# When the Cat Is Near, the Mice Won't Play:
## The Effect of External Examiners in Italian Schools[*]

Marco Bertoni
University of Padova and Centre for Economic Performance, LSE

Giorgio Brunello[**]
University of Padova, IZA and CEsifo

Lorenzo Rocco
University of Padova

## Abstract

We use a natural experiment to show that the presence of an external examiner has both a direct and an indirect negative effect on the performance of monitored classes in standardized educational tests. The direct effect is the difference in the test performance between classes of the same school with and without external examiners. The indirect effect is the difference in performance between un-monitored classes in schools with an external examiner and un-monitored classes in schools without external monitoring. We find that the overall effect of having an external examiner in the class is to reduce the proportion of correct answers by 5.5 to 8.5% - depending on the grade and the test - with respect to classes in schools with no external monitor. The direct and indirect effects range between 4.3 and 6.6% and between 1.2 and 1.9% respectively. Using additional supporting evidence, we argue that the negative impact of the presence of an external examiner on measured test scores is due to reduced cheating (by students and/or teachers) rather than to the negative effects of anxiety or distraction from having a stranger in the class.

Keywords: education, testing, external monitoring, indirect treatment effects.
JEL codes: C31, H52, I2.

[*] The authors are grateful to Erich Battistin, Thomas Breda, Daniele Checchi, David Figlio, Ifty Hussain, Edwin Leuven, Marco Manacorda, Guy Michaels, Hessel Oosterbeek, Steve Pischke, Olmo Silva and to the audiences at the 2012 LSE-CEP Annual Conference, the 2012 HECER Economics of Education Summer Meeting in Helsinki, the APPAM-INVALSI conference "Improving Education through Accountability and Evaluation" in Rome, and at seminars in Padova and CIDE (Bertinoro) for comments and suggestions. We also thank Patrizia Falzetti, Roberto Ricci and Paolo Sestito (INVALSI) for helping us with data collection and for explaining several technical features of the administration of the SNV tests. Financial support by the Ministry of Italian Universities (PRIN contract n. 2009MAATFS_002) is gratefully acknowledged. All errors are our own.

[**] Corresponding author. Address: via del Santo 33, 35100 Padova. E-mail: giorgio.brunello@unipd.it

## 1. Introduction

A problem with test – based accountability systems in education is that they generate incentives for teachers, students and school administrators to "game" the system in order to obtain better results. The manipulation of test outcomes generates efficiency losses both when these outcomes are used to allocate resources to schools and teachers and when – more modestly – they provide valuable benchmarking information which can affect the choices of schools and their stakeholders.

One mechanism for inflating test scores is outright cheating. Empirical analysis of cheating behaviour is scarce[1]. In their influential study, Jacob and Levitt (2003) develop an algorithm for detecting teachers' cheating that combines information on unexpected test score fluctuations and suspicious patterns of answers for students in a class. They find that a small fraction of Chicago teachers responded to accountability pressures by completing student examinations in an attempt to improve outcomes.

A possible deterrent of forms of cheating that may occur during the test – e.g. students copying from one another or teachers communicating the correct answers – or during the scoring – e.g. teachers changing students' answers or filling in missing answers – is monitoring by external examiners. External monitoring has costs and benefits. Costs increase with the desired level of coverage. Benefits depend both on the efficiency gain associated to a reduction in cheating and on how effective monitoring is in influencing test scores and reducing cheating.

In this paper, we estimate the impact of external monitoring on test scores, using a rather unique natural experiment designed by the Italian central test administrator (INVALSI), which assigned external examiners to randomly selected classes and schools with the task of monitoring students taking the test and reporting results[2]. We compare test outcomes in the classes with an external examiner with the outcomes in other classes, where the test was administered by a local teacher, and find that the rate of correct answers is lower in the former than in the latter. Using additional supporting evidence, we argue that the negative impact of the presence of an external examiner on measured test scores is due to reduced cheating (by students and/or teachers) rather than to the negative effects of anxiety or distraction from having a stranger in the class.

[1] See Figlio and Loeb, 2011, for a review of the recent literature.

[2] These tests are taken by the universe of primary second and fifth grade students. INVALSI sampled a number of classes and schools for external monitoring to obtain reliable data, speed up data collection and verification and prepare an annual report on the state of primary education in Italy.

Our study contributes to the literature on school accountability in two main directions. First, we show that the introduction of external examiners has a significant effect on measured test scores in an environment where there are incentives to manipulate results. Second, we document that the monitoring effects of having an external examiner spill over to un-monitored classes of the same school. We decompose the overall effect of external monitoring - which we measure as the difference in the average rate of correct answers in monitored classes and in classes of un-monitored schools - into a direct and an indirect effect. The direct effect is the difference in the test performance between classes with and without external examiners belonging to schools selected for external monitoring. The indirect effect is instead the difference in performance between un-monitored classes in a school with an external examiner and un-monitored classes in schools without external examiners.

We estimate that having an external examiner reduces the percentage of correct answers by 3.6 to 5.4 percentage points - depending on the grade and the test - which corresponds to 5.5 to 8.5% of the average score in classes belonging to schools with no external examiner. The estimated direct effect ranges from 2.8 to 4.2 percentage points (4.3 to 6.6%), and the residual indirect effect from 0.8 to 1.2 percentage points (1.2 to 1.9%). We discuss two alternative reasons why the effects of monitoring spread from the monitored class to the other classes in the same school. The first is that the presence of an external examiner in the school acts as a disciplinary device also on students and teachers in other classes of the same school because of the fear that the examiner may roam about. The second is that teachers dislike excessive dispersion in average class scores within the same school, because of the conflicts it could generate.

We find that the estimated overall effect of external supervision is significantly higher in the schools located in Southern Italy than in Northern schools and in schools where class size is smaller and the proportion of tenured teachers is higher. We show that territorial differences are associated to differences in social capital, even after controlling for territorial differences in GDP per capita and unemployment rates.

Studying the Italian experience with external monitoring has both advantages and disadvantages. The key advantage is that the random allocation of examiners to schools and classes allows us to bypass the selection problems that typically plague the evaluation of monitoring effects. A potential disadvantage is that in the Italian context there is limited accountability of schools and teachers. In this environment, the incentives to cheat may be weaker than in high-stakes contexts. In this case, our estimates can be interpreted as lower bounds of the effect of external monitoring in contexts where the incentives to manipulate results are stronger.

The paper is organized as follows: Section 2 reviews the relevant literature and Section 3 describes the design of the INVALSI test and the dataset. The empirical strategy is presented in Section 4. The main empirical results, a few robustness checks and extensions are reported in Section 5, 6 and 7, respectively. Conclusions follow.

## 2. Review of the literature

Aside from outright cheating studied by Jacob and Levitt (2003), the literature has identified several indirect ways that teachers and school administrators can use to manipulate student results. On the one hand, Jacob (2005), Figlio (2006), Figlio and Getzler (2006), Cullen and Reback (2006) and Hussain (2012) investigate whether schools engage in strategic manipulation of the composition of the pool of tested students by excluding low ability students, either by reclassifying them as disabled or by strategically using grade retention and disciplinary suspensions. On the other hand, Figlio and Winicki (2005) show that during testing periods some schools increase the caloric intake provided by school cafeterias so as to boost students' performance. Attempts to increase test scores by taking psycho-stimulant drugs are documented for the US by Bokhari and Schneider (2011), who show that the diagnosis of "attention deficit/hyperactivity disorder" is more frequent in states where there are stronger accountability laws.

To our knowledge, we are the first in this literature to investigate both the direct and the indirect effects of external examiners as deterrents of cheating in standardized tests. That indirect treatment effects can occur has been already pointed out by a broader literature. Heckman, Lalonde and Smith (1999), for instance, discuss how policy effects may spread to those not directly participating in the programme mainly because of general equilibrium or spill-over effects. Miguel and Kremer (2004) evaluate both direct and external effects of a Kenyan programme aimed at treating intestinal worms infection among primary school kids. In a similar fashion, Angelucci and De Giorgi (2009) evaluate the effects of *Progresa*, a Mexican aid programme based on cash transfers, and stress the importance of estimating indirect treatment effects on the ineligibles when there are social interactions between eligible and ineligible individuals.

## 3. The Design of INVALSI *Servizio Nazionale di Valutazione* (SNV) Tests and the Data

INVALSI[3] standardized tests in Italian and Math were introduced in Italian primary schools in 2008[4] to evaluate school productivity. The purposes of the evaluation[5] are to inform the central government about the general performance of the school system, and to offer schools a standardised reference to self-assess their strengths and weaknesses, using a value added approach. These tests are not formally high-stakes, because the allocation of resources to schools, the salary of teachers and the school career of students do not explicitly depend on test outcomes. Even so, pressure to perform well in the tests has been high because of the widespread expectations that they might be used at some point to evaluate teachers and schools. These expectations were fostered by the Ministry of Education, who in an intervention at the Lower House of the Italian Parliament (June 10th 2008) when the tests were introduced, made explicit reference to the need to establish within a few years a system of evaluation and incentives for teachers and schools based on student performance in the tests. Schools have an incentive to perform well also because results affect their reputation. Although the outcomes of the tests are not made public by INVALSI, schools have access to the results of their own students and can disclose them to parents and other stakeholders, in an effort to build their reputation and attract good students[6].

Since 2008 the tests have been administered every year. In this paper, we focus on the 2010 wave because of its peculiar design features. First, this wave was the first to test and collect data for the entire population of Italian primary school students in their second and fifth grade. Second, and most important for our purposes, in 2,000 randomly selected classes - out of a population of about 30,000 - the test was administered in the presence of an external examiner[7], who had two main tasks: a) be present in the class during the test and monitor its correct implementation; b) report student answers on the dedicated answer sheets and

[3] INVALSI is the National Institute for the Evaluation of the Education System, in charge of the design and administration of standardized education tests in Italy.

