6	0	10	12	7	28	6	1

TR and CO respectively denote the treatment and control groups. 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 14: Correlation between RE Logit estimates and FE Logit estimates

	2nd grade		5th grade		6th grade		8th grade		10th grade	
	M	I	M	I	M	I	M	I	M	I
TR	1.00	0.87	1.00	0.98	1.00	0.97	0.97	1.00	1.00	0.95
CO	0.99	1.00	0.98	0.95	1.00	1.00	1.00	1.00	1.00	1.00

TR and CO respectively denote the treatment and control groups.

Table 15: Comparison between RE Logit estimates and Logit estimates

	2nd grade		5th grade		6th grade		8th grade		10th grade	
	M	I	M	I	M	I	M	I	M	I
TR,≠	30	35	40	77	33	70	37	75	43	83
TR,=	2	4	7	5	15	1	8	3	7	5
CO,≠	32	39	46	82	45	71	45	78	48	88
CO,=	0	0	1	0	3	0	0	0	2	0

TR and CO respectively denote the treatment and control groups; = and ≠ 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 16: RE logit estimates, 2nd grade mathematics exam

		$\hat{\bar{\xi}}$	FE	CE	SI	IT	$SMALL$	$\hat{\sigma}_\eta$	$\hat{\rho}$
Treatment	(1)	0.21***	-	-	-	-	-	1.05***	-
		(0.01)						(0.01)	
	(2)	0.26***	-0.07***	-	-	-	-	1.09***	-
		(0.01)	(0.02)					(0.01)	
	(3)	0.30***	-0.08***	0.00	-0.13***	-	-	1.10***	-
		(0.02)	(0.02)	(0.02)	(0.02)			(0.01)	
	(4)	-0.03	-0.09***	0.00	-0.22***	0.44***	0.01***	1.12***	-
		(0.04)	(0.02)	(0.02)	(0.02)	(0.04)	(0.00)	(0.01)	
	(5)	0.22***	-	-	-	-	-	0.92***	3.03***
		(0.01)						(0.00)	(0.03)
	(6)	0.15***	-0.07***	-0.05***	-0.21***	0.42***	-0.39***	0.95***	2.29***
		(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.02)
Control	(1)	0.57***	-	-	-	-	-	1.21***	-
		(0.00)						(0.00)	
	(2)	0.69***	-0.02***	-	-	-	-	1.25***	-
		(0.00)	(0.00)					(0.00)	
	(3)	0.58***	-0.03***	0.14***	0.30***	-	-	1.24***	-
		(0.00)	(0.00)	(0.01)	(0.00)			(0.00)	
	(4)	0.23***	-0.03***	0.16***	0.36***	0.35***	0.00***	1.24***	-
		(0.01)	(0.00)	(0.01)	(0.00)	(0.01)	(0.00)	(0.00)	
	(5)	0.73***	-	-	-	-	-	0.90***	1.82***
		(0.00)						(0.00)	(0.00)
	(6)	0.28***	-0.02***	0.23***	0.53***	0.41***	-0.03***	1.02***	0.80***
		(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)

FE represents the average of the interaction between a dummy for female students and the question dummies, CE, SI, IT, and SMALL are dummies for Center region, South & Islands region, natives, and small class. *, **, and *** denote statistically significant at the 90%, 95%, and 99% level, respectively, standard errors in parentheses.

Table 17: RE logit estimates, 5th grade mathematics exam

		$\hat{\bar{\xi}}$	FE	CE	SI	IT	$SMALL$	$\hat{\sigma}_\eta$	$\hat{\rho}$
Treatment	(1)	0.31***	-	-	-	-	-	0.87***	-
		(0.01)						(0.01)	
	(2)	0.30***	-0.17***	-	-	-	-	0.96***	-
		(0.01)	(0.02)					(0.01)	
	(3)	0.40***	-0.18***	-0.05***	-0.21***	-	-	0.95***	-
		(0.01)	(0.02)	(0.02)	(0.01)			(0.01)	
	(4)	0.08***	-0.22***	-0.07***	-0.30***	0.43***	0.01***	0.93***	-
		(0.03)	(0.02)	(0.02)	(0.01)	(0.03)	(0.00)	(0.01)	
	(5)	0.30***	-	-	-	-	-	0.78***	3.39***
		(0.00)						(0.00)	(0.02)
	(6)	0.25***	-0.22***	-0.09***	-0.36***	0.41***	-0.38***	0.80***	2.47***
		(0.01)	(0.00)	(0.01)	(0.01)	(0.01)	(0.01)	(0.00)	(0.03)
Control	(1)	0.46***	-	-	-	-	-	1.02***	-
		(0.00)						(0.00)	
	(2)	0.53***	-0.15***	-	-	-	-	1.11***	-
		(0.00)	(0.00)					(0.00)	
	(3)	0.55***	-0.17***	0.02***	-0.01***	-	-	1.10***	-
		(0.00)	(0.00)	(0.00)	(0.00)			(0.00)	
	(4)	0.20***	-0.19***	0.02***	-0.03***	0.41***	0.00***	1.09***	-
		(0.01)	(0.00)	(0.00)	(0.00)	(0.01)	(0.00)	(0.00)	
	(5)	0.47***	-	-	-	-	-	0.83***	2.48***
		(0.00)						(0.00)	(0.00)
	(6)	0.24***	-0.14***	0.09***	0.03***	0.40***	-0.15***	0.85***	0.93***
		(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)

FE represents the average of the interaction between a dummy for female students and the question dummies, CE, SI, IT, and SMALL are dummies for Center region, South & Islands region, natives, and small class. *, **, and *** denote statistically significant at the 90%, 95%, and 99% level, respectively, standard errors in parentheses.

Table 18: RE logit estimates, 6th grade mathematics exam

		$\hat{\bar{\xi}}$	FE	CE	SI	IT	$SMALL$	$\hat{\sigma}_\eta$	$\hat{\rho}$
Treatment	(1)	-0.25***	-	-	-	-	-	0.82***	-
		(0.01)						(0.01)	
	(2)	-0.16***	-0.17***	-	-	-	-	0.81***	-
		(0.01)	(0.01)					(0.01)	
	(3)	0.10***	-0.12***	-0.19***	-0.67***	-	-	0.56***	-
		(0.01)	(0.01)	(0.01)	(0.01)			(0.00)	
	(4)	-0.58***	-0.20***	-0.08***	-0.27***	0.71***	0.02***	0.53***	-
		(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.00)	(0.00)	
	(5)	-0.30***	-	-	-	-	-	0.68***	2.21***
		(0.01)						(0.00)	(0.02)
	(6)	-0.55***	-0.19***	-0.09***	-0.28***	0.71***	-0.20***	0.50***	2.03***
		(0.01)	(0.00)	(0.01)	(0.01)	(0.01)	(0.00)	(0.00)	(0.02)
Control	(1)	-0.28***	-	-	-	-	-	0.61***	-
		(0.00)						(0.00)	
	(2)	-0.17***	-0.13***	-	-	-	-	0.82***	-
		(0.00)	(0.00)					(0.00)	
	(3)	0.13***	-0.13***	-0.19***	-0.56***	-	-	0.54***	-
		(0.00)	(0.00)	(0.00)	(0.00)			(0.00)	
	(4)	-0.56***	-0.15***	-0.07***	-0.21***	0.63***	0.03***	0.52***	-
		(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	
	(5)	-0.31***	-	-	-	-	-	0.53***	2.90******
		(0.00)						(0.00)	(0.01)
	(6)	-0.50***	-0.14***	-0.10***	-0.26***	0.62***	-0.23***	0.48***	2.26***
		(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)

FE represents the average of the interaction between a dummy for female students and the question dummies, CE, SI, IT, and SMALL are dummies for Center region, South & Islands region, natives, and small class. *, **, and *** denote statistically significant at the 90%, 95%, and 99% level, respectively, standard errors in parentheses.

Table 19: RE logit estimates, 8th grade mathematics exam

		$\hat{\bar{\xi}}$	FE	CE	SI	IT	$SMALL$	$\hat{\sigma}_{\eta}$	$\hat{\rho}$
Treatment	(1)	0.07***	-	-	-	-	-	0.91***	-
		(0.01)						(0.01)	
	(2)	0.13***	-0.19***	-	-	-	-	0.97***	-
		(0.01)	(0.02)					(0.01)	
	(3)	0.16***	-0.21***	-0.01	-0.09***	-	-	0.96***	-
		(0.01)	(0.01)	(0.02)	(0.01)			(0.01)	
	(4)	-0.25***	-0.22***	-0.03*	-0.15***	0.52***	0.02***	0.94***	-
		(0.03)	(0.01)	(0.02)	(0.01)	(0.02)	(0.00)	(0.01)	
	(5)	0.14***	-	-	-	-	-	0.72***	2.14***
		(0.00)						(0.00)	(0.01)
	(6)	-0.10***	-0.17***	0.06***	0.03**	0.39***	-0.21***	0.78***	0.76***
		(0.02)	(0.01)	(0.02)	(0.01)	(0.01)	(0.01)	(0.00)	(0.01)
Control	(1)	0.16***	-	-	-	-	-	0.83***	-
		(0.00)						(0.00)	
	(2)	0.17***	-0.14***	-	-	-	-	0.93***	-
		(0.00)	(0.00)					(0.00)	
	(3)	0.16***	-0.16***	0.01**	0.02***	-	-	0.93***	-
		(0.00)	(0.00)	(0.00)	(0.00)			(0.00)	
	(4)	-0.20***	-0.17***	0.00	-0.01**	0.46***	0.02***	0.92***	-
		(0.01)	(0.00)	(0.00)	(0.00)	(0.01)	(0.00)	(0.00)	
	(5)	0.16***	-	-	-	-	-	0.67***	2.45***
		(0.00)						(0.00)	(0.00)
	(6)	-0.16***	-0.14***	0.11***	0.13***	0.38***	-0.27***	0.77***	1.41***
		(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)

FE represents the average of the interaction between a dummy for female students and the question dummies, CE, SI, IT, and SMALL are dummies for Center region, South & Islands region, natives, and small class. *, **, and *** denote statistically significant at the 90%, 95%, and 99% level, respectively, standard errors in parentheses.

