Discussion Paper No. 845

**WHEN THE CAT IS NEAR,
THE MICE WON'T PLAY:
THE EFFECT OF EXTERNAL EXAMINERS
IN ITALIAN SCHOOLS**

Marco Bertoni
Giorgio Brunello
Lorenzo Rocco

May 2012

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

Marco Bertoni[**]
University of Padova and CEP

Giorgio Brunello
University of Padova, IZA and CEsifo

Lorenzo Rocco
University of Padova

Abstract

Using a natural experiment designed by the Italian national test administrator (INVALSI) to monitor test procedures in Italian primary schools, this paper shows that the presence of an external examiner who monitors test procedures has both a direct and an indirect effect on the measured performance of monitored classes and schools. 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 instead the difference in performance between un-monitored classes in a school with an external examiner and un-monitored classes in schools without external monitoring. We find that having an external examiner in the class reduces the proportion of correct answers by 5.5 to 8.6 percent compared to classes in schools with no external monitor. Comparing results of un-monitored classes across schools with and without external examiners, we estimate indirect effects that range between 1.2 and 1.9 percent. The size of the overall effect of external supervision varies significantly across regions and it is higher in Southern Italy.

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

[**] Corresponding author. Department of Economics and Management "M. Fanno", University of Padova, Via del Santo 33, 35123 Padova, Italy. E-mail: marco.bertoni@studenti.unipd.it.

1. Introduction

Test – based accountability systems in education are becoming widespread (see Figlio and Loeb, 2011, for a review of the literature on accountability in education). On the one hand, formal quantitative assessments of educational achievements are increasingly informing policy making. On the other hand, test-based accountability systems generate incentives for teachers, students and school administrators to "game" the system in order to obtain better scores.

One way to address the threat of strategic manipulation is to have external examiners monitoring entirely or in part the test process. The purpose of this paper is to evaluate the effects of external monitoring on test outcomes. For this purpose, we exploit an Italian natural experiment which assigned external examiners to randomly selected classes and schools, in charge of monitoring students taking the test and of reporting results to the central test administrator (INVALSI). We compare the results of the test in the classes with an external examiner with the results in other classes of the same school - where the test was administered by a local teacher - and in classes of other schools with no external examiner.

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 of the same school with and without external examiners. 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 monitoring. In the absence of spill-over effects from monitored classes, the latter effect should be zero.

We find that the presence of external examiners reduces the average rate of correct answers in monitored classes with respect both to un-monitored classes of the same school and to un-monitored classes in schools with no external examiner. The effect of monitoring spills over un-monitored classes of schools with external examiners. We estimate that having an external examiner reduces the percentage of correct answers by 3.6 to 5.4 percentage points (equivalent to 5.5 to 8.6 percent with respect to the average score in classes belonging to schools with no external examiner), depending on the grade and the test. The estimated direct effect ranges from 2.8 to 4.2 percentage points (4.3 to 6.6 percent), and the residual indirect effect from 0.8 to 1.2 percentage points (1.2 to 1.9 percent). There is also evidence

that the reduction in measured performance is significantly higher in the schools located in less developed Southern Italy than in Northern schools.

We argue that the negative effects of the presence of an external examiner on measured test scores are more likely to be due to reduced cheating than to the negative effects of distraction from having a stranger in the class. Interestingly, the effects of monitoring spread from the monitored class to the other classes in the school. We propose two alternative reasons why this might be the case. 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 based on the idea that teachers in the same school share a concern both for average and for relative performance within the school. Both explanations predict that the indirect effect falls with the number of classes in the monitored school. Our evidence confirms this prediction.

Our study contributes to the literature in several 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 spills over to un-monitored classes of the same school. This is important when considering the costs and benefits of having external examiners.

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

2. Review of the literature

In their seminal work, Jacob and Levitt (2003) show that a fraction of Chicago teachers reacted to accountability pressures by fraudulently completing student tests in order to improve outcomes. They develop algorithms to detect cheating that rely on the presence of unusual response patterns within the answers given by students and by looking at the performance of the same class over different years. Aside from outright cheating, 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 student performance. Attempts to increase test scores by taking psycho-stimulant drugs are documented for the US by Bokhari and Schneider (2011), who show a higher rate of diagnosis of "attention deficit/hyperactivity disorder" in states where there are stronger accountability laws.