[4] See Law n.147 – 2007, and Ministry of Education and Research Decree n.74 and 76 – 2009.

[5] See article 2 of the INVALSI statute (Ministry of Education and Research Decree n. 11-2011) and the Ministry of Education and Research Directive n. 88-2011.

[6] "INVALSI does not provide public rankings of schools based on the outcomes of the test. The main purpose of the tests is to provide each single school and its stakeholders with valuable information that can help them to benchmark and improve their performance. Each school is free to advertise its own results, using the tools provided by the Ministry of Education…" (free translation by the authors of Ricci and Sestito, 2012).

[7] External examiners were selected by INVALSI and the Regional Schooling Authorities mainly among retired teachers and active teachers employed in non-primary schools. Each examiner was paid 200 euro per working day. Details on the criteria adopted to select external examiners are reported in the Appendix.

transmit them to INVALSI. In the other classes, the test was administered by teachers of the school (but not of the class and not in the subject tested), and reporting was done jointly with the teacher of the class. We use the random selection of classes as a natural experiment to estimate the effects of external monitoring on test outcomes.

Classes assigned to external monitoring were sampled using the same two-stage sampling scheme adopted by the IEA TIMSS survey, with stratification by region[8]. In the first stage, a pre-determined number of schools in each region were randomly selected by probabilistic sampling, with probability of inclusion proportional to school size, measured by the total number of students enrolled in the tested grades (second and fifth). In the second stage, and depending on school size, one or two classes for each tested grade within each treated school were selected by simple random sampling[9]. In each selected class, the test was administered in the presence of an external examiner. Table 1 shows for each grade the total and sampled number of primary schools, classes and pupils: about 18% of all primary schools and close to 7% of all classes and pupils in the second and fifth grade were selected to have an external examiner during the test.

We have access to the records containing the individual answers to the questions of the test taken in 2010 by students in classes with and without external monitoring, as they were transmitted to INVALSI by teachers and external examiners[10]. For each student, we also have information - provided by school offices - on her marks in Italian and Math during the semester before the test and on parental background. We add to these data the results of a questionnaire administered by INVALSI exclusively to fifth graders in order to collect additional information both on parental background and on student feelings and motivation during the tests. Finally, we have obtained from INVALSI additional information on school and class characteristics, including the number of students enrolled in each class and in each school for each tested grade, the proportion of tenured teachers in each school and, only for fifth grade students, an index of individual economic, social and cultural status (ESCS) [11].

[8] Region Valle d'Aosta and the Province of Bolzano autonomously decided to have all classes assigned to external monitoring. For this reason, we exclude them from our analysis. Our management of the data from the original to the final dataset is described in the Appendix.

[9] The average number of classes per school in sampled schools is 5.3, with a standard deviation of 1.9. Further details on the sampling procedure are reported in the Appendix.

[10] All questions were either multiple choice or open questions with a univocally correct answer, and were coded by INVALSI as correct, incorrect or missing.

[11] Available information includes the following variables: 1) at the school level: whether the school offers a full time schedule; 2) at the class level: class size measured both as the number of students enrolled in the class and as the number of students who were present at the test, full or part-time schedule (measured in term of the schedule of the median student in the class, to avoid measurement errors); 3) at the individual level: gender, place of birth, citizenship, attendance of pre-primary school, age, employment, education and nationality of

We test for successful randomization by checking whether observables are balanced between sampled and non-sampled schools and classes. Reflecting the sampling strategy adopted by INVALSI, we verify balancing in two steps, first between sampled and non-sampled schools and second between sampled and non-sampled classes within the set of sampled schools. Since sampling is stratified by region and sampling probabilities depend on school size, our balancing tests are conditional on regional effects, school size and, in the second step, the number of classes in the school. Although we have data for second and fifth graders, we focus hereinafter on the latter for brevity. Selected results for second graders are shown in the Appendix.

For each variable *X* in Table 2 we first test between – school balancing by running

$$X_j = \alpha + \beta t_j + \rho RD_r + \sigma RS_{rj} + \varepsilon_j \qquad (1)$$

where the subscript *r* is for the region where the school is located, $X_j$ is the average value of *X* in school *j*, $t_j$ is a dummy taking the value 1 if school *j* has been sampled and 0 otherwise, $RD_r$ is the full set of regional dummies, $RS_{rj}$ is school size interacted with regional dummies and $\varepsilon_j$ is the error term.

Next, we test within-schools balancing by running

$$X_{ij} = \gamma + \delta t_{ij} + \varsigma R_{rj} + v_{ij} \qquad (2)$$

where $X_{ij}$ is the average value of *X* in the class *i* of school *j*, $t_{ij}$ is a dummy that indicates whether class *i* in school *j* has been sampled and $R_{rj} = [RD_j, RS_{rj}, RC_{rj}]$ is a vector which includes the controls used in equation (1) as well as $RC_{rj}$, the number of fifth (or second) grade classes in school *j* interacted with regional dummies. We estimate equation (2) only for the classes belonging to the schools with external examiners and, since the second stage randomization took place within each school, we add school fixed effects and cluster standard errors at the school level.

parents. For fifth grade students only we have information on: whether the student at home has own bedroom, internet access, an encyclopaedia, own desk, a computer and a place for doing homework, the number of books in the house, the number of siblings, whether she lives with both parents or not, the language spoken at home, whether she gets help with her homework or not.

Table 3 reports the point estimates of the β and δ coefficients in (1) and (2) and their statistical significance. Since balancing is not attained for the number of students enrolled in a class, which is greater among treated classes, we include this variable as a covariate in all our regressions. Turning to individual variables, although for some covariates we detect statistically significant differences across the various groups, the point estimates show that these differences are very close to zero in almost all cases. Prudentially, we add these variables as covariates in our regressions to eliminate the risk of unbalancing and to increase precision[12].

## 4. Identification and Estimation

We define the following three potential outcomes at the class level: $Y_{00}$ if the class was assigned to a school with no external observer (an untreated class in an untreated school), $Y_{11}$ in case of direct monitoring (a treated class in a treated school) and $Y_{01}$ if the class was not monitored by an external examiner but belonged to a school where at least one other class was monitored (an untreated class in a treated school). By design, all classes of untreated schools are un-monitored.

Let the dummy variable $S_j$ take the value one if school $j$ has been assigned to school-level treatment (and zero otherwise) and the dummy $C_i$ take value one if class $i$ has been assigned to class-level treatment (and zero otherwise). The observed outcome $Y_{ij}$ for class $i$ in school $j$ can be represented in terms of potential outcomes as follows:

$$Y_{ij} = (1 - S_j)Y_{00} + S_j C_i Y_{11} + S_j(1 - C_i)Y_{01} \qquad (3)$$

We are interested in the identification and estimation of a) the average direct effect of monitoring $E[Y_{11}-Y_{01}]$; b) the average indirect effect of monitoring $E[Y_{01}-Y_{00}]$; c) the average overall effect of monitoring $E[Y_{11}-Y_{00}]$, where $E[.]$ is the mean operator.

The sampling procedure described in Section 3 is characterized by conditional randomization, which implies that a) in each region, the assignment to school - level treatment is random, conditional on the size of the school, measured by the number of students enrolled in the second and fifth grade; b) the assignment to class - level treatment for

[12] We notice that the proportion of missing data is slightly smaller in sampled schools and classes. This might be due to a more careful reporting of administrative information by secretaries in the schools and classes assigned to external monitoring.

a class of a given grade in a treated school is random conditional on the size of the school, measured both by the number of students enrolled in the second and fifth grade and by the number of classes in the selected grade. Conditional randomization in each grade implies that

$$Y_{00},\ Y_{01},\ Y_{11} \perp S_j,\ C_i \mid R \tag{4}$$

When (4) holds, the average direct, indirect and overall effects of external monitoring are given by

$$E[Y_{11} - Y_{01} \mid R] = E[Y_{ij} \mid C_i = 1, S_j = 1, R] - E[Y_{ij} \mid C_i = 0, S_j = 1, R] \tag{5}$$

$$E[Y_{01} - Y_{00} \mid R] = E[Y_{ij} \mid C_i = 0, S_j = 1, R] - E[Y_{ij} \mid C_i = 0, S_j = 0, R] \tag{6}$$

$$E[Y_{11} - Y_{00} \mid R] = E[Y_{ij} \mid C_i = 1, S_j = 1, R] - E[Y_{ij} \mid C_i = 0, S_j = 0, R] \tag{7}$$

We aggregate our data at the class level and evaluate the effects of external monitoring on average class performance in the Math test by estimating

$$Y_{ij} = \theta_0 + \theta_1 C_{ij} S_j + \theta_2 S_j + \theta_3 R_{rj} + \theta_4 \Omega_{ij} + u_{ij} \tag{8}$$

where the dependent variable is the average percentage of correct answers in the class. We allow errors *u* to be correlated among the classes of the same school and weigh each class-level observation with the number of students in the class. The vector $\Omega$ includes for all grades the number of students enrolled in a class, which is greater among treated classes, and the following covariates: type of school (public or private), full or part-time schedule, average (in the class) gender, place of birth, citizenship, attendance of pre-primary school, age, grades in Italian and Math in the previous semester, employment, education and nationality of parents, and only for the fifth grade the percentage (in the class) of students who have their own bedroom, internet access, an encyclopaedia, own desk, a computer and a place for doing homework, the average number of books in the house, the average number of siblings, the percentage of students living with both parents, the language spoken at home, and whether they receive help with her homework. The summary statistics of these covariates are in Panel A of Table 2. The direct, indirect and overall effect of external monitoring are given by $\theta_1$, $\theta_2$ and $\theta_1 + \theta_2$ respectively.