Table 20: RE logit estimates, 2nd grade Italian exam

		$\hat{\bar{\xi}}$	FE	CE	SI	IT	$SMALL$	$\hat{\sigma}_{\eta}$	$\hat{\rho}$
Treatment	(1)	0.52***	-	-	-	-	-	0.78***	-
		(0.01)						(0.01)	
	(2)	0.55***	0.09***	-	-	-	-	0.85***	-
		(0.01)	(0.01)					(0.01)	
	(3)	0.62***	0.09***	-0.03*	-0.1***3	-	-	0.83***	-
		(0.01)	(0.01)	(0.02)	(0.01)			(0.01)	
	(4)	0.27***	0.10***	-0.04**	-0.14***	0.41***	0.01***	0.81***	-
		(0.03)	(0.01)	(0.02)	(0.01)	(0.03)	(0.00)	(0.01)	
	(5)	0.62***	-	-	-	-	-	0.62***	1.65***
		(0.00)						(0.00)	(0.01)
	(6)	0.52***	0.10***	-0.03***	-0.13***	0.38***	-0.29***	0.62***	0.44***
		(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.00)	(0.01)
Control	(1)	0.92***	-	-	-	-	-	0.95***	-
		(0.00)						(0.00)	
	(2)	0.86***	0.14***	-	-	-	-	0.92***	-
		(0.00)	(0.00)					(0.00)	
	(3)	0.79***	0.12***	0.07***	0.20***	-	-	0.91***	-
		(0.00)	(0.00)	(0.00)	(0.00)			(0.00)	
	(4)	0.50***	0.12***	0.07***	0.20***	0.30***	0.00***	0.90***	-
		(0.01)	(0.00)	(0.00)	(0.00)	(0.01)	(0.00)	(0.00)	
	(5)	1.10***	-	-	-	-	-	0.72***	1.52***
		(0.00)						(0.00)	(0.00)
	(6)	0.62***	0.12***	0.08***	0.28***	0.33***	-0.08***	0.78***	0.66***
		(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)

FE represents the average of the interaction between a dummy for female students and the question dummies, CE, SI, IT, and SMALL are dummies for Center region, South & Islands region, natives, and small class. *, **, and *** denote statistically significant at the 90%, 95%, and 99% level, respectively, standard errors in parentheses.

Table 21: RE logit estimates, 5th grade Italian exam

		$\hat{\bar{\xi}}$	FE	CE	SI	IT	$SMALL$	$\hat{\sigma}_\eta$	$\hat{\rho}$
Treatment	(1)	1.26***	-	-	-	-	-	1.07***	-
		(0.01)						(0.01)	
	(2)	1.31***	0.11***	-	-	-	-	1.08***	-
		(0.01)	(0.01)					(0.01)	
	(3)	1.44***	0.17***	-0.01	-0.18***	-	-	1.07***	-
		(0.01)	(0.01)	(0.01)	(0.01)			(0.01)	
	(4)	1.26***	0.20***	-0.03*	-0.29***	0.30***	0.01***	1.05***	-
		(0.02)	(0.01)	(0.01)	(0.01)	(0.02)	(0.00)	(0.01)	
	(5)	1.25***	-	-	-	-	-	1.01***	3.50***
		(0.01)						(0.01)	(0.02)
	(6)	1.25***	0.21***	-0.02***	-0.29***	0.30***	-0.26***	1.00***	2.72***
		(0.01)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.01)	(0.02)
Control	(1)	1.49***	-	-	-	-	-	1.09***	-
		(0.00)						(0.00)	
	(2)	1.53***	0.15***	-	-	-	-	1.09***	-
		(0.00)	(0.00)					(0.00)	
	(3)	1.58***	0.22***	0.01***	-0.06***	-	-	1.08***	-
		(0.00)	(0.00)	(0.00)	(0.00)			(0.00)	
	(4)	1.45***	0.24***	0.02***	-0.08***	0.14***	0.01***	1.03***	-
		(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	
	(5)	1.52***	-	-	-	-	-	1.03***	2.79***
		(0.00)						(0.00)	(0.00)
	(6)	1.50***	0.25***	0.02***	-0.07***	0.14***	-0.07***	0.98***	2.68***
		(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)

FE represents the average of the interaction between a dummy for female students and the question dummies, CE, SI, IT, and SMALL are dummies for Center region, South & Islands region, natives, and small class. *, **, and *** denote statistically significant at the 90%, 95%, and 99% level, respectively, standard errors in parentheses.

Table 22: RE logit estimates, 6th grade Italian exam

		$\hat{\bar{\xi}}$	FE	CE	SI	IT	$SMALL$	$\hat{\sigma}_\eta$	$\hat{\rho}$
Treatment	(1)	0.82***	-	-	-	-	-	1.06***	-
		(0.01)						(0.01)	
	(2)	0.92***	0.06***	-	-	-	-	1.05***	-
		(0.01)	(0.01)					(0.01)	
	(3)	1.03***	0.09***	-0.03*	-0.16***	-	-	0.98***	-
		(0.01)	(0.01)	(0.02)	(0.01)			(0.01)	
	(4)	0.66***	0.10***	-0.04***	-0.27***	0.41***	0.01***	0.84***	-
		(0.02)	(0.01)	(0.01)	(0.01)	(0.02)	(0.00)	(0.01)	
	(5)	0.45***	-	-	-	-	-	1.95***	0.95***
		(0.12)						(0.02)	(0.05)
	(6)	0.86***	0.16***	-0.09***	-0.35***	0.50***	-0.31***	0.71***	0.58***
		(0.01)	(0.00)	(0.00)	(0.00)	(0.01)	(0.00)	(0.00)	(0.01)
Control	(1)	0.93***	-	-	-	-	-	0.96***	-
		(0.00)						(0.00)	
	(2)	0.87***	0.11***	-	-	-	-	0.91***	-
		(0.00)	(0.00)					(0.00)	
	(3)	0.97***	0.17***	-0.03***	-0.17***	-	-	0.95***	-
		(0.00)	(0.00)	(0.00)	(0.00)			(0.00)	
	(4)	0.67***	0.20***	-0.05***	-0.27***	0.39***	0.02***	0.89***	-
		(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	
	(5)	0.85***	-	-	-	-	-	0.87***	2.73***
		(0.00)						(0.00)	(0.00)
	(6)	0.72***	0.21***	-0.05***	-0.24***	0.41***	-0.26***	0.81***	2.34***
		(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)

FE represents the average of the interaction between a dummy for female students and the question dummies, CE, SI, IT, and SMALL are dummies for Center region, South & Islands region, natives, and small class. *, **, and *** denote statistically significant at the 90%, 95%, and 99% level, respectively, standard errors in parentheses.

Table 23: RE logit estimates, 8th grade Italian exam

		$\hat{\bar{\xi}}$	FE	CE	SI	IT	$SMALL$	$\hat{\sigma}_\eta$	$\hat{\rho}$
Treatment	(1)	1.30***	-	-	-	-	-	0.89***	-
		(0.01)						(0.01)	
	(2)	1.27***	0.23***	-	-	-	-	0.93***	-
		(0.01)	(0.01)					(0.01)	
	(3)	1.32***	0.27***	0.03**	-0.07***	-	-	0.92***	-
		(0.01)	(0.01)	(0.01)	(0.01)			(0.01)	
	(4)	0.96***	0.28***	0.01	-0.14***	0.50***	0.02***	0.88***	-
		(0.02)	(0.01)	(0.01)	(0.01)	(0.02)	(0.00)	(0.01)	
	(5)	1.31***	-	-	-	-	-	0.85***	3.80***
		(0.00)						(0.00)	(0.01)
	(6)	0.99***	0.28***	0.05***	-0.07***	0.49***	-0.26***	0.79***	2.98***
		(0.01)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.01)
Control	(1)	1.35***	-	-	-	-	-	0.88***	-
		(0.00)						(0.00)	
	(2)	1.33***	0.24***	-	-	-	-	0.92***	-
		(0.00)	(0.00)					(0.00)	
	(3)	1.33***	0.29***	0.04***	0.00***	-	-	0.91***	-
		(0.00)	(0.00)	(0.00)	(0.00)			(0.00)	
	(4)	1.01***	0.29***	0.03***	-0.05***	0.42***	0.02***	0.88***	-
		(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	
	(5)	1.37***	-	-	-	-	-	0.83***	3.66***
		(0.00)						(0.00)	(0.00)
	(6)	1.07******	0.29***	0.03***	0.01***	0.41***	-0.23***	0.80***	2.94***
		(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)

FE represents the average of the interaction between a dummy for female students and the question dummies, CE, SI, IT, and SMALL are dummies for Center region, South & Islands region, natives, and small class. *, **, and *** denote statistically significant at the 90%, 95%, and 99% level, respectively, standard errors in parentheses.