To our knowledge, little has been done in this literature to investigate the indirect effects of monitoring teachers and students. These effects are discussed instead in the broad literature on labour market policy evaluation. Heckman, Lalonde and Smith (1999) review this literature and 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[1] tests were introduced in Italian primary schools in 2008[2], with the purpose of evaluating school productivity (in terms of value added) by using standardized tests in Italian and maths. 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 depend on the test outcomes. Nonetheless, the government in charge when they were introduced explicitly supported the view that the tests should matter in the evaluation of the performance of teachers. Even in the absence of explicit high stakes, pressure to perform well was high because the school reputation was also at stake. While results of the tests were not publicly

---

[1] INVALSI is the National Institute for the Evaluation of the Educational System, in charge of the design and administration of standardized educational tests in Italy.
[2] See Law Decree n.147 – 2007, and Ministry of Education and Research Decree n.74 and 76 – 2009.

available, schools and their principals could access the results of their students and decide to make them public, which created another incentive to perform well. Since 2008 the tests have been administered every year. In this paper, we focus on the 2009/2010 wave because of its peculiar design features. First, the wave was the first to test the entire population of Italian primary school students in their second and fifth grade. Second, in order to monitor the correct implementation of the mandated testing procedures, INVALSI selected ex-ante a random sample of classes where the test was administered in the presence of an external observer (see INVALSI, 2010a). This observer had two main tasks: to be present in the class during the test administration and to report student answers on the dedicated answer sheets and transmit them to INVALSI for correction. In the other classes, instead, 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.

Classes were selected using a two-stages sampling scheme, stratified by region[3]. In the first stage, a given number of schools in each region were randomly selected by probabilistic sampling, with the probability of inclusion proportional to school size, measured by the total number of students enrolled in the tested grades. In the second stage, one or two classes within each treated school were selected by simple random sampling. Figure 1 summarizes the sampling scheme and Table 1 shows the characteristics of the national sample for each grade[4].

We have access to micro data for the entire population of second and fifth grade primary school students who took the INVALSI tests in 2009/2010. The data includes information on the answers given by each student to each question in the test, on past individual marks in Italian and maths and on parental background. The latter information was filled in by school offices. Exclusively for fifth graders, INVALSI used a student questionnaire asking further information on their family background and on feelings and motivation during the tests. We also 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 and whether the school is public or private[5].

---

[3] 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 the following analysis.

[4] INVALSI (2010a) has developed its own procedure to detect cheating among monitored classes and concluded that there is no cheating problem within this group of classes. The cheating detection algorithm is described in INVALSI (2010b).

[5] Summing up, the following variables are available: 1) at the school level: private or public school, whether the school offers full time schedule; 2) at the class level: class size, full- or part- time schedule (measured in term of the schedule of the median student in the class, to avoid measurement errors); 3) at individual level: gender,

4. Identification and Estimation

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

Let the dummy variable S_j take value 1 if school j has been assigned to school-level treatment (and zero otherwise) and the dummy C_i take value 1 if class i has been assigned to class-level treatment. The observed outcome Y_{ij} of class i in school j may be represented in terms of potential outcomes as follows:

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

We are interested in the identification and estimation of the following quantities:
- The average direct effect of monitoring: $E[Y_{11}-Y_{01}]$,
- The average indirect effect of monitoring: $E[Y_{01}-Y_{00}]$,
- The average total effect of monitoring: $E[Y_{11}-Y_{00}]$.

where E is the mean operator.

The sampling procedure – described in INVALSI (2010a) – has the following features: a) within a region, two schools of the same size (i.e., same number of students enrolled in the second and fifth grade) have the same probability of being assigned to school-level treatment; b) two treated schools of same size have the same probability of being assigned to the selection of one or two classes per grade for external monitoring; c) two classes of a given grade belonging to two different treated schools with the same size have the same probability of being monitored if the number of classes in the grade is the same in the two schools. We thus have conditional randomization, implying that:

---

place of birth, citizenship, attendance of pre-primary school, age, maths and Italian grade in previous semester, employment, education and nationality of parents. Moreover, for fifth grade students we also have information on the following variables: motivation and feelings during the test, whether at home the student has his own bedroom, internet access, an encyclopedia, his own desk, a computer and a place for doing his homework, the number of books that are present at home, the number of siblings, whether he lives with both parents or not, the language spoken at home, whether he gets help with his homework or not.