## 5. Results

Table 4 shows our baseline estimates of Eq. (8). Standard errors in this and the next tables are clustered at the school level. The first column in the table considers all Italian regions, and the remaining columns show the estimates by macro area (North, Centre and South). We find that having an external examiner in the class reduces the percentage of correct answers by 3.59 percentage points, which corresponds to a 5.5 percent decline with respect to the mean score in untreated schools[13]. Close to 80 percent (2.79/3.59) of this total effect is direct, and the remaining 20 percent (0.81/3.59) is indirect. The size of the total, direct and indirect effects varies with the macro area and is highest in Southern regions, where the total effect is -8.9%, and lowest in Northern Italy, where it is -2.6%.

Why are test results worse in classes with the external examiner? One possibility is that young students under-perform because they are distracted by the presence of a stranger in the class and are more anxious that students in un-monitored classes. The other possibility is that either students or teachers in classes without the external examiner engage in outright cheating[14]. We believe that the second one is the explanation, for the following reasons.

First, there is no evidence that students in classes with the external examiner are negatively affected in their feelings and motivation to complete the test properly. In a questionnaire filled up by fifth graders participating to the test in classes with and without the external examiner, INVALSI asked a set of motivational questions aimed at capturing the psychological status of students during the test, which included agreement or disagreement with the following sentences: a) I was already anxious before starting the test; b) I was so nervous I couldn't find the answers; c) while answering, I felt like I was doing badly; d) while answering, I was calm. Table 5 presents the results of estimating Eq. (8) when the dependent variable is the percentage of students in the class agreeing with each of the four statements above. We find no evidence that being in a class with an external examiner increased anxiety

[13] As shown in Table A.2 in the Appendix, the total effect is somewhat larger for second graders (5.4 percentage points, or 8.5% of the average score in untreated schools).

[14] We assume that cheating is unlikely in classes with the external examiner. On the one hand, since schools are informed of having been selected to receive an external examiner only about one week prior to the date of the test, there is little room of *manoeuvre* for teachers to react and adopt strategies that manipulate student performances in the presence of the examiner. On the other hand, we assume that external examiners have no incentive to cheat and collude with school teachers and principals in order to boost school results. In support of this assumption, INVALSI (2010a) used a procedure to detect cheating in monitored classes and concluded that there was no evidence of cheating. The cheating detection algorithm is described in INVALSI (2010b).

or nervousness. Quite the opposite, there is evidence that students in these classes were less nervous and calmer during the test.

Second, we examine the distribution of results within classes. In the absence of external controls, the teacher can communicate the correct answers to students or change their answers in the answer sheet, or students can simply copy from each other. If outright cheating by students and/or teachers was taking place in the classes without the external examiner, we should find that in these classes – *ceteris paribus* - the standard deviation and the coefficient of variation of test results are lower than in classes with the external examiner, where cheating is minimized or altogether absent. While distraction and anxiety can reduce average performance, it is not obvious that they reduce its variability. Table 6 shows for the entire country the effects of the presence of an external examiner on the within – class standard deviation and coefficient of variation of the percentage of correct answers, as well as on the bottom quartile, median and top quartile of the distribution of test scores within classes.

We find that in classes with the external examiner the standard deviation and the coefficient of variation of results are about 6% and 11% higher than in un-monitored classes. There is also evidence that the presence of the external examiner affects to a higher extent the performance of students in the lower quartile of the distribution of outcomes, in line with the expectation that cheating typically helps low performers, or that low performing students are those more prone to copy. When compared with students in untreated schools, having an external examiner reduces the score of these students by about 8% (-4.26/55.6). This effect is strongest for second grade students in Southern Italy, where if reaches a striking 18.7%[15].

Third, we compute an index of heterogeneity in the pattern of answers given by students in each class. For each question, we use a modified version of the Herfindahl Index

$$H = \frac{1 - \sum_{a=1}^{A} s_a^2}{1 - \frac{1}{A}}. \qquad (12)$$

where $s_a$ is the within-class share of students who chose answer "a" in the set A of possible answers[16]. Index $H$ ranges between 0 and 1, with higher values signalling a more heterogeneous pattern of answers to a given question. We obtain an overall measure of the

[15] Detailed results by macro area are available from the authors upon request.
[16] We treat missing values as a separate category.

heterogeneity of answers in the class by averaging $H$ across all questions in the test. While we expect this measure to decline in classes without the external examiner in the presence of cheating, it is not clear whether it declines or increases if anxiety or distraction play a role. Table 7 reports the estimates of Eq. (8) when the dependent variable is $H$, and shows that heterogeneity is significantly higher in classes with the external examiner. We also find that, as in the case of the percentage of correct answers in the class, the effects of external monitoring on the heterogeneity of answers increase significantly moving from Northern to Southern Italy (columns (2) to (4)).

Finally, the correlation between test score outcomes and teacher grades in the semester before the tests should be lower in the presence of cheating. Using individual student data, we examine the correlation between the rank in the test and the rank in teacher grades in classes with and without the external examiners. In line with our expectations, we find a higher correlation for students taking the test in classes with the external examiner[17].

While these results are suggestive of the presence of cheating, we cannot say whether cheating occurs because teachers change answers in their report to INVALSI, or because they suggest the correct answers to students in the class, or because students are given extra time or are allowed to copy from each other in classes without the external examiner. Since all these cheating strategies generate a higher proportion of correct answers and a lower within - class dispersion of results, they are observationally equivalent in our data. To distinguish between some of these strategies, we would need to observe both the answers directly chosen by students and the answers reported by teachers to INVALSI. Unfortunately, we only observe the latter. We can only speculate that since in un-monitored classes teachers are responsible for supervision in class, collection of the tests, filling-in of the answer sheets on the basis of the responses given by the students and transmission of the answer sheets to INVALSI, they have certainly plenty of opportunities to modify test results.

An interesting and novel result of our analysis is that external examiners affect performance not only in the class they supervise but also in other classes of the same school. This indirect effect of monitoring in school tests has not been detected before and deserves further explanation. One interpretation is that teachers administering the test in the same school where the external examiner is present are afraid to be monitored by this supervisor and therefore restrain their cheating activities. This interpretation relies on irrational

[17] We regress the individual within-class rank in the test on the individual within-class rank in teacher grades and its interaction with the presence of an external examiner and find that the interaction attracts a positive and statistically significant coefficient, especially in the South, where cheating appears to be more widespread. Detailed results are available from the authors upon request.

behaviour, because teachers were informed before the test that the external examiner's mandate was restricted to the randomly selected class.

An alternative explanation is that teachers dislike excessive dispersion in average test scores within the same school, because such dispersion could generate conflicts with other teachers. To illustrate, consider a school where a single class is supervised by an external examiner. If teachers administering the test in the other classes cheat freely, these classes will look much better than the supervised class, where cheating is restrained. This may generate conflicts with the teacher in charge of the supervised class. To reduce these conflicts, teachers in un-monitored classes may be induced to restrain their cheating.

## 6. Robustness checks

In this section we investigate whether our main results are robust to several sensitivity checks. First, since the dependent variable of our main estimates is a fraction (the percentage of correct answers in the class) we implement the GLM estimator proposed by Papke and Wooldridge (1996) to deal with fractional dependent variables. Estimated marginal effects, shown in Table A.4 in the Appendix, are in line with the baseline estimates in Table 4.

Second, we exploit the census nature of our data and the fact that we observe almost the entire population of students in each grade to apply a finite population correction to statistical inference. Results (Table A.5 in the Appendix) are qualitatively unchanged with respect to the baseline, but precision increases significantly.

Third, we drop all observable covariates not required for the implementation of conditional randomization[18]. Since assignment to treatment does not depend on observables, finding differences between the estimates that include and exclude covariates is a symptom of strategic manipulation of the composition of the pool of tested students. Results in Table A.6 in the Appendix do not provide any strong evidence in this direction. Finally, we test for differences in absenteeism across treatment statuses, using as dependent variable the percentage of students absent from the test in each class. Again, differences in behaviour across the three groups are minimal (see Table A.7 in the Appendix).

[18] We still include regional dummies, regional dummies interacted with school size and with the number of fifth grade classes in the school, and the number of students enrolled in the class.

## 7. Extensions

So far, we have allowed treatment effects to vary across the different macro areas of the country. Yet there might be other relevant sources of heterogeneity to be considered. In this section we do two things. We start by exploring what these other sources could be – without pretending to be exhaustive - and then examine whether regional heterogeneity is related to regional differences in social capital.

Our candidate sources of heterogeneous treatment effects are a) class size; b) the percentage of tenured teachers in the school; c) an indicator of average parental background for the students in the class[19]. On the one hand, if student cheating is easier in larger classes, we should find that the overall effect of having an external examiner increases with class size. On the other hand, larger classes could increase the cost of cheating by teachers or could reduce the effectiveness of external supervision. In this case, the overall effect should be smaller in larger classes. Column (1) in Table 8 presents our estimates when both the direct and the indirect effect are interacted with class size[20]. The evidence suggests that the overall effect of external supervision is smaller in larger classes, in line with the second hypothesis.

Column (2) in the table shows that both the direct and the overall effect of external monitoring are higher in schools where the percentage of tenured teachers is higher. Typically, these are senior teachers with very secure jobs, who are less willing to adjust their teaching style to the needs of standardized tests and may therefore be more likely to engage in cheating and sabotaging.

Column (3) examines the interactions of the overall, direct and indirect effects with *ESCS*, the indicator of the average parental background in the class. If the incentives to engage in cheating were higher in classes with poor parental background, perhaps because teachers wish to altruistically compensate their students for their unfavourable initial conditions, we should find that the negative effect of external supervision is higher in these classes. Yet, there is no statistical evidence that this is the case[21].

Next, we ask whether the regional differences in the size of the effects of external monitoring are associated to the differences in the level of social capital[22]. Guiso, Sapienza and

[19] Descriptive statistics for these variables are shown in Table 2 – Panel B.

[20] In this and in the following regressions the interacted variable is included also as an independent control.

[21] One possible explanation is that not only teachers, but also external examiners may be induced to engage in compensatory behaviour.