Table 24: RE logit estimates, 10th grade Italian exam

		$\hat{\bar{\xi}}$	FE	CE	SI	IT	$SMALL$	$\hat{\sigma}_\eta$	$\hat{\rho}$
Treatment	(1)	0.47***	-	-	-	-	-	0.97***	-
		(0.00)						(0.01)	
	(2)	0.60***	0.05***	-	-	-	-	0.92***	-
		(0.01)	(0.01)					(0.01)	
	(3)	0.83***	0.11***	-0.11***	-0.24***	-	-	0.93***	-
		(0.01)	(0.01)	(0.01)	(0.01)			(0.01)	
	(4)	0.88***	0.08***	-0.20***	-0.45***	0.42***	0.04***	0.81***	-
		(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.00)	(0.00)	
	(5)	0.52***	-	-	-	-	-	0.91***	3.77***
		(0.00)						(0.00)	(0.01)
	(6)	0.99***	0.09***	-0.23***	-0.50***	0.42***	-0.48***	0.75***	3.27***
		(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.01)
Control	(1)	0.66***	-	-	-	-	-	1.04***	-
		(0.00)						(0.00)	
	(2)	0.76***	0.04***	-	-	-	-	1.01***	-
		(0.00)	(0.00)					(0.00)	
	(3)	0.89***	0.10***	-0.07***	-0.11***	-	-	1.01***	-
		(0.00)	(0.00)	(0.00)	(0.00)			(0.00)	
	(4)	0.92***	0.11***	-0.13***	-0.20***	0.19***	0.02***	0.95***	-
		(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	
	(5)	0.78***	-	-	-	-	-	0.97***	2.85***
		(0.00)						(0.00)	(0.00)
	(6)	0.80***	0.20***	-0.13***	-0.25***	0.37***	-0.30***	1.03***	1.07***
		(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)

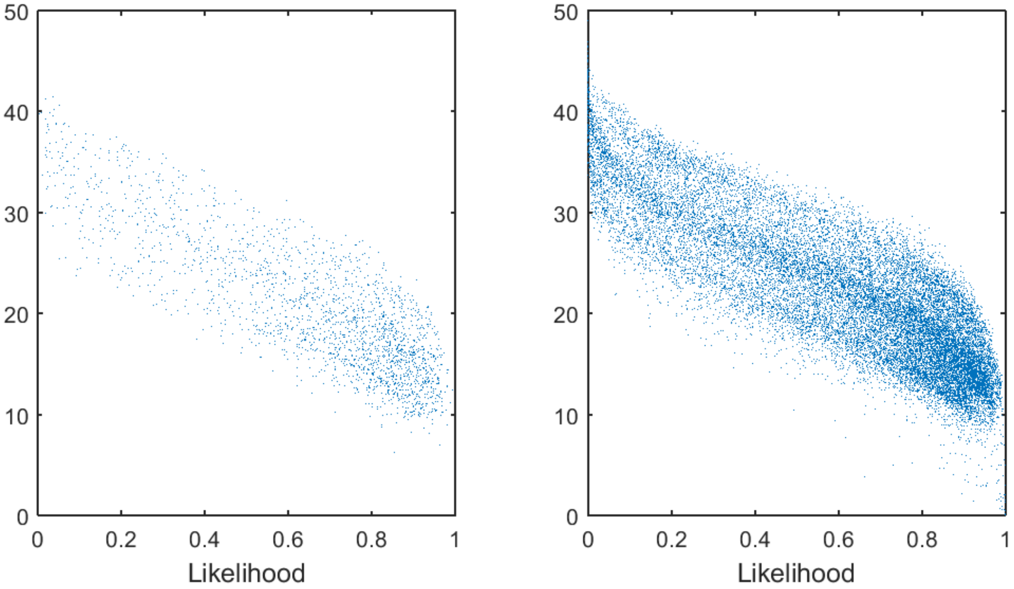
FE represents the average of the interaction between a dummy for female students and the question dummies, CE, SI, IT, and SMALL are dummies for Center region, South & Islands region, natives, and small class. *, **, and *** denote statistically significant at the 90%, 95%, and 99% level, respectively, standard errors in parentheses.

Table 25: 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
TR	0.53	0.55	0.50	0.27	0.54	0.18	0.58	0.22	0.60	0.62
CO	0.29	0.32	0.53	0.41	0.49	0.25	0.57	0.54	0.60	0.35

TR and CO denote treatment and control groups, respectively. 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 12: Likelihood and mean test scores



The left and right figures respectively show the scatter plot of the estimated likelihood of the test scores of each class (equation 7) and the class mean test scores for the treatment and control groups.

Figure 13: Correction for cheating, provincial variation, 2nd grade mathematics exam

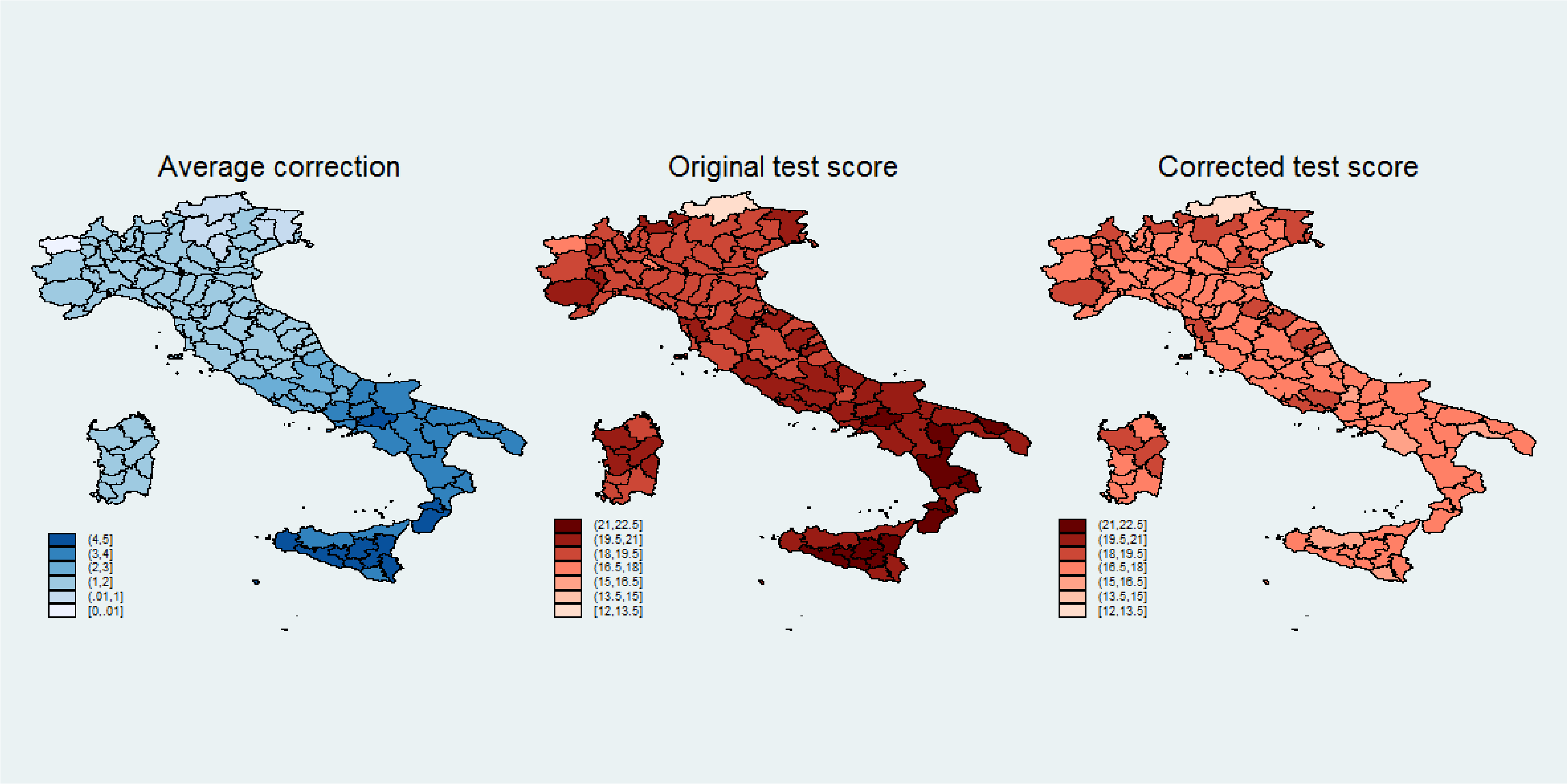


Figure 14: Correction for cheating, provincial variation, 5th grade mathematics exam