- 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;
- assignment to class - level treatment for a class of a given grade in a treated school is random conditional on the size of the school, measured by the number of students enrolled in the second and fifth grade and by the number of classes in the selected grade.

Let RD be the vector of regional dummies, RS_j the vector of regional dummies interacted with the size of school j and RC_j the vector of regional dummies interacted with the number of classes in a given grade in school j. Letting R = [RD, RS, RC], conditional independence in each considered grade requires that

$$Y_{00}, Y_{01}, Y_{11} \perp S_j, C_i \mid R \tag{2}$$

Under this assumption, the effects of monitoring are given by

$$E[Y_{ij} \mid C_i = 1, S_j = 1, R] - E[Y_{ij} \mid C_i = 0, S_j = 1, R] = E[Y_{11} - Y_{01} \mid R] \tag{3}$$

$$E[Y_{ij} \mid C_i = 0, S_j = 1, R] - E[Y_{ij} \mid C_i = 0, S_j = 0, R] = E[Y_{01} - Y_{00} \mid R] \tag{4}$$

and

$$E[Y_{ij} \mid C_i = 1, S_j = 1, R] - E[Y_{ij} \mid C_i = 0, S_j = 0, R] = E[Y_{11} - Y_{00} \mid R] \tag{5}$$

We test for successful randomization by checking whether the vector *X* of pre-determined covariates at the school, class and individual level are balanced in the treatment and control sub-samples[6]. While we have data on second and fifth graders, we focus hereinafter on the latter to save space. Some results for second graders are presented in the Appendix. Table 2 shows the means and standard deviations of the variables in the vector *X* for the full sample.

---

[6] The covariates considered are those listed in footnote 5.

To test for balancing we first focus on differences between treated and untreated schools, and then consider differences between treated and untreated classes within treated schools. The tests are based on the following regression models:

$$X_j = \alpha + \beta t_j + \rho RD + \sigma RS_j + \varepsilon_j \tag{6}$$

and

$$X_{ij} = \alpha + \beta t_{ij} + \rho R + \varepsilon_{ij} \tag{7}$$

for the first and the second stage randomization respectively, where t are dummy variables for treatment at the school and class level. Table 3 reports the point estimates of the β coefficients and the p-values of the test of the hypothesis H_0: $\beta=0$ for each covariate.[7] Results do not suggest that unbalancing is a problem for school and class-level variables. 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 the following regressions. Turning to individual variables, although in some cases we detect statistically significant differences across the various groups, the point estimates show that these differences are virtually zero in almost all cases. Prudentially, we introduce the variables with statistically significant differences as covariates in our regressions to eliminate the risk of unbalancing and increase precision.

We estimate the effects of monitoring using the following regression

$$Y_{ij} = \alpha + \beta C_{ij} S_j + \gamma S_j + \delta R_j + \theta X_{ij} + \varepsilon_{ij} \tag{8}$$

Standard errors are robust and weighted by class size whenever the dependent variable is the average test performance in the class. From this regression we obtain the direct, indirect and total effects of external monitoring as follows

---

[7] Results are shown for the fifth grade maths test only, but are very similar for grade II (available from the authors upon requests). In these and the other estimates of this paper we drop from the sample hereafter the schools where there is a different number of second and fifth grade classes assigned to monitoring, because this outcome is not in line with the sampling scheme and may be a symptom of manipulation in the selection of treated classes. We also drop the classes with less than five students and the schools with a single class per grade or with two classes if both were assigned to monitoring.

$$\beta = E[Y_{ij} \mid C_i = 1, S_j = 1, R, X] - E[Y_{ij} \mid C_i = 0, S_j = 1, R, X] \qquad (9)$$

$$\gamma = E[Y_{ij} \mid C_i = 0, S_j = 1, R, X] - E[Y_{ij} \mid C_i = 0, S_j = 0, R, X] \qquad (10)$$

and

$$\gamma + \beta = E[Y_{ij} \mid C_i = 1, S_j = 1, R, X] - E[Y_{ij} \mid C_i = 0, S_j = 0, R, X] \qquad (11)$$

4. Results

Table 4 presents our estimates for fifth graders and the maths test. The dependent variable is the mean percentage of correct answers given by each student in the class[8]. While the first column in the table considers all Italian regions, the remaining columns present separate 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. 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. As shown in Table A.2 in the Appendix, the total effect is somewhat larger for second graders (5.4 percentage points, or 8.6% of the average score in untreated schools).