[22] In their seminal work, Putnam et al. (1993) links differences in the performance of local Italian governments to regional heterogeneity in social capital, measured in terms of local patterns of associationism, newspaper readership and political participation. Guiso, Sapienza and Zingales (2004) show that social capital is a key

Zingales (2010) define social capital as civic capital, or as "...those persistent and shared beliefs and values that help a group overcome free rider outcomes..."(p.8). They report higher levels of social capital in Northern and Central Italy compared to the South.

We interact both the direct and the indirect effect of external monitoring with two measures of social capital at the provincial level taken from Guiso, Sapienza and Zingales (2004), the number of blood donations per 10,000 inhabitants in 1995 and the average electoral participation in the referenda held in Italy between 1946 and 1987. Since social capital is strongly correlated with local economic conditions, as shown in Figures 1.a-1.d, we also interact both effects with provincial GDP per capita and unemployment rates in 2009.

Results are shown in Table 9[23]. Column (1) in the table reports the estimates of the baseline model in the sub-sample of provinces for which data on social capital are available. These estimates are in line with those presented in Table 4. Column (2) and (4) show the interactions of the direct, indirect and overall effect of external monitoring with the two selected measures of social capital (blood donations and turnout at referenda, measured as deviations from sample means). We find that both the direct and the overall effect are smaller in schools located in provinces with a higher social capital. These qualitative results remain when we add to the regressions the interactions with provincial unemployment and GDP per capita (also measured as deviations from sample means, see columns (3) and (5)), although the effect of social capital is smaller.

Starting with Putnam's seminal contribution, several studies have suggested that Southern Italy has a lower endowment of "bridging" social capital, the form of social capital supportive of a more cohesive society and higher civicness[24], and is richer at the same time of "bonding" social capital, the type of social capital which reinforces family and clan ties in competition with the market and overall society and which is at the roots of the so called *amoral familism* (in the words of Banfield, 1958)[25]. We interpret the higher level of cheating observed in Southern Italy as the outcome of lower marginal costs of cheating due to lower "bridging" social capital, and/or of higher marginal benefits due to higher "bonding" social capital.

determinant of financial development, and Nannicini et al. (2012) study the impact of social capital on political accountability. Finally, Ichino and Maggi (2000) measure civicness in terms of shirking behaviour in the workplace and document large shirking differentials between Northern and Southern Italy.

[23] Descriptive statistics for these variables are shown in Table 2 – Panel B.

[24] Blood donations and referenda turnout measure bridging social capital.

[25] See Alesina and Ichino, 2009, for recent evidence.

## Conclusions

Test-based accountability systems in education may be gamed by students, teachers and school administrators in order to obtain higher measured levels of performance. This paper shows that having an external examiner who monitors test procedures has negative effects on the measured performance of tested classes and schools.These results are based on a natural experiment designed by the Italian national test administrator (INVALSI) to monitor test procedures in a random sample of Italian primary school classes. We have used random assignment to treatment to estimate both the direct and indirect effects of external monitoring. The former is based on the comparison of monitored and un-monitored classes within the same school and the latter on the comparison of un-monitored classes in schools with and without the external examiner.

The overall effect (direct plus indirect) of external monitoring is statistically significant and sizeable: depending on the grade, the presence of an external examiner reduces the percentage of correct answers in the class by 5.5 to 8.5 percent with respect to classes in schools with no external monitor. External monitoring spills over to un-monitored classes of the same school, but the size of this beneficial effect is rather small (about 20 percent of the overall effect).

Using additional supporting evidence on the psychological conditions of students before and during the test and on the distribution of answers within classes, we have concluded that the better performance of classes without the external examiner is due to the manipulation of test outcomes by teachers and/or students, and that the performance gap between monitored and un-monitored classes can be interpreted as a measure of the average intensity of cheating taking place in the latter.

While the direct negative effect of external supervision on test performance is not surprising, the presence of a small but statistical significant indirect negative effect is less expected. We have argued that this effect can be explained either by (irrational) fear of supervision or by a model where rational teachers administering the tests dislike excessive dispersion of test results within the school.

We believe that our results are useful for an economic assessment of external monitoring, which requires the evaluation of costs and benefits. To measure benefits, we need to ascertain whether external monitoring reduces cheating and by how much. Needless to say, using external examiners is not the only deterrence tool. Alternatives include re-shuffling the questions assigned to each students and computer – based tests. Reshuffling questions deters students from copying but does not strongly prevent cheating by teachers. Computer-based testing virtually eliminates cheating by teachers but it is quite costly, as it requires that each student is endowed with a computer. At the cost of 200 euro per workday, external examiners are rather cost-effective at reducing the manipulation of tests in a random sample of Italian schools. Yet, extending their use to the universe of tested schools seems complicated, not only because of the monetary costs involved but also because of the difficulty of finding enough qualified examiners.

TABLES AND FIGURES

Table 1. Total and Sampled Number of Schools, Classes and Students. INVALSI SNV Test 2010

| | Number of schools (total) | Number of classes (total) | Number of students (total) | Number of sampled schools | Number of sampled classes | Number of sampled students |
|---|---|---|---|---|---|---|
| Second Grade | 7,700 | 30,175 | 555,347 | 1,385 | 2,000 | 39,299 |
| Fifth Grade | 7,700 | 30,476 | 565,064 | 1,385 | 2,000 | 39,643 |

Table 2. Mean and Standard Deviation of Covariates - Math Tests - V Graders

Panel A

| | Mean | St Dev |
|---|---|---|
| Gender | | |
| Missing (%) | 0.01 | 0.10 |
| Male (%) | 0.50 | 0.50 |
| Place of birth | | |
| Missing (%) | 0.04 | 0.20 |
| Italy (%) | 0.89 | 0.31 |
| Citizenship | | |
| Missing (%) | 0.02 | 0.15 |
| Italian (%) | 0.89 | 0.32 |
| First generation foreigner (%) | 0.05 | 0.22 |
| Second generation foreigner (%) | 0.04 | 0.20 |
| Pre-primary school | | |
| Missing (%) | 0.15 | 0.35 |
| Yes (%) | 0.83 | 0.37 |
| Age | | |
| Missing (%) | 0.01 | 0.10 |
| Older than regular (%) | 0.03 | 0.16 |
| Regular (%) | 0.87 | 0.33 |
| Younger than regular (%) | 0.09 | 0.29 |
| Math grade in previous semester (range:1-10) | | |
| Missing (%) | 0.07 | 0.26 |
| 1-4 (%) | 0.00 | 0.04 |
| 5 (%) | 0.04 | 0.20 |
| 6-7 (%) | 0.38 | 0.48 |
| 8-10 (%) | 0.51 | 0.50 |
| Italian grade in previous semester (range:1-10) | | |
| Missing (%) | 0.07 | 0.25 |
| 1-4 (%) | 0.00 | 0.04 |
| 5 (%) | 0.04 | 0.19 |
| 6-7 (%) | 0.41 | 0.49 |
| 8-10 (%) | 0.48 | 0.50 |
| Has own bedroom | | |
| Missing (%) | 0.03 | 0.17 |
| Yes (%) | 0.55 | 0.50 |
| Has internet access | | |
| Missing (%) | 0.03 | 0.16 |
| Yes (%) | 0.76 | 0.43 |
| Has an encyclopedia | | |
| Missing (%) | 0.03 | 0.16 |
| Missing (%) | 0.71 | 0.46 |
| Has own desk | | |
| Missing (%) | 0.02 | 0.15 |
| Yes (%) | 0.85 | 0.36 |
| Has a PC | | |
| Missing (%) | 0.03 | 0.16 |
| Yes (%) | 0.75 | 0.43 |
| Has a place for homework | | |
| Missing (%) | 0.03 | 0.16 |
| Yes (%) | 0.84 | 0.37 |
| Number of books at home | | |
| Missing (%) | 0.04 | 0.20 |
| 0-10 (%) | 0.12 | 0.33 |
| 11-25 (%) | 0.25 | 0.43 |
| 26-100 (%) | 0.31 | 0.46 |
| 101-200 (%) | 0.15 | 0.36 |
| >200 (%) | 0.12 | 0.33 |

| | Mean | St Dev |
|---|---|---|
| Mother occupation | | |
| Missing (%) | 0.20 | 0.40 |
| Unemployed or retired (%) | 0.35 | 0.48 |
| Employee (%) | 0.31 | 0.46 |
| Entrepreneur (%) | 0.08 | 0.28 |
| Middle manager (%) | 0.06 | 0.23 |
| Father occupation | | |
| Missing (%) | 0.22 | 0.41 |
| Unemployed or retired (%) | 0.04 | 0.19 |
| Employee (%) | 0.39 | 0.49 |
| Entrepreneur (%) | 0.25 | 0.43 |
| Middle manager (%) | 0.11 | 0.31 |
| Mother education | | |
| Missing (%) | 0.21 | 0.41 |
| Primary (%) | 0.39 | 0.49 |
| Secondary (%) | 0.29 | 0.45 |
| Tertiary (%) | 0.11 | 0.32 |
| Father education | | |
| Missing (%) | 0.22 | 0.42 |
| Primary (%) | 0.43 | 0.49 |
| Secondary (%) | 0.25 | 0.43 |
| Tertiary (%) | 0.10 | 0.30 |
| Mother nationality | | |
| Missing (%) | 0.09 | 0.28 |
| Italian (%) | 0.80 | 0.40 |
| Father nationality | | |
| Missing(%) | 0.09 | 0.29 |
| Italian (%) | 0.82 | 0.39 |
| Private school | 0.05 | 0.23 |
| Full time schedule class | 0.23 | 0.42 |
| Number of students enrolled in class | 19.00 | 4.65 |
| Number of siblings | | |
| Missing (%) | 0.02 | 0.15 |
| 0 (%) | 0.15 | 0.36 |
| 1 (%) | 0.55 | 0.50 |
| 2 (%) | 0.20 | 0.40 |
| 3 (%) | 0.05 | 0.21 |
| 4 or more (%) | 0.03 | 0.17 |
| Lives with | | |
| Missing (%) | 0.02 | 0.15 |
| Both parents (%) | 0.86 | 0.35 |
| One parent only (%) | 0.06 | 0.24 |
| Both parents alternatively(%) | 0.05 | 0.22 |
| Others (%) | 0.01 | 0.08 |
| Language spoken at home | | |
| Missing (%) | 0.04 | 0.21 |
| Italian (%) | 0.73 | 0.44 |
| Dialect (%) | 0.15 | 0.36 |
| Other (%) | 0.07 | 0.25 |
| Help with homework | | |
| Missing (%) | 0.07 | 0.26 |
| No homework (%) | 0.01 | 0.07 |
| No help needed (%) | 0.20 | 0.40 |
| Parents (%) | 0.45 | 0.50 |
| Siblings (%) | 0.12 | 0.32 |
| Private teacher (%) | 0.03 | 0.17 |
| Other (%) | 0.04 | 0.20 |
| No one (%) | 0.09 | 0.28 |