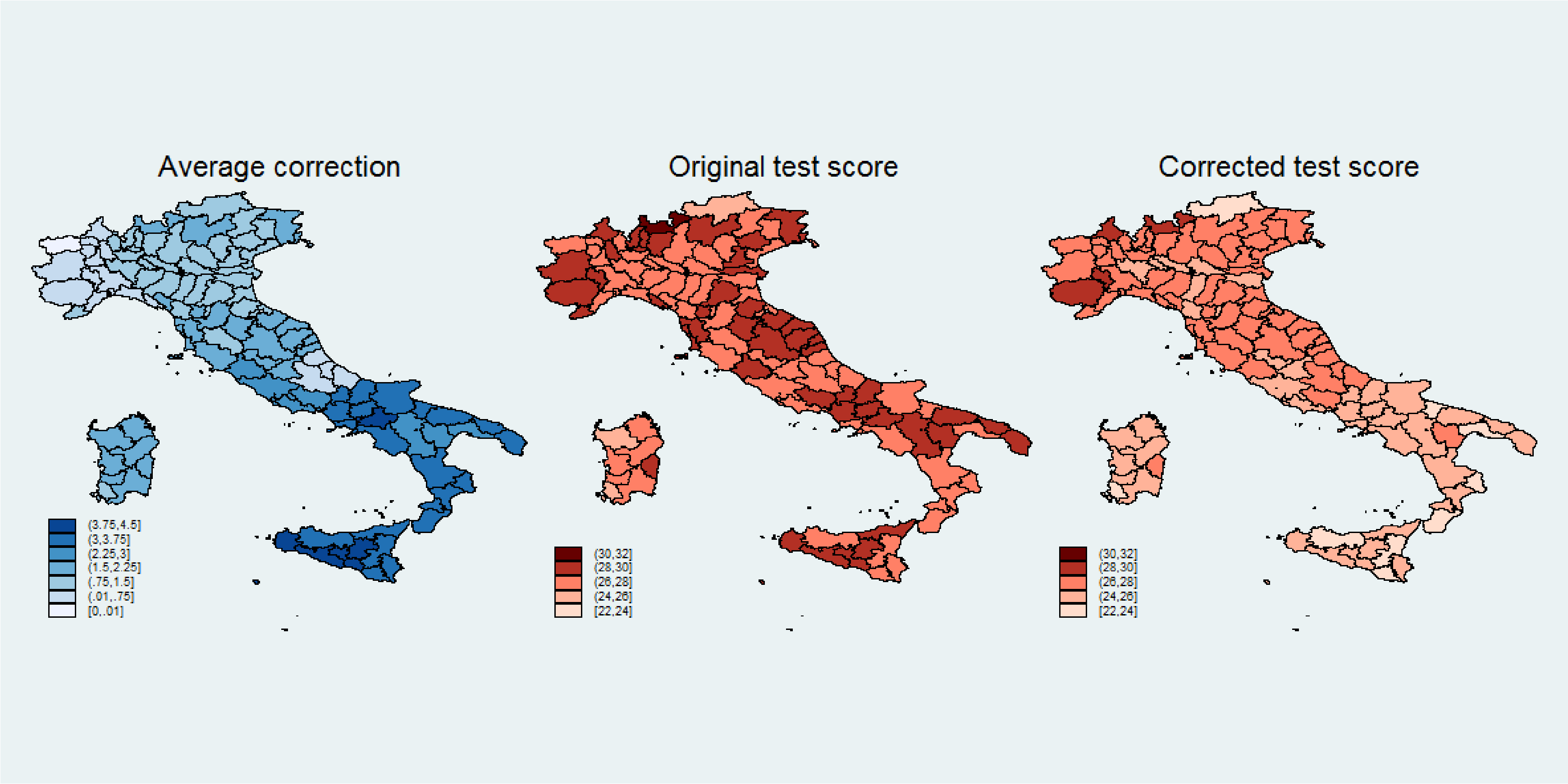


Figure 15: Correction for cheating, provincial variation, 6th grade mathematics exam

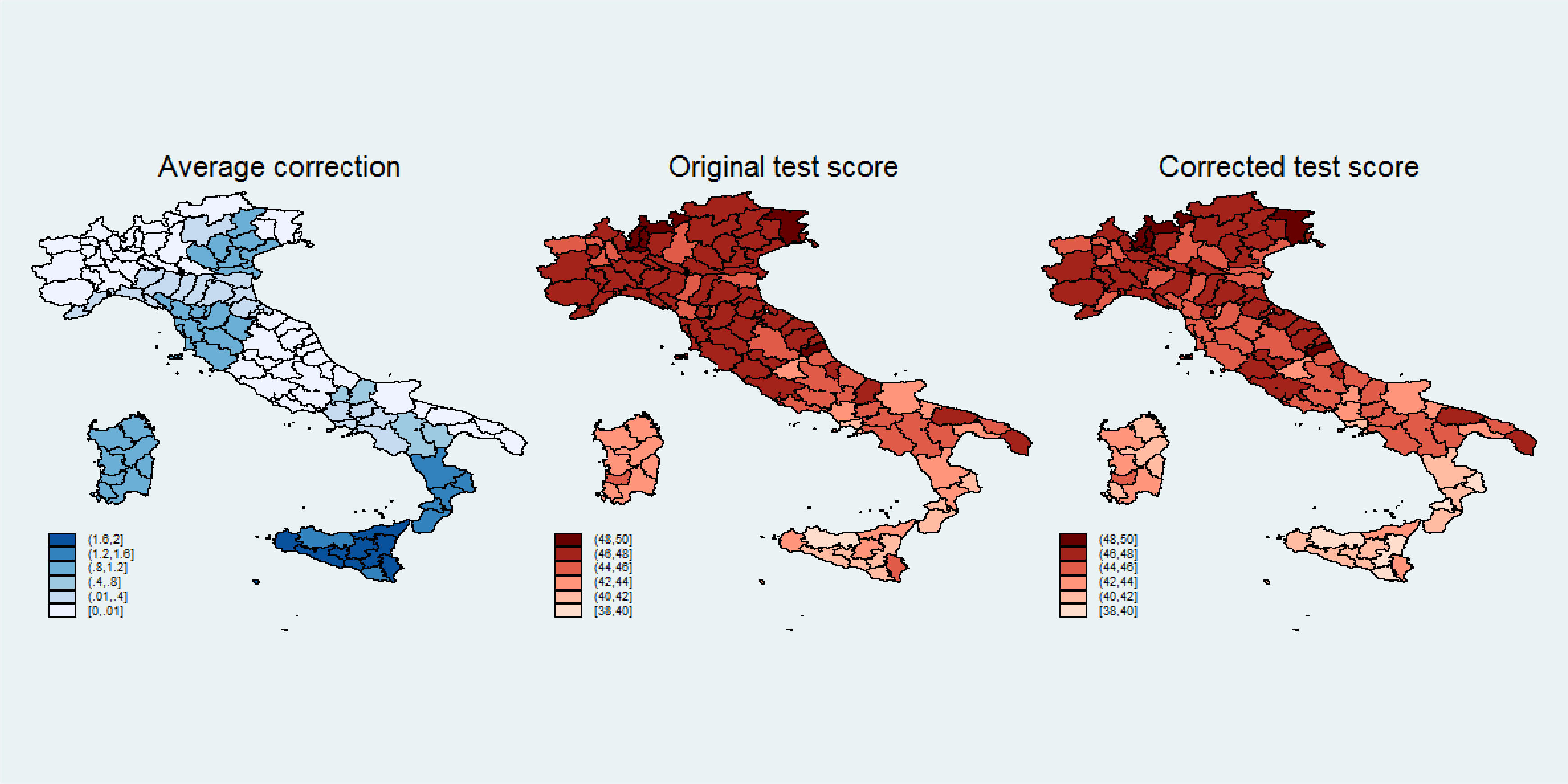


Figure 16: Correction for cheating, provincial variation, 8th grade mathematics exam