Panel B (continued)

| | Mean | St. Dev. |
|---|---|---|
| Tenured teachers in the school (%) | 90.33 | 9.13 |
| Class average ESCS index | -0.045 | 0.51 |
| Class size | 16.93 | 4.64 |

| | Mean | St. Dev. |
|---|---|---|
| Blood donations | 2.81 | 2.17 |
| Average turnout at referenda (%) | 80.28 | 8.37 |
| Provincial unemployment rate (2009) | 7.95 | 3.69 |
| Provincial per capita GDP (2009) | 23.84 | 5.60 |

Panel C

| | Mean | St. Dev. |
|---|---|---|
| Math Test – V grade | | |
| Score | 0.65 | 0.19 |
| Within-class standard deviation | 0.14 | 0.04 |
| Within-class coefficient of variation | 0.23 | 0.09 |
| Within-class bottom quartile | 0.55 | 0.14 |
| Within-class median | 0.65 | 0.13 |
| Within-class top quartile | 0.75 | 0.12 |
| Within-class Herfindal Index | 0.53 | 0.15 |
| Ranking based on Math scores | 9.82 | 5.84 |
| Ranking based on Math grades given by teachers in the previous semester | 10.44 | 6.96 |

| | Mean | St. Dev. |
|---|---|---|
| Anxiety Questions | | |
| I was already anxious before starting the test | 0.61 | 0.49 |
| I was so nervous I couldn't find the answers | 0.19 | 0.39 |
| While answering , I felt like I was doing badly | 0.50 | 0.50 |
| While answering, I was calm | 0.53 | 0.50 |
| Absences from test (%) | 0.11 | 0.10 |
| Maths Test – II grade | | |
| Score | 0.62 | 0.20 |
| Italian Test – V grade | | |
| Score | 0.70 | 0.18 |
| Italian Test – II grade | | |
| Score | 0.65 | 0.23 |

Notes: The table reports the mean and standard deviation of the covariates included in the regressions (Panel A), the variables used in Section 7 (Panel B) and the dependent variables (Panel C). All numbers refer to the entire country. These statistics are based on individual, school and class level data. Except for the number of students enrolled in each class, the variables in Panel A have been categorized as dummy variables. Class size in Panel B refers to the number of students attending the test. Blood donations are the number of blood bags per 10,000 inhabitants in the province. Per capita GDP is measured in thousand euro. See the Appendix for further details.

Table 3 - Balancing Tests. First (between schools) and Second Stage (within schools) Randomization. - Math tests - V Graders.

Panel A

| | Between schools | Within schools |
|---|---|---|
| Private school (%) | 0.003 | . |
| Full time schedule (%) | 0.015 | 0.011 |
| Number of students enrolled in class | 0.079 | 0.425*** |
| Gender | | |
| Missing (%) | 0.007*** | 0.020*** |
| Male (%) | -0.005** | -0.004 |
| Place of birth | | |
| Missing (%) | -0.014*** | -0.027*** |
| Italy (%) | 0.014*** | 0.027*** |
| Citizenship | | |
| Missing (%) | -0.008*** | -0.013*** |
| Italian (%) | 0.008** | 0.010*** |
| First generation foreigner (%) | -0.001 | 0.000 |
| Second generation foreigner (%) | 0.001 | 0.002 |
| Pre-primary school | | |
| Missing (%) | -0.027*** | -0.009* |
| Yes (%) | 0.027*** | 0.010* |
| Age | | |
| Missing (%) | 0.007*** | 0.018*** |
| Older than regular (%) | 0.000 | 0.000 |
| Regular (%) | -0.008*** | -0.014*** |
| Younger than regular (%) | 0.002 | -0.004** |
| Math grade in semester before the test | | |
| Missing (%) | -0.021*** | -0.009* |
| 1-4 (%) | 0.000 | 0.000* |
| 5 (%) | 0.001 | 0.000 |
| 6-7 (%) | 0.010** | 0.008* |
| 8-10 (%) | 0.011* | 0.001 |
| Italian grade in semester before the test | | |
| Missing (%) | -0.021*** | -0.008 |
| 1-4 (%) | 0.000 | 0.000 |
| 5 (%) | 0.000 | 0.001 |
| 6-7 (%) | 0.006 | 0.003 |
| 8-10 (%) | 0.014*** | 0.004 |

| | Between schools | Within schools |
|---|---|---|
| Mother occupation | | |
| Missing (%) | -0.014 | -0.024*** |
| Unemployed or retired (%) | 0.008 | 0.012*** |
| Employee (%) | 0.003 | 0.004 |
| Entrepreneur (%) | 0.001 | 0.006** |
| Middle manager (%) | 0.003 | 0.002 |
| Father occupation | | |
| Missing (%) | -0.014 | -0.023*** |
| Unemployed or retired (%) | 0.001 | 0.001 |
| Employee (%) | 0.002 | 0.016*** |
| Entrepreneur (%) | 0.009* | 0.005 |
| Middle manager (%) | 0.002 | 0.002 |
| Mother education | | |
| Missing (%) | -0.017 | -0.028*** |
| Primary (%) | 0.008 | 0.019*** |
| Secondary (%) | 0.005 | 0.009** |
| Tertiary (%) | 0.004 | 0.000 |
| Father education | | |
| Missing (%) | -0.018* | -0.025*** |
| Primary (%) | 0.013* | 0.016*** |
| Secondary (%) | 0.001 | 0.008** |
| Tertiary (%) | 0.003 | 0.001 |
| Mother nationality | | |
| Missing (%) | -0.018*** | -0.014*** |
| Italian (%) | 0.015*** | 0.012** |
| Father nationality | | |
| Missing (%) | -0.017*** | -0.013*** |
| Italian (%) | 0.015*** | 0.009* |

Panel B (continued)

| | Between schools | Within schools |
|---|---|---|
| Has own bedroom | | |
| Missing (%) | -0.006** | -0.009*** |
| Yes (%) | 0.000 | 0.004 |
| Has internet access | | |
| Missing (%) | -0.006** | -0.008*** |
| Yes (%) | 0.007** | 0.008** |
| Has an encyclopedia | | |
| Missing (%) | -0.006** | -0.008*** |
| Yes (%) | 0.005 | 0.016*** |
| Has own desk | | |
| Missing (%) | -0.005** | -0.008*** |
| Yes (%) | 0.005* | 0.009*** |
| Has a PC | | |
| Missing (%) | -0.005** | -0.008*** |
| Yes (%) | 0.007** | 0.011*** |
| Has a place for homework | | |
| Missing (%) | -0.006** | -0.008*** |
| Yes (%) | 0.006** | 0.008** |
| Number of books at home | | |
| Missing (%) | -0.007*** | -0.008*** |
| 0-10 (%) | 0.000 | 0.001 |
| 11-25 (%) | -0.004 | -0.001 |
| 26-100 (%) | 0.001 | 0.006* |
| 101-200 (%) | 0.004** | 0.003 |
| >200 (%) | 0.006*** | -0.001 |

| | Between schools | Within schools |
|---|---|---|
| Number of siblings | | |
| Missing (%) | -0.007*** | -0.009*** |
| 0 (%) | -0.001 | 0.000 |
| 1 (%) | 0.005* | 0.008** |
| 2 (%) | 0.001 | 0.000 |
| 3 (%) | 0.001 | 0.000 |
| 4 or more (%) | 0.001 | 0.001 |
| Lives with | | |
| Missing (%) | -0.008*** | -0.010*** |
| Both parents (%) | 0.008*** | 0.007** |
| One parent only (%) | -0.001 | 0.000 |
| Both parents alternatively (%) | 0.000 | 0.002 |
| Others (%) | 0.000 | 0.000 |
| Language spoken at home | | |
| Missing (%) | -0.008*** | -0.009*** |
| Italian (%) | 0.004 | 0.007* |
| Dialect (%) | 0.003 | 0.001 |
| Other (%) | 0.001 | 0.001 |
| Help with homework | | |
| Missing | -0.008*** | -0.006** |
| No homework (%) | -0.001** | -0.001*** |
| No help needed (%) | -0.001 | 0.005 |
| Parents (%) | 0.006* | 0.001 |
| Siblings (%) | 0.003** | -0.002 |
| Private teacher (%) | 0.000 | 0.002 |
| Other (%) | 0.002 | -0.001 |
| No one (%) | -0.001 | 0.002 |

Notes: the table shows the point estimates of the balancing tests between and within schools. We compute school or class averages of individual variables and test for balancing using regressions (1) and (2). Full time schedule refers to schools offering this option in the between schools analysis and to the schedule of the single class in the within school analysis. While variables in Panel A are available for students in both grades, variables in Panel B are only available for fifth grade students. Standard errors for the second stage are adjusted for clustering at the school level. One, two and three stars for statistical significance at the 10, 5 and 1 percent level.

Table 4. The Effects of External Monitoring. Math Tests – V Grade. Dependent variable: Percentage of Correct Answers in the Class.

| | (1)<br>Italy | (2)<br>North | (3)<br>Centre | (4)<br>South |
|---|---|---|---|---|
| Direct Effect | -2.79*** | -0.99*** | -2.27*** | -4.92*** |
| | (0.25) | (0.28) | (0.48) | (0.50) |
| Indirect Effect | -0.81*** | -0.70*** | -0.73 | -1.04* |
| | (0.28) | (0.27) | (0.45) | (0.61) |
| Overall Effect | -3.59*** | -1.69*** | -2.99*** | -5.96*** |
| | (0.29) | (0.31) | (0.54) | (0.60) |
| Observations | 27,325 | 11,541 | 4,886 | 10,898 |
| R-squared | 0.15 | 0.2 | 0.15 | 0.14 |
| Additional covariates | Yes | Yes | Yes | Yes |
| Mean - Untreated Schools | 65.1 | 63.9 | 64.0 | 66.8 |

Notes: all regressions include the number of students enrolled in the class, regional dummies and regional dummies interacted with school size and with the number of fifth grade classes in the school. Additional covariates are shown in Table 2 - panel A. Estimates are weighted by class size. Standard errors adjusted for clustering at the school level in parentheses. One, two and three stars for statistical significance at the 10, 5 and 1 percent level of confidence.

Table 5. The Effects of External Monitoring on Student Psychological Conditions. Math Tests – V Grade. Dependent variable: Percentage of Positive Answers in the Class.

| | (1)<br>I was already anxious before starting the test | (2)<br>I was so nervous I couldn't find the answers | (3)<br>While answering , I felt like I was doing badly | (4)<br>While answering, I was calm |
|---|---|---|---|---|
| Direct Effect | 0.25 | -0.92*** | -0.08 | 0.64 |
| | (0.42) | (0.29) | (0.39) | (0.39) |
| Indirect Effect | 0.25 | 0.01 | 0.36 | -0.01 |
| | (0.31) | (0.21) | (0.28) | (0.29) |
| Overall Effect | 0.50 | -0.90*** | 0.28 | 0.63* |
| | (0.41) | (0.28) | (0.38) | (0.38) |
| Observations | 27,141 | 27,142 | 27,141 | 27,140 |
| R-squared | 0.07 | 0.11 | 0.1 | 0.07 |
| Additional covariates | Yes | Yes | Yes | Yes |
| Mean - Untreated Schools | 61.0 | 19.2 | 50.7 | 53.1 |

Notes: see Table 4. In each column, the dependent variable is the percentage of students in the class who agreed with the sentence reported at the top of the column. Students with missing answers have been dropped from the estimation sample (about 2 percent of the total). The estimates refer to the entire country.

Table 6. The Effects of External Monitoring on the Standard Deviation, the Coefficient of Variation and the Quartiles of the Distribution of Correct Answers within the Class. Math tests – V Grade.

| | (1)<br>Standard Deviation | (2)<br>Coefficient of Variation | (3)<br>Bottom quartile | (4)<br>Median | (5)<br>Top quartile |
|---|---|---|---|---|---|
| Direct Effect | 0.76*** | 2.14*** | -3.70*** | -3.07*** | -2.26*** |
| | (0.09) | (0.21) | (0.31) | (0.29) | (0.27) |
| Indirect Effect | 0.03 | 0.30 | -0.55* | -0.56* | -0.61** |
| | (0.08) | (0.18) | (0.31) | (0.29) | (0.26) |
| Overall Effect | 0.79*** | 2.44*** | -4.26*** | -3.63*** | -2.88*** |
| | (0.09) | (0.22) | (0.33) | (0.32) | (0.30) |
| Observations | 27,325 | 27,325 | 27,325 | 27,325 | 27,325 |
| R-squared | 0.18 | 0.15 | 0.12 | 0.1 | 0.09 |
| Additional covariates | Yes | Yes | Yes | Yes | Yes |
| Mean - Untreated Schools | 14.1 | 22.8 | 55.6 | 65.6 | 75.2 |

Notes: see Table 4. The estimates refer to the entire country.

Table 7. The Effects of External Monitoring on the Heterogeneity of Answers in each Class. Math Tests – V Grade. Dependent Variable: Average Herfindhal Index in Each Class x 100.

| | (1)<br>Italy | (2)<br>North | (3)<br>Centre | (4)<br>South |
|---|---|---|---|---|
| Direct Effect | 3.93*** | 1.24*** | 2.63*** | 7.32*** |
| | (0.32) | (0.32) | (0.60) | (0.64) |
| Indirect Effect | 0.82** | 0.64** | 0.51 | 1.22* |
| | (0.34) | (0.31) | (0.58) | (0.73) |
| Overall Effect | 4.75*** | 1.88*** | 3.14*** | 8.54*** |
| | (0.35) | (0.35) | (0.62) | (0.719) |
| Observations | 27,325 | 11,541 | 4,886 | 10,898 |
| R-squared | 0.2 | 0.17 | 0.13 | 0.15 |
| Additional covariates | Yes | Yes | Yes | Yes |
| Mean - Untreated Schools | 52.8 | 57.3 | 55.7 | 46.9 |

Notes: see Table 4.

Table 8. Heterogeneous Effects of External Monitoring. Math Tests – V Grade. Dependent variable: Percentage of Correct Answers in the Class.

| | (1) | (2) | (3) |
|---|---|---|---|
| | Interacted with Class Size | Interacted with % Tenured Teachers | Interacted with ESCS |
| Direct Effect | -3.41*** | -1.34*** | -2.65*** |
| | (0.41) | (0.29) | (0.33) |
| Interacted Direct Effect | 0.98* | -2.98*** | -0.15 |
| | (0.53) | (0.50) | (0.54) |
| Indirect Effect | -0.94*** | -0.66** | -0.67** |
| | (0.36) | (0.29) | (0.31) |
| Interacted Indirect Effect | 0.22 | -0.19 | -0.30 |
| | (0.41) | (0.54) | (0.44) |
| Overall Effect | -4.35*** | -2.00*** | -3.32*** |
| | (0.43) | (0.33) | (0.36) |
| Interacted Overall Effect | 1.20** | -3.17*** | -0.45 |
| | (0.51) | (0.57) | (0.51) |
| Observations | 27,325 | 26,313 | 27,323 |
| R-squared | 0.15 | 0.15 | 0.15 |
| Additional covariates | Yes | Yes | Yes |
| Mean - Untreated Schools | 65.1 | 64.9 | 65.1 |

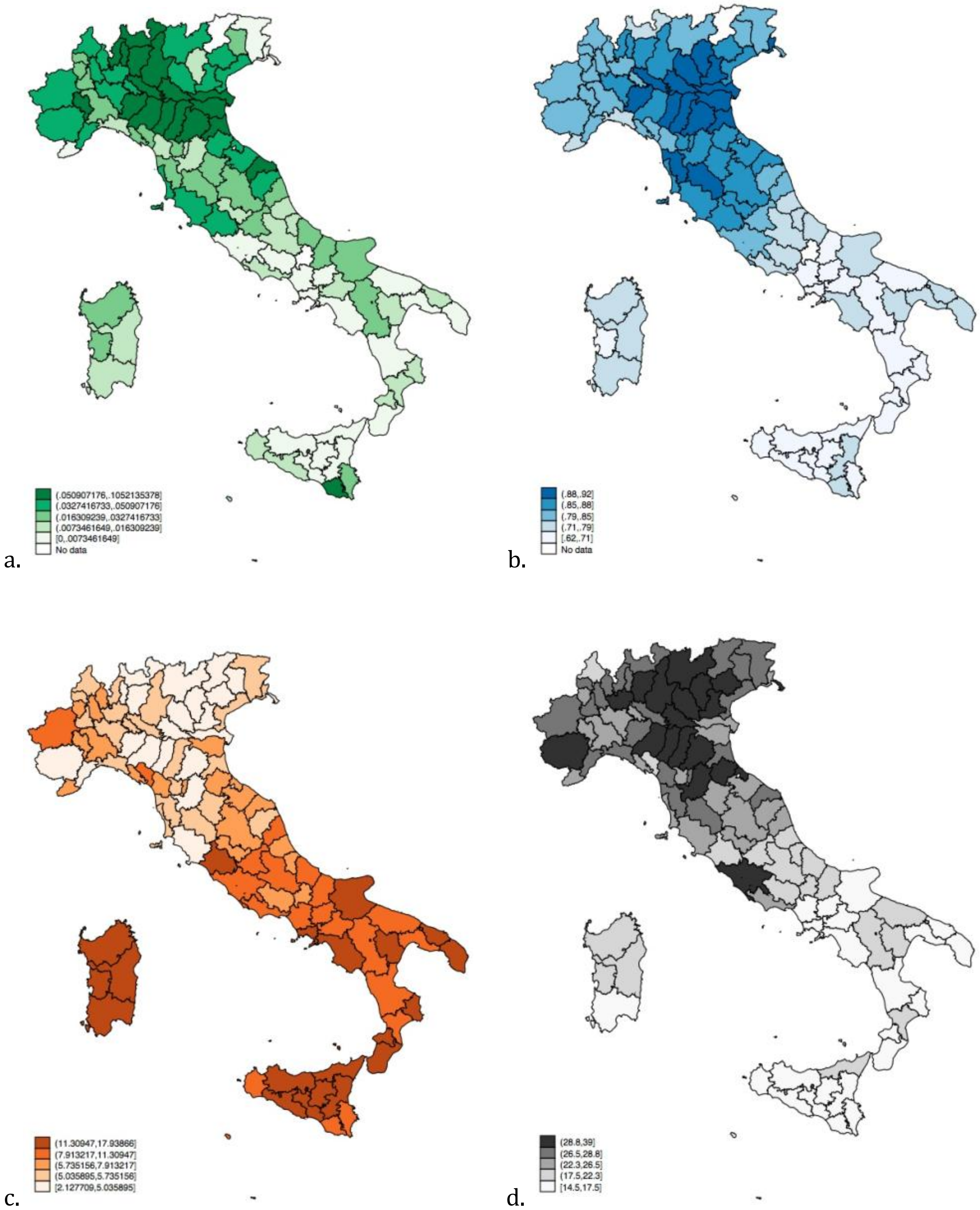
Notes: Interacted effects refer to the interactions between direct, indirect and overall effects and the variable listed at the top of each column. The interacting variable enters also as an independent covariate in the regression. Class size and the percentage of tenured teachers in the school are coded as dummy variables taking value one and zero when above and below the median. ESCS is coded as a dummy taking value one when below median and zero when above. The proportion of tenured teachers is not available for private schools (729 classes), for the public schools located in the Province of Trento (263 classes) and for five Sicilian public schools who did not transmit the information (20 classes). Average ESCS is not available for 2 classes. All regressions include the number of students enrolled in the class, regional dummies and regional dummies interacted with school size and with the number of fifth grade classes in the school. Estimates are weighted by class size. Standard errors adjusted for clustering at the school level in parentheses. One, two and three stars for statistical significance at the 10, 5 and 1 percent level of confidence.