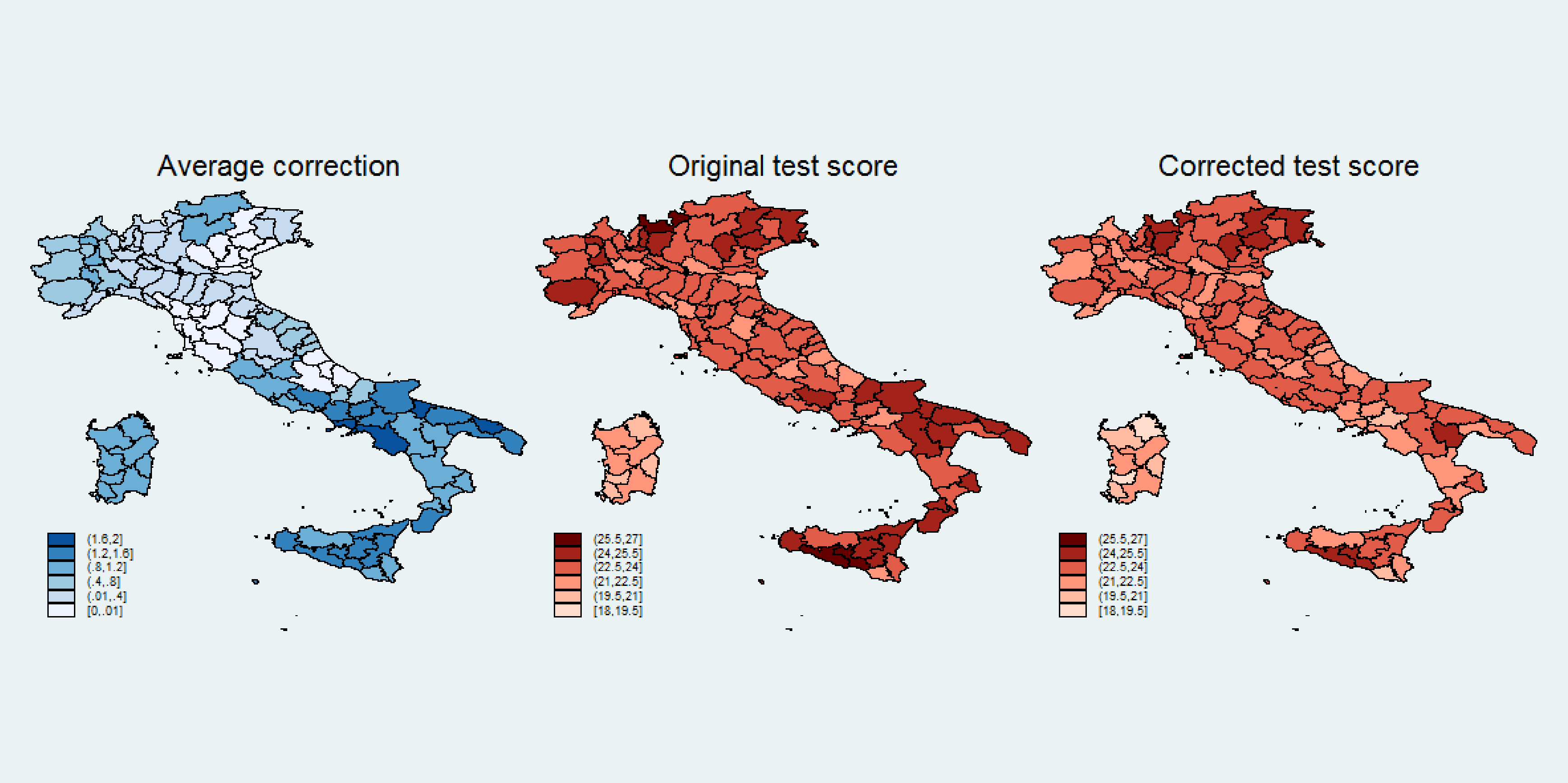


Figure 17: Correction for cheating, provincial variation, 2nd grade Italian exam

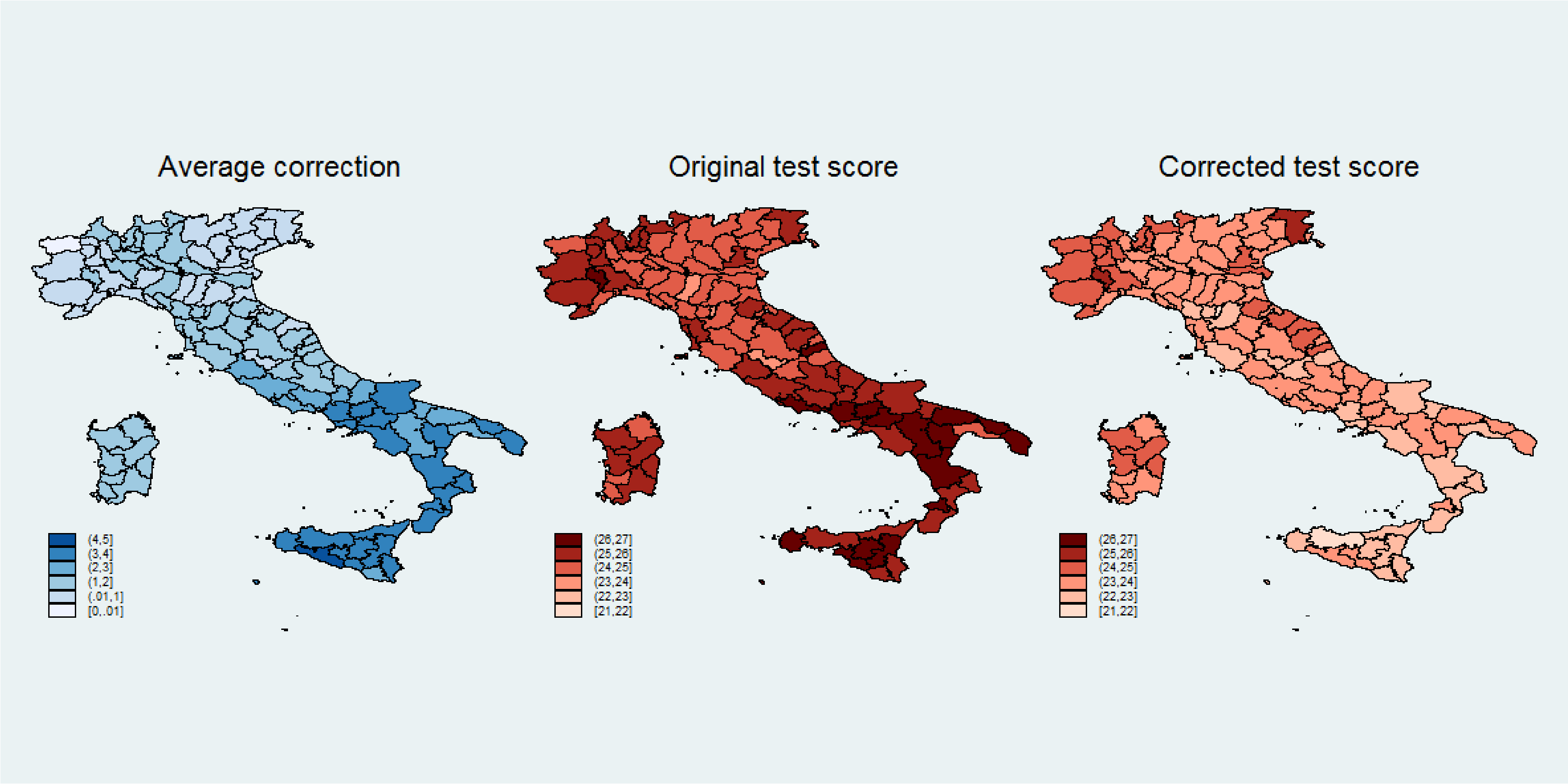


Figure 18: Correction for cheating, provincial variation, 5th grade Italian exam

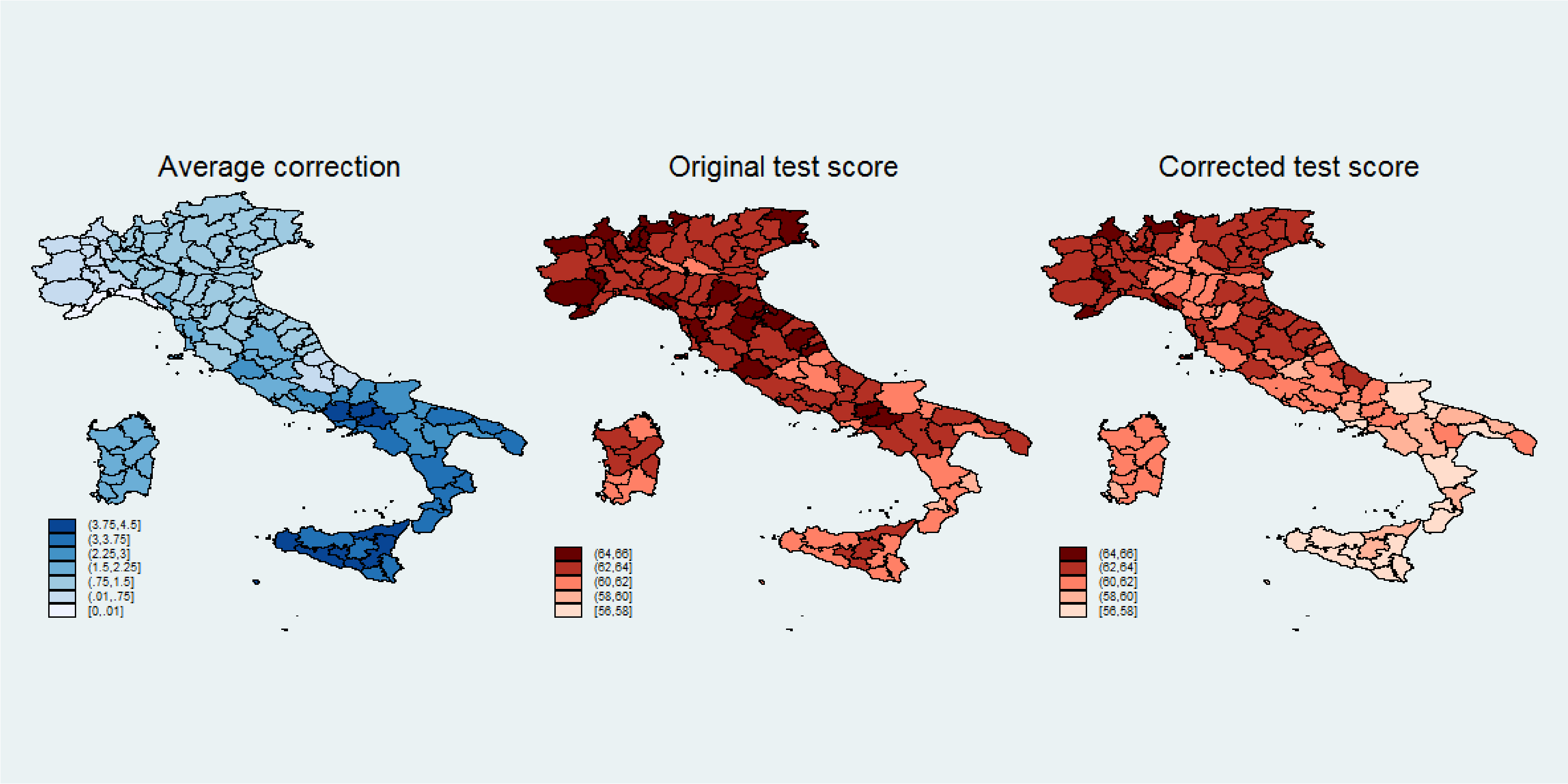


Figure 19: Correction for cheating, provincial variation, 6th grade Italian exam

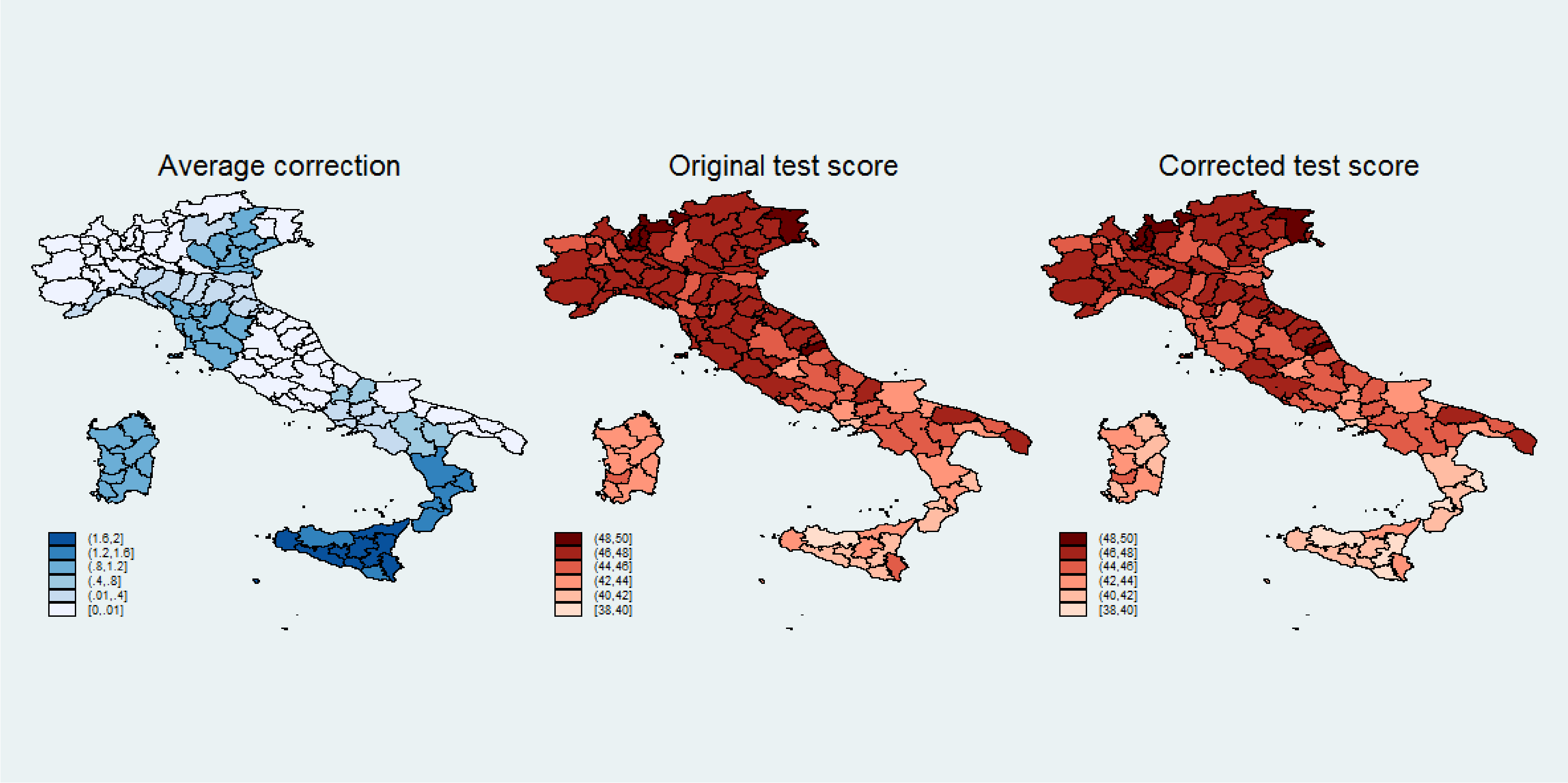


Figure 20: Correction for cheating, provincial variation, 8th grade Italian exam

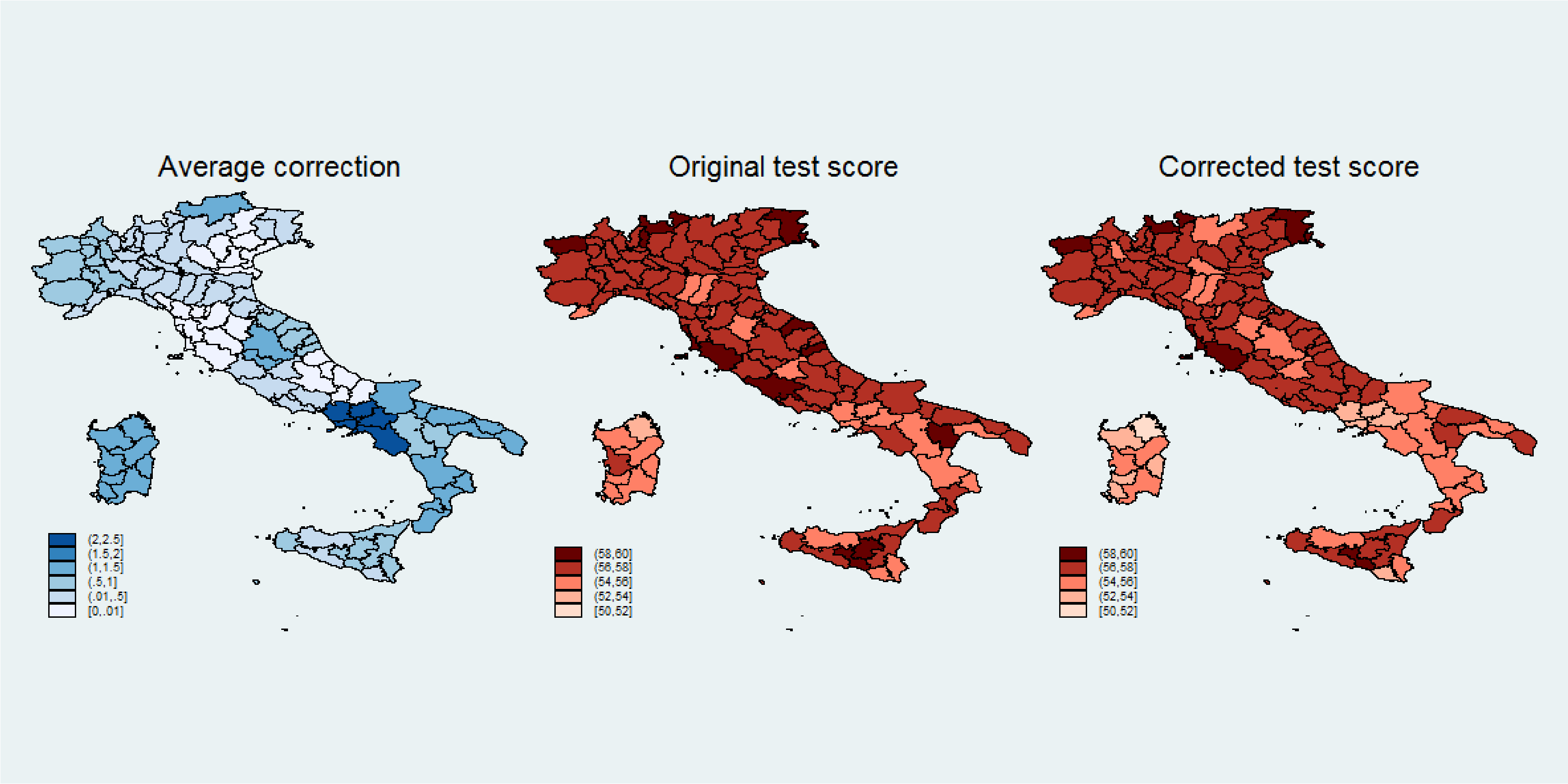
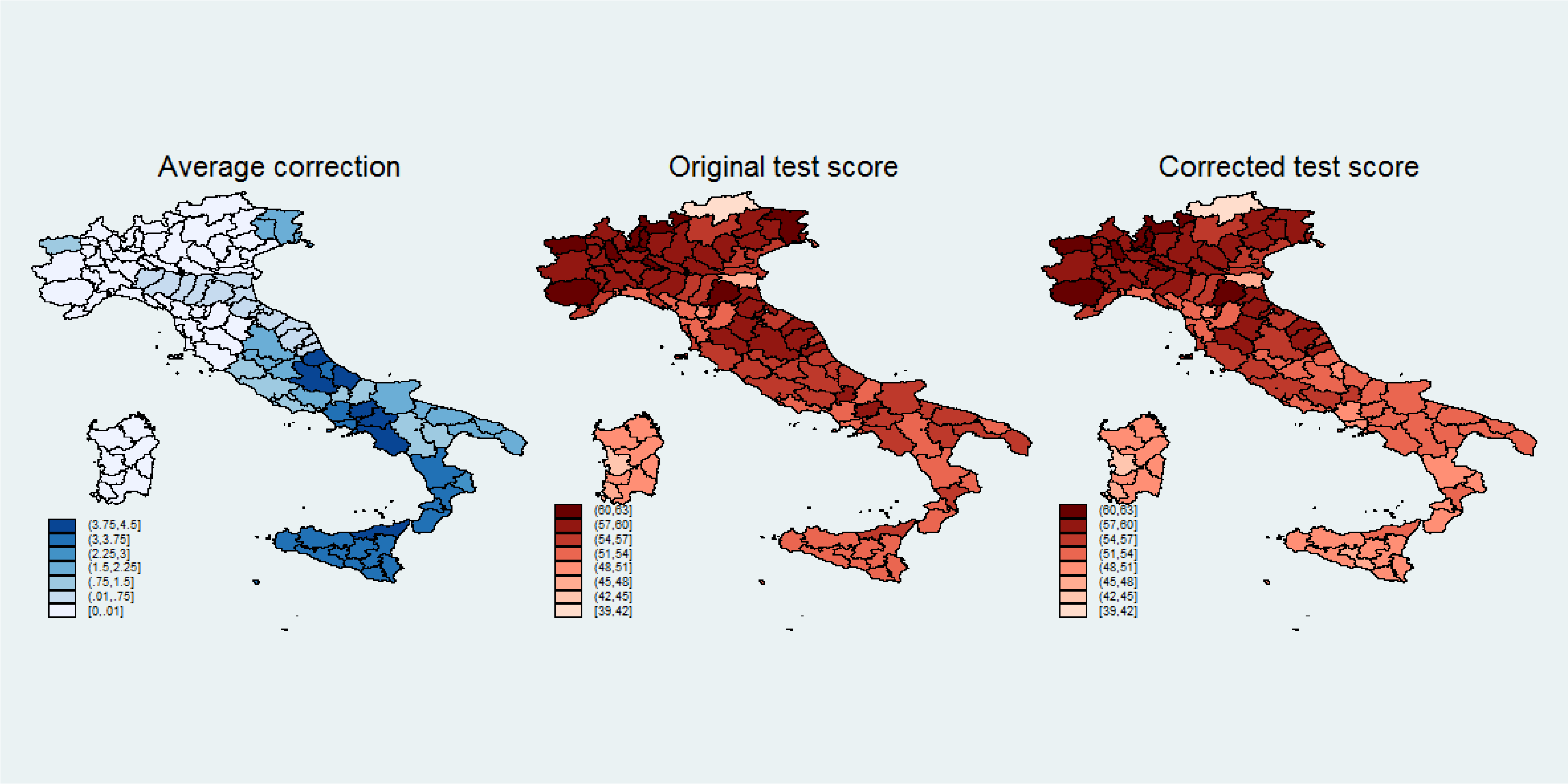


Figure 21: Correction for cheating, provincial variation, 10th grade Italian exam



RECENTLY PUBLISHED "TEMI" (*)

N. 1023 – *Understanding policy rates at the zero lower bound: insights from a Bayesian shadow rate model*, by Marcello Pericoli and Marco Taboga (July 2015).

N. 1024 – *Accessorizing. The effect of union contract renewals on consumption*, by Effrosyni Adamopoulou and Roberta Zizza (July 2015).

N. 1025 – *Tail comovement in option-implied inflation expectations as an indicator of anchoring*, by Sara Cecchetti, Filippo Natoli and Laura Sigalotti (July 2015).

N. 1026 – *Follow the value added: bilateral gross export accounting*, by Alessandro Borin and Michele Mancini (July 2015).

N. 1027 – *On the conditional distribution of euro area inflation forecast*, by Fabio Busetti, Michele Caivano and Lisa Rodano (July 2015).

N. 1028 – *The impact of CCPs' margin policies on repo markets*, by Arianna Miglietta, Cristina Picillo and Mario Pietrunti (September 2015).

N. 1029 – *European structural funds during the crisis: evidence from Southern Italy*, by Emanuele Ciani and Guido de Blasio (September 2015).

N. 1030 – *Female employment and pre-kindergarten: on the uninteded effects of an Italian reform*, by Francesca Carta and Lucia Rizzica (September 2015).

N. 1031 – *The predictive content of business survey indicators: evidence from SIGE*, by Tatiana Cesaroni and Stefano Iezzi (September 2015).

N. 1032 – *Sovereign debt exposure and the bank lending channel: impact on credit supply and the real economy*, by Margherita Bottero, Simone Lenzu and Filippo Mezzanotti (September 2015).

N. 1033 – *Does trend inflation make a difference?*, by Michele Loberto and Chiara Perricone (September 2015).

N. 1034 – *Procyclicality of credit rating systems: how to manage it*, by Tatiana Cesaroni (September 2015).

N. 1035 – *The time varying effect of oil price shocks on euro-area exports*, by Marianna Riggi and Fabrizio Venditti (September 2015).

N. 1036 – *Domestic and international macroeconomic effects of the Eurosystem expanded asset purchase programme*, by Pietro Cova, Patrizio Pagano and Massimiliano Pisani (September 2015).

N. 1037 – *Deconstructing the gains from trade: selection of industries vs. reallocation of workers*, by Stefano Bolatto and Massimo Sbracia (November 2015).

N. 1038 – *Young adults living with their parents and the influence of peers*, by Effrosyni Adamopoulou and Ezgi Kaya (November 2015).

N. 1039 – *Shoe-leather costs in the euro area and the foreign demand for euro banknotes*, by Alessandro Calza and Andrea Zaghini (November 2015).

N. 1040 – *The macroeconomic effects of low and falling inflation at the zero lower bound*, by Stefano Neri and Alessandro Notarpietro (November 2015).

N. 1041 – *The use of fixed-term contracts and the (adverse) selection of public sector workers*, by Lucia Rizzica (November 2015).

N. 1042 – *Multitask agents and incentives: the case of teaching and research for university professors*, by Marta De Philippis (November 2015).

N. 1043 – *Exposure to media and corruption perceptions*, by Lucia Rizzica and Marco Tonello (November 2015).

N. 1044 – *The supply side of household finance*, by Gabriele Foà, Leonardo Gambacorta, Luigi Guiso and Paolo Emilio Mistrulli (November 2015).

(*) Requests for copies should be sent to:
Banca d'Italia – Servizio Studi di struttura economica e finanziaria – Divisione Biblioteca e Archivio storico – Via Nazionale, 91 – 00184 Rome – (fax 0039 06 47922059). They are available on the Internet www.bancaditalia.it.

2013

A. Mercatanti, *A likelihood-based analysis for relaxing the exclusion restriction in randomized experiments with imperfect compliance*, Australian and New Zealand Journal of Statistics, v. 55, 2, pp. 129-153, **TD No. 683 (August 2008).**

F. Cingano and P. Pinotti, *Politicians at work. The private returns and social costs of political connections*, Journal of the European Economic Association, v. 11, 2, pp. 433-465, **TD No. 709 (May 2009).**

F. Busetti and J. Marcucci, *Comparing forecast accuracy: a Monte Carlo investigation*, International Journal of Forecasting, v. 29, 1, pp. 13-27, **TD No. 723 (September 2009).**

D. Dottori, S. I-Ling and F. Estevan, *Reshaping the schooling system: The role of immigration*, Journal of Economic Theory, v. 148, 5, pp. 2124-2149, **TD No. 726 (October 2009).**

A. Finicelli, P. Pagano and M. Sbracia, *Ricardian Selection*, Journal of International Economics, v. 89, 1, pp. 96-109, **TD No. 728 (October 2009).**

L. Monteforte and G. Moretti, *Real-time forecasts of inflation: the role of financial variables*, Journal of Forecasting, v. 32, 1, pp. 51-61, **TD No. 767 (July 2010).**

R. Giordano and P. Tommasino, *Public-sector efficiency and political culture*, FinanzArchiv, v. 69, 3, pp. 289-316, **TD No. 786 (January 2011).**

E. Gaiotti, *Credit availablility and investment: lessons from the "Great Recession"*, European Economic Review, v. 59, pp. 212-227, **TD No. 793 (February 2011).**

F. Nucci and M. Riggi, *Performance pay and changes in U.S. labor market dynamics*, Journal of Economic Dynamics and Control, v. 37, 12, pp. 2796-2813, **TD No. 800 (March 2011).**

G. Cappelletti, G. Guazzarotti and P. Tommasino, *What determines annuity demand at retirement?,* The Geneva Papers on Risk and Insurance – Issues and Practice, pp. 1-26, **TD No. 805 (April 2011).**

A. Accetturo e L. Infante, *Skills or Culture? An analysis of the decision to work by immigrant women in Italy,* IZA Journal of Migration, v. 2, 2, pp. 1-21, **TD No. 815 (July 2011).**

A. De Socio, *Squeezing liquidity in a "lemons market" or asking liquidity "on tap"*, Journal of Banking and Finance, v. 27, 5, pp. 1340-1358, **TD No. 819 (September 2011).**

S. Gomes, P. Jacquinot, M. Mohr and M. Pisani, *Structural reforms and macroeconomic performance in the euro area countries: a model-based assessment,* International Finance, v. 16, 1, pp. 23-44, **TD No. 830 (October 2011).**

G. Barone and G. de Blasio, *Electoral rules and voter turnout,* International Review of Law and Economics, v. 36, 1, pp. 25-35, **TD No. 833 (November 2011).**

O. Blanchard and M. Riggi, *Why are the 2000s so different from the 1970s? A structural interpretation of changes in the macroeconomic effects of oil prices*, Journal of the European Economic Association, v. 11, 5, pp. 1032-1052, **TD No. 835 (November 2011).**

R. Cristadoro and D. Marconi, *Household savings in China,* in G. Gomel, D. Marconi, I. Musu, B. Quintieri (eds), The Chinese Economy: Recent Trends and Policy Issues, Springer-Verlag, Berlin, **TD No. 838 (November 2011).**

A. Anzuini, M. J. Lombardi and P. Pagano, *The impact of monetary policy shocks on commodity prices*, International Journal of Central Banking, v. 9, 3, pp. 119-144, **TD No. 851 (February 2012).**

R. Gambacorta and M. Iannario, *Measuring job satisfaction with CUB models,* Labour, v. 27, 2, pp. 198-224, **TD No. 852 (February 2012).**

G. Ascari and T. Ropele, *Disinflation effects in a medium-scale new keynesian model: money supply rule versus interest rate rule,* European Economic Review, v. 61, pp. 77-100, **TD No. 867 (April 2012)**

E. Beretta and S. Del Prete, *Banking consolidation and bank-firm credit relationships: the role of geographical features and relationship characteristics,* Review of Economics and Institutions, v. 4, 3, pp. 1-46, **TD No. 901 (February 2013).**

M. Andini, G. de Blasio, G. Duranton and W. Strange, *Marshallian labor market pooling: evidence from Italy,* Regional Science and Urban Economics, v. 43, 6, pp.1008-1022, **TD No. 922 (July 2013).**

G. Sbrana and A. Silvestrini, *Forecasting aggregate demand: analytical comparison of top-down and bottom-up approaches in a multivariate exponential smoothing framework,* International Journal of Production Economics, v. 146, 1, pp. 185-98, **TD No. 929 (September 2013).**

A. Filippin, C. V, Fiorio and E. Viviano, *The effect of tax enforcement on tax morale,* European Journal of Political Economy, v. 32, pp. 320-331, **TD No. 937 (October 2013).**

2014

G. M. TOMAT, *Revisiting poverty and welfare dominance*, Economia pubblica, v. 44, 2, 125-149, **TD No. 651 (December 2007).**

M. TABOGA, *The riskiness of corporate bonds*, Journal of Money, Credit and Banking, v.46, 4, pp. 693-713, **TD No. 730 (October 2009).**

G. MICUCCI and P. ROSSI, *Il ruolo delle tecnologie di prestito nella ristrutturazione dei debiti delle imprese in crisi*, in A. Zazzaro (a cura di), Le banche e il credito alle imprese durante la crisi, Bologna, Il Mulino, **TD No. 763 (June 2010).**

F. D'AMURI, *Gli effetti della legge 133/2008 sulle assenze per malattia nel settore pubblico,* Rivista di politica economica, v. 105, 1, pp. 301-321, **TD No. 787 (January 2011).**

R. BRONZINI and E. IACHINI, *Are incentives for R&D effective? Evidence from a regression discontinuity approach,* American Economic Journal : Economic Policy, v. 6, 4, pp. 100-134, **TD No. 791 (February 2011).**

P. ANGELINI, S. NERI and F. PANETTA, *The interaction between capital requirements and monetary policy*, Journal of Money, Credit and Banking, v. 46, 6, pp. 1073-1112, **TD No. 801 (March 2011).**

M. BRAGA, M. PACCAGNELLA and M. PELLIZZARI, *Evaluating students' evaluations of professors,* Economics of Education Review, v. 41, pp. 71-88, **TD No. 825 (October 2011).**

M. FRANCESE and R. MARZIA, *Is there Room for containing healthcare costs? An analysis of regional spending differentials in Italy,* The European Journal of Health Economics, v. 15, 2, pp. 117-132, **TD No. 828 (October 2011).**

L. GAMBACORTA and P. E. MISTRULLI, *Bank heterogeneity and interest rate setting: what lessons have we learned since Lehman Brothers?,* Journal of Money, Credit and Banking, v. 46, 4, pp. 753-778, **TD No. 829 (October 2011).**

M. PERICOLI, *Real term structure and inflation compensation in the euro area*, International Journal of Central Banking, v. 10, 1, pp. 1-42, **TD No. 841 (January 2012).**