Table 9. Interacting External Monitoring with Measures of Social Capital. Math Tests – V Grade. Dependent variable: Percentage of Correct Answers in the Class

| | (1)<br>Baseline | (2)<br>Interacted with Blood Donations | (3)<br>Interacted with Blood Donations and Macro Variables | (4)<br>Interacted with Turnover at Referenda | (5)<br>Interacted with Turnover at Referenda and Macro Variables |
|---|---|---|---|---|---|
| Direct Effect | -2.78*** | -2.48*** | -2.64*** | -2.63*** | -2.69*** |
| | (0.25) | (0.24) | (0.26) | (0.24) | (0.25) |
| Interacted Direct Effect | | 0.81*** | 0.41*** | 0.25*** | 0.14** |
| | | (0.11) | (0.12) | (0.04) | (0.06) |
| Indirect Effect | -0.82*** | -0.85*** | -0.93*** | -0.80*** | -0.88*** |
| | (0.28) | (0.26) | (0.29) | (0.26) | (0.20) |
| Interacted Indirect Effect | | -0.06 | -0.13 | 0.01 | -0.02 |
| | | (0.12) | (0.13) | (0.04) | (0.07) |
| Overall Effect | -3.60*** | -3.33*** | -3.57*** | -3.43*** | -3.57** |
| | (0.30) | (0.28) | (0.31) | (0.28) | (0.30) |
| Interacted Overall Effect | | 0.75*** | 0.28** | 0.26*** | 0.12* |
| | | (0.13) | (0.14) | (0.04) | (0.07) |
| Observations | 27,178 | 27,178 | 27,178 | 27,178 | 27,178 |
| R-squared | 0.15 | 0.15 | 0.15 | 0.15 | 0.15 |
| Additional covariates | Yes | Yes | Yes | Yes | Yes |
| Mean - Untreated Schools | 65.1 | 65.1 | 65.1 | 65.1 | 65.1 |

Notes: Interacted effects are the interactions between direct, indirect and overall effects and the variables listed at the top of each column. These variables enter as deviations from their sample means both in the interaction term and as an independent covariates in the regression. Social capital measures are not available for the provinces of Belluno and Isernia (147 classes). Macro variables: Per capita GDP and the unemployment rate in the province. All regressions include the number of students enrolled in the class, regional dummies and regional dummies interacted with school size and with the number of fifth grade classes in the school. Additional covariates are shown in Table 2 – panel A. Estimates are weighted by class size. Standard errors adjusted for clustering at the school level in parentheses. One, two and three stars for statistical significance at the 10, 5 and 1 percent level of confidence.

Figure 1. Geographical Distribution of Blood Donations, Average Turnout at Referenda, the Unemployment Rate and GDP per capita in the Italian Provinces.



Notes: Panel a): number of blood donations per 10,000 inhabitants in 1995. Panel b): average turnover at the referenda that took place between 1946 and 1989. Panel c): unemployment rate in 2009. Panel d) GDP per capita in 2009. The data are ordered by quintiles, with darker colours referring to the top quintile of the distribution.

## Appendix

*1) Tables*

Table A.1. The Effects of External Monitoring. Italian Tests – V Grade. Dependent variable: Percentage of Correct Answers in the Class.

| | (1)<br>Italy | (2)<br>North | (3)<br>Centre | (4)<br>South |
|---|---|---|---|---|
| Direct Effect | -2.61*** | -1.03*** | -2.17*** | -4.39*** |
| | (0.20) | (0.21) | (0.42) | (0.39) |
| Indirect Effect | -0.67*** | -0.38* | -0.81** | -0.99** |
| | (0.21) | (0.21) | (0.35) | (0.46) |
| Overall Effect | -3.28*** | -1.41*** | -2.98*** | -5.37*** |
| | (0.23) | (0.22) | (0.45) | (0.45) |
| Observations | 27,369 | 11,557 | 4,894 | 10,918 |
| R-squared | 0.19 | 0.28 | 0.22 | 0.17 |
| Additional covariates | Yes | Yes | Yes | Yes |
| Mean - Untreated Schools | 70.0 | 70.2 | 70.1 | 69.7 |

Notes: see Table 4.

Table A.2. The Effects of External Monitoring. Math Tests – II Grade. Dependent variable: Percentage of Correct Answers in the Class.

| | (1)<br>Italy | (2)<br>North | (3)<br>Centre | (4)<br>South |
|---|---|---|---|---|
| Direct Effect | -4.20*** | -1.57*** | -3.09*** | -7.50*** |
| | (0.29) | (0.32) | (0.54) | (0.58) |
| Indirect Effect | -1.22*** | -0.91*** | -1.37** | -1.53** |
| | (0.33) | (0.34) | (0.60) | (0.74) |
| Overall Effect | -5.42*** | -2.48*** | -4.47*** | -9.03*** |
| | (0.34) | (0.36) | (0.58) | (0.69) |
| Observations | 27,012 | 11,724 | 4,905 | 10,383 |
| R-squared | 0.11 | 0.08 | 0.09 | 0.08 |
| Additional covariates | Yes | Yes | Yes | Yes |
| Mean - Untreated Schools | 62.9 | 59.9 | 61.8 | 66.7 |

Notes: see Table 4.

Table A.3. The Effects of External Monitoring. Italian Tests – II Grade. Dependent variable: Percentage of Correct Answers in the Class.

| | (1)<br>Italy | (2)<br>North | (3)<br>Centre | (4)<br>South |
|---|---|---|---|---|
| Direct Effect x 100 | -3.40*** | -1.36*** | -2.17*** | -6.21*** |
| | (0.28) | (0.34) | (0.51) | (0.54) |
| Indirect Effect x 100 | -1.04*** | -0.71** | -1.25** | -1.33** |
| | (0.28) | (0.31) | (0.53) | (0.62) |
| Overall Effect x 100 | -4.44*** | -2.07*** | -3.42 | -7.54*** |
| | (0.29) | (0.34) | (0.56) | (0.58) |
| Observations | 27,025 | 11,721 | 4,911 | 10,393 |
| R-squared | 0.13 | 0.2 | 0.16 | 0.11 |
| Additional covariates | Yes | Yes | Yes | Yes |
| Mean - Untreated Schools | 65.9 | 65.0 | 66.2 | 66.7 |

Notes: see Table 4.

Table A.4. GLM estimates of the Effects of External Monitoring. Math Tests – V Grade. Dependent variable: Percentage of Correct Answers in the Class.

| | (1)<br>Italy | (2)<br>North | (3)<br>Centre | (4)<br>South |
|---|---|---|---|---|
| Direct Effect | -2.74*** | -0.97*** | -2.25*** | -4.73*** |
| | (0.25) | (0.28) | (0.47) | (0.48) |
| Indirect Effect | -0.80*** | -0.70*** | -0.72 | -1.04* |
| | (0.28) | (0.27) | (0.45) | (0.60) |
| Overall Effect | -3.54*** | -1.67*** | -2.97*** | -5.77*** |
| | (0.29) | (0.30) | (0.53) | (0.57) |
| Observations | 27,325 | 11,541 | 4,886 | 10,898 |
| Additional covariates | Yes | Yes | Yes | Yes |
| Mean - Untreated Schools | 65.1 | 63.9 | 64.0 | 66.8 |

Notes: see Table 4.

Table A.5. The Effects of External Monitoring. Math Tests – V Grade. Dependent variable: Percentage of Correct Answers in the Class. Finite Population Correction.

| | (1)<br>Italy | (2)<br>North | (3)<br>Centre | (4)<br>South |
|---|---|---|---|---|
| Direct Effect | -2.89*** | -1.08*** | -2.35*** | -5.05*** |
| | (0.12) | (0.14) | (0.23) | (0.24) |
| Indirect Effect | -0.83*** | -0.71*** | -0.70*** | -1.06*** |
| | (0.13) | (0.13) | (0.21) | (0.27) |
| Overall Effect | -3.72*** | -1.79*** | -3.05*** | -6.11*** |
| | (0.14) | (0.14) | (0.25) | (0.28) |
| Observations | 27,325 | 11,541 | 4,886 | 10,898 |
| R-squared | 0.15 | 0.19 | 0.15 | 0.15 |
| Additional covariates | Yes | Yes | Yes | Yes |
| Mean - Untreated Schools | 65.1 | 63.9 | 64.0 | 66.8 |

Notes: see Table 4.