E. GENNARI and G. MESSINA, *How sticky are local expenditures in Italy? Assessing the relevance of the flypaper effect through municipal data,* International Tax and Public Finance, v. 21, 2, pp. 324-344, **TD No. 844 (January 2012).**

V. DI GACINTO, M. GOMELLINI, G. MICUCCI and M. PAGNINI, *Mapping local productivity advantages in Italy: industrial districts, cities or both?*, Journal of Economic Geography, v. 14, pp. 365–394, **TD No. 850 (January 2012).**

A. ACCETTURO, F. MANARESI, S. MOCETTI and E. OLIVIERI, *Don't Stand so close to me: the urban impact of immigration,* Regional Science and Urban Economics, v. 45, pp. 45-56, **TD No. 866 (April 2012).**

M. PORQUEDDU and F. VENDITTI, *Do food commodity prices have asymmetric effects on euro area inflation,* Studies in Nonlinear Dynamics and Econometrics, v. 18, 4, pp. 419-443, **TD No. 878 (September 2012).**

S. FEDERICO, *Industry dynamics and competition from low-wage countries: evidence on Italy*, Oxford Bulletin of Economics and Statistics, v. 76, 3, pp. 389-410, **TD No. 879 (September 2012).**

F. D'AMURI and G. PERI, *Immigration, jobs and employment protection: evidence from Europe before and during the Great Recession,* Journal of the European Economic Association, v. 12, 2, pp. 432-464, **TD No. 886 (October 2012).**

M. TABOGA, *What is a prime bank? A euribor-OIS spread perspective,* International Finance, v. 17, 1, pp. 51-75, **TD No. 895 (January 2013).**

G. CANNONE and D. FANTINO, *Evaluating the efficacy of european regional funds for R&D,* Rassegna italiana di valutazione, v. 58, pp. 165-196, **TD No. 902 (February 2013).**

L. GAMBACORTA and F. M. SIGNORETTI, *Should monetary policy lean against the wind? An analysis based on a DSGE model with banking,* Journal of Economic Dynamics and Control, v. 43, pp. 146-74, **TD No. 921 (July 2013).**

M. BARIGOZZI, CONTI A.M. and M. LUCIANI, *Do euro area countries respond asymmetrically to the common monetary policy?,* Oxford Bulletin of Economics and Statistics, v. 76, 5, pp. 693-714, **TD No. 923 (July 2013).**

U. ALBERTAZZI and M. BOTTERO, *Foreign bank lending: evidence from the global financial crisis,* Journal of International Economics, v. 92, 1, pp. 22-35, **TD No. 926 (July 2013).**

R. De Bonis and A. Silvestrini, *The Italian financial cycle: 1861-2011,* Cliometrica, v.8, 3, pp. 301-334, **TD No. 936 (October 2013).**

G. Barone and S. Mocetti, *Natural disasters, growth and institutions: a tale of two earthquakes,* Journal of Urban Economics, v. 84, pp. 52-66, **TD No. 949 (January 2014).**

D. Pianeselli and A. Zaghini, *The cost of firms' debt financing and the global financial crisis,* Finance Research Letters, v. 11, 2, pp. 74-83, **TD No. 950 (February 2014).**

J. Li and G. Zinna, *On bank credit risk: sytemic or bank-specific? Evidence from the US and UK,* Journal of Financial and Quantitative Analysis, v. 49, 5/6, pp. 1403-1442, **TD No. 951 (February 2015).**

A. Zaghini, *Bank bonds: size, systemic relevance and the sovereign,* International Finance, v. 17, 2, pp. 161-183, **TD No. 966 (July 2014).**

G. Sbrana and A. Silvestrini, *Random switching exponential smoothing and inventory forecasting,* International Journal of Production Economics, v. 156, 1, pp. 283-294, **TD No. 971 (October 2014).**

M. Silvia, *Does issuing equity help R&D activity? Evidence from unlisted Italian high-tech manufacturing firms,* Economics of Innovation and New Technology, v. 23, 8, pp. 825-854, **TD No. 978 (October 2014).**

2015

M. Bugamelli, S. Fabiani and E. Sette, *The age of the dragon: the effect of imports from China on firm-level prices*, Journal of Money, Credit and Banking, v. 47, 6, pp. 1091-1118, **TD No. 737 (January 2010).**

R. Bronzini, *The effects of extensive and intensive margins of FDI on domestic employment: microeconomic evidence from Italy*, B.E. Journal of Economic Analysis & Policy, v. 15, 4, pp. 2079-2109, **TD No. 769 (July 2010).**

G. Bulligan, M. Marcellino and F. Venditti, *Forecasting economic activity with targeted predictors,* International Journal of Forecasting, v. 31, 1, pp. 188-206, **TD No. 847 (February 2012).**

A. Ciarlone, *House price cycles in emerging economies,* Studies in Economics and Finance, v. 32, 1, **TD No. 863 (May 2012).**

D. Fantino, A. Mori and D. Scalise, *Collaboration between firms and universities in Italy: the role of a firm's proximity to top-rated departments,* Rivista Italiana degli economisti, v. 1, 2, pp. 219-251, **TD No. 884 (October 2012).**

D. Depalo, R. Giordano and E. Papapetrou, *Public-private wage differentials in euro area countries: evidence from quantile decomposition analysis,* Empirical Economics, v. 49, 3, pp. 985-1115, **TD No. 907 (April 2013).**

G. Barone and G. Narciso, *Organized crime and business subsidies: Where does the money go?,* Journal of Urban Economics, v. 86, pp. 98-110, **TD No. 916 (June 2013).**

P. Alessandri and B. Nelson, *Simple banking: profitability and the yield curve,* Journal of Money, Credit and Banking, v. 47, 1, pp. 143-175, **TD No. 945 (January 2014).**

M. Taneli and B. Ohl, *Information acquisition and learning from prices over the business cycle,* Journal of Economic Theory, 158 B, pp. 585–633, **TD No. 946 (January 2014).**

R. Aaberge and A. Brandolini, *Multidimensional poverty and inequality,* in A. B. Atkinson and F. Bourguignon (eds.), Handbook of Income Distribution, Volume 2A, Amsterdam, Elsevier, **TD No. 976 (October 2014).**

V. Cuciniello and F. M. Signoretti, *Large banks,loan rate markup and monetary policy,* International Journal of Central Banking, v. 11, 3, pp. 141-177, **TD No. 987 (November 2014).**

M. Fratzscher, D. Rimec, L. Sarnob and G. Zinna, *The scapegoat theory of exchange rates: the first tests,* Journal of Monetary Economics, v. 70, 1, pp. 1-21, **TD No. 991 (November 2014).**

A. Notarpietro and S. Siviero, *Optimal monetary policy rules and house prices: the role of financial frictions,* Journal of Money, Credit and Banking, v. 47, S1, pp. 383-410, **TD No. 993 (November 2014).**

R. Antonietti, R. Bronzini and G. Cainelli, *Inward greenfield FDI and innovation,* Economia e Politica Industriale, v. 42, 1, pp. 93-116, **TD No. 1006 (March 2015).**

T. Cesaroni, *Procyclicality of credit rating systems: how to manage it,* Journal of Economics and Business, v. 82. pp. 62-83, **TD No. 1034 (October 2015).**

M. Riggi and F. Venditti, *The time varying effect of oil price shocks on euro-area exports,* Journal of Economic Dynamics and Control, v. 59, pp. 75-94, **TD No. 1035 (October 2015).**

FORTHCOMING

G. DE BLASIO, D. FANTINO and G. PELLEGRINI, *Evaluating the impact of innovation incentives: evidence from an unexpected shortage of funds,* Industrial and Corporate Change, **TD No. 792 (February 2011).**

A. DI CESARE, A. P. STORK and C. DE VRIES, *Risk measures for autocorrelated hedge fund returns,* Journal of Financial Econometrics, **TD No. 831 (October 2011).**

E. BONACCORSI DI PATTI and E. SETTE, *Did the securitization market freeze affect bank lending during the financial crisis? Evidence from a credit register,* Journal of Financial Intermediation, **TD No. 848 (February 2012).**

M. MARCELLINO, M. PORQUEDDU and F. VENDITTI, *Short-Term GDP Forecasting with a mixed frequency dynamic factor model with stochastic volatility,* Journal of Business & Economic Statistics, **TD No. 896 (January 2013).**

M. ANDINI and G. DE BLASIO, *Local development that money cannot buy: Italy's Contratti di Programma,* Journal of Economic Geography, **TD No. 915 (June 2013).**

F BRIPI, *The role of regulation on entry: evidence from the Italian provinces,* World Bank Economic Review, **TD No. 932 (September 2013).**

G. ALBANESE, G. DE BLASIO and P. SESTITO, *My parents taught me. evidence on the family transmission of values,* Journal of Population Economics, **TD No. 955 (March 2014).**

A. L. MANCINI, C. MONFARDINI and S. PASQUA, *Is a good example the best sermon? Children's imitation of parental reading,* Review of Economics of the Household, **TD No. 958 (April 2014).**

R. BRONZINI and P. PISELLI, *The impact of R&D subsidies on firm innovation,* Research Policy, **TD No. 960 (April 2014).**

L. BURLON, *Public expenditure distribution, voting, and growth,* Journal of Public Economic Theory, **TD No. 961 (April 2014).**

L. BURLON and M. VILALTA-BUFI, *A new look at technical progress and early retirement,* IZA Journal of Labor Policy, **TD No. 963 (June 2014).**

A. BRANDOLINI and E. VIVIANO, *Behind and beyond the (headcount) employment rate,* Journal of the Royal Statistical Society: Series A, **TD No. 965 (July 2015).**

G. ZINNA, *Price pressures on UK real rates: an empirical investigation,* Review of Finance, **TD No. 968 (July 2014).**

A. CALZA and A. ZAGHINI, *Shoe-leather costs in the euro area and the foreign demand for euro banknotes,* International Journal of Central Banking, **TD No. 1039 (December 2015).**