Table A.6. The Effects of External Monitoring. Math Tests – V Grade. Dependent variable: Percentage of Correct Answers in the Class. Without Covariates.

| | (1)<br>Italy | (2)<br>North | (3)<br>Centre | (4)<br>South |
|---|---|---|---|---|
| Direct Effect | -2.82*** | -0.85*** | -2.04*** | -5.29*** |
| | (0.26) | (0.30) | (0.49) | (0.52) |
| Indirect Effect | -0.70** | -0.82*** | -0.46 | -0.70 |
| | (0.30) | (0.31) | (0.51) | (0.65) |
| Overall Effect | -3.52*** | -1.68*** | -2.50*** | -5.99*** |
| | (0.31) | (0.34) | (0.58) | (0.64) |
| Observations | 27,325 | 11,541 | 4,886 | 10,898 |
| R-squared | 0.03 | 0.01 | 0.01 | 0.03 |
| Additional covariates | No | No | No | No |
| Mean - Untreated Schools | 65.1 | 63.9 | 64.0 | 66.8 |

Notes: see Table 4. Each regression includes the number of students enrolled in the class, regional dummies and regional dummies interacted with school size and with the number of fifth grade classes in the school.

Table A.7. The Effects of External Monitoring. Math Tests – V grade. Dependent variable: Percentage Absent from the Test

| | (1)<br>Italy | (2)<br>North | (3)<br>Centre | (4)<br>South |
|---|---|---|---|---|
| Direct Effect | -0.53** | -0.50 | -0.47 | -0.55 |
| | (0.24) | (0.40) | (0.47) | (0.40) |
| Indirect Effect | -0.10 | 0.44 | -0.44 | -0.51 |
| | (0.24) | (0.36) | (0.42) | (0.44) |
| Overall Effect | -0.63** | -0.06 | -0.91** | -1.06** |
| | (0.25) | (0.40) | (0.46) | (0.42) |
| Observations | 27,325 | 11,541 | 4,886 | 10,898 |
| R-squared | 0.03 | 0.02 | 0.03 | 0.03 |
| Additional covariates | No | No | No | No |
| Mean - Untreated Schools | 11.0 | 10.4 | 11.7 | 11.4 |

Notes: see Table 4. The only covariates still included in the models are the number of students enrolled in the class, regional dummies and regional dummies interacted with school size and with the number of fifth grade classes in the school.

2) *External examiners.*

External examiners are selected by the regional education offices using criteria defined at the national level, from a pool of potential candidates composed by teachers and school principals, most of them retired. Eligible candidates must have personal characteristics that facilitate a fair collaboration with the school principal and the teachers involved in the test, should have a good knowledge of the evaluation procedure and should be familiar with the software and the procedure to transmit data to INVALSI.

Eligibility requires that examiners did not work during the two years before the test in the same municipality or in the same school they are going to supervise. If they are still active as teachers, they must be employed in a non-primary school. INVALSI conducted some investigation about possible cases of collusion between external examiners and school principals or teachers and did not find evidence of misconduct. Once appointed, external examiners need to coordinate with the school principal to prepare for the test. External examiners generally worked for two days and earned 200 euro per working day.

3) *Sampling procedure*.

The sampling procedure is a two-stage design and was taken from the IEA TIMMS survey, which INVALSI manages for Italy. Sampling takes place separately in each region. In the first stage, a pre-specified number of schools was randomly drawn from the population of schools located in the region. Schools with less than 10 students were excluded from the population and the rest were listed in a spreadsheet with the corresponding number of enrolled students in the second and fifth grades, which is the relevant measure of school size. The sampling method adopted is a PPS – probability proportional to size: the probability that each school is randomly sampled is proportional to school size. Practically, a software randomly samples schools from the sampling frame.[26] Only 5 schools have been replaced from the original sample. This low replacement rate is due to the fact that participation and compliance with INVALSI procedures are compulsory because of the law. The second stage of the sampling procedure is a simple random sampling of classes within the sampled schools. One or two classes per grade, depending uniquely on school size, were randomly selected from each

[26] Additional details on the sampling of schools can be found at the IEA TIMSS and PIRLS 2011 webpage http://timssandpirls.bc.edu/methods/pdf/TP_Sampling_Design.pdf

sampled school. No negotiation between school principals and INVALSI occurred to determine the selected classes.

The PPS technique implies that larger schools have a higher probability of being sampled than smaller schools. However, this difference in the selection probabilities is largely offset at the second stage of sampling by selecting a fixed number of classes with equal probability from the sampled school. Classes in large schools with many classes in the target grade have a lower probability of selection than classes in smaller schools that have just one or two classes.

### *4) From the initial dataset to the final sample*

Our data are drawn from the 2010 wave of the INVALSI SNV survey of educational achievements in Italian primary schools. These data are freely available from INVALSI. In this section of the Appendix we briefly describe our handling of the data.

1) We exclude Valle d'Aosta and the Province of Bolzano, because all classes in these areas were assigned to external monitoring.
2) We drop schools where there is a different number of second and fifth grade classes assigned to monitoring, because this outcome is inconsistent with the sampling scheme.
3) We drop classes with less than five students and schools with a single class per grade or with two classes if both were assigned to monitoring.

To illustrate the effects of these actions, we consider the Math test for fifth graders. For this group, the population consists of 7,700 schools, 30,476 classes and 565,064 students. Our initial dataset includes 7,541 schools, 29,811 classes and 491,421 non-disabled students in schools with more than ten students (smaller schools are excluded from testing) who were present during the testing day. Dropping data for the provinces of Aosta and Bolzano reduces the total number of schools to 7,502, with 29,647 classes and 489,396 students. Elimination of treated schools where there is a different number of second and fifth grade classes leaves us with 489,126 students allocated in 29,629 classes of 7,498 schools. Purging out classes with less than 5 students leaves us with 28,677 classes in 7,452 schools and a total of 486,531 students. After dropping schools with a single class in the grade or with two classes if both are treated we obtain our estimation sample, which consists of 6,108 schools, 27,325 classes and 462,570 students.

*5)Other data*

Unemployment and per capita GDP data refer to year 2009 and are drawn from EUROSTAT regional statistics database. Data on blood donations and the average turnout at referenda are from Guiso, Sapienza and Zingales (2004). The original data have been re-classified to match INVALSI classification, which includes 103 provinces

## References

Alesina, A. and Ichino, A., 2009. *L' Italia fatta in casa. Indagine sulla vera ricchezza degli italiani*. 1st ed. Milano: Mondadori.

Angelucci, M. and De Giorgi, G., 2009. Indirect Effects of an Aid Program: How Do Cash Transfers Affect Ineligibles' Consumption? *American Economic Review*, 99(1), pp. 486-508.

Banfield, E. C. (with L. Fasano), 1958. *The Moral Basis of a Backward Society*. 1st ed. Glencoe, IL: The Free Press.

Bokhari, F. A. S. and Schneider, H., 2011. School Accountability Laws and the Consumption of Psycho-stimulants. *Journal of Health Economics*, 30(2), pp. 355-372.

Cullen, J.B. and Reback, R., 2006. Tinkering Toward Accolades: School Gaming under a Performance Accountability System. In: Gronberg, T.J. and Jansen, D. W. (eds.), *Advances in Applied Microeconomics, 14*, pp.1-34.

Figlio, D. N., 2006. Testing, Crime and Punishment. *Journal of Public Economics,* 90(4), pp. 837-851.

Figlio, D. N. and Getzler, S.G, 2006. Accountability, Ability and Disability: Gaming the System. In: Gronberg, T.J. and Jansen, D. W. (eds.), *Advances in Applied Microeconomics, 14,* pp.35-49

Figlio, D. N. and Loeb, S., 2011. School Accountability. In: Hanushek, E. A., Machin, S. and Woessmann, L. (eds.), *Handbook of the Economics of Education*, 3, pp. 383-421.

Figlio, D. N., Winicki, J., 2005. Food for thought: the effects of school accountability plans on school nutrition, *Journal of Public Economics*, 89(2), pp. 381-394.

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

Guiso, L., Sapienza, P. and Zingales, L., 2010. Civic Capital as the Missing Link. NBER working Paper 15845.

Heckman, J.J., Lalonde, R. J. and Smith, J.A., 1999. The Economics and Econometrics of Active Labor Market Programs. In: Ashenfelter, O. C. and Card, D. (eds.), *Handbook of Labor Economics*, 3(1), pp. 1865-2097.

Hussain, I., 2012. Subjective Performance Evaluation in the Public Sector: Evidence from School Inspections. CEE Discussion Paper 135, London School of Economics.

Ichino, A. and Maggi, G. 2000. Work Environment and Individual Background: Explaining Regional Shirking Differentials in a Large Italian Firm". *Quarterly Journal of Economics*, 115(3), pp. 1057-1090.

INVALSI, 2010a. Sistema Nazionale di Valutazione – A.S. 2009/2010 - Rilevazione degli apprendimenti.

INVALSI, 2010b. Esami di Stato Primo Ciclo – A.S. 2009/2010 – Prova Nazionale. Prime Analisi.

Jacob, B. A., 2005. Accountability, incentives and behavior: the impact of high-stakes testing in the Chicago Public Schools. *Journal of Public Economics*, 89(5), pp. 761-796.

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

Miguel, E. and Kremer, M., 2004. Worms: Identifying Impacts on Education and Health in the Presence of Treatment Externalities. *Econometrica*, 72(1), pp.159-217.

Nannicini, T. et al., 2012. Social Capital and Political Accountability. *American Economic Journal: Economic Policy, forthcoming.*

Papke, L. E. and Wooldridge, J. M., 1996. Econometric Methods for Fractional Response Variables with an Application to 401(K) Plan Participation Rates. *Journal of Applied Econometrics*, 11(6), pp. 619-32.

Putnam, R. D. et al., 1993. *Making Democracy Work: Civic Traditions in Modern Italy.* 1st ed. Princeton, NJ: Princeton University Press.

Ricci, R. and Sestito, P., 2012. Il senso delle prove, *La Voce.Info*, July 6. 2012.