The size of the total, direct and indirect effects varies with the macro area and is highest in Southern regions (-8.9%) and lowest in Northern Italy (-2.6%). How do we explain these negative effects? One possibility is that young students are distracted by the presence of a stranger in the class and under-perform as a consequence. The other possibility is that either students or teachers in classes without the external examiner engage in outright cheating.

We believe that the second explanation is more likely for two 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 the student during the test, which included agreeing or not with the following sentences: a) I was

---

[8] Results for Italian and second graders are qualitatively similar and are shown in Appendix 1.

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 equation (8) on the full sample 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 or nervousness. Quite the opposite, there is some evidence that students in these classes were less nervous and calmer.

Second, we examine the distribution of results within classes. If outright cheating by students and/or teachers was taking place in the classes without the external examiner, we should find that – ceteris paribus - the coefficient of variation and the standard deviation of results are lower than in classes with the examiner, where cheating is precluded by definition[9]. 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. In either case, the within – class coefficient of variation and the standard deviation of the percentage of correct answers are bound to decline. While distraction can reduce average performance, it is not clear that is should reduce also its variability.

Table 6 shows the effects of the presence of an external examiner on the within – class coefficient of variation and standard deviation of the percentage of correct answers, as well as on the bottom quartile, median and top quartile of the distribution of test scores. The table only considers Southern Italy and the maths scores of fifth graders. We find that both the coefficient of variation and the standard deviation are significantly higher in classes with an external examiner. 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. When compared with students in untreated schools, having an external examiner reduces the score of these students by a 13.7 percent (-7.62/55.6). This effect is stronger for second grade students, where it reaches a striking 20.7 percent.

We further document that our results are the outcome of cheating by computing an index of heterogeneity in the pattern of answers given in each class. For each class and question we compute a modified version of the Herfindahl Index[10]

---

[9] We are assuming that the external examiner does not collude with school teachers and principals to boost school results.

[10] See INVALSI (2010b)

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

where s_a is the within-class share of students who chose "a" in the set A of possible choices[11]. This indicator varies between 0 and 1, with higher values signalling a more heterogeneous pattern of answers to a given question. We obtain an overall measure of heterogeneity by averaging H across all questions in the test. The lower this measure the higher the likelihood that systematic cheating has occurred in the class. Table (7) reports the results for fifth graders engaged in the maths test. The estimates in column (1), which refer to the entire country, show that monitored classes have a more heterogeneous pattern of answers than un-monitored classes. There is also evidence of a statistically significant indirect effect. Again, the effects of external monitoring on the heterogeneity of answers increase significantly as we move from Northern to Southern Italy (columns (2) to (4)).

5. 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 average percentage of correct answers in the class) we implement the GLM estimator proposed by Papke and Wooldridge (1996) to deal with fractional dependent variables. Results, shown in Table A.4 in Appendix 1, are in line with the baselines 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 (see 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 balancing. 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 Appendix 1 do not provide any strong evidence in this direction. Finally, we test directly for differences in absenteeism across treatment statuses, using as dependent variable the percentage of students absent from the

---

[11] We treat missing values as a separate category.

test in each class. Again, differences in behaviour across the three groups are minimal (see Table A.7 in the Appendix).

6. Discussion

The paper has three empirical findings: 1) having an external examiner reduces the average test score in the class with respect to classes where the test is administered by local teachers; 2) the effect of the presence of an external examiner spills over to other classes in the same school; 3) the size of the external monitoring effects is larger in Southern Italy than in the rest of the country. We have argued that the first result is consistent with teachers and students manipulating test results in the classes without external examiners. We have also ruled out an alternative explanation based on student distraction in treated classes.

The indirect effect of monitoring in school tests has not been detected before and deserves some 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 fear is rather irrational, as the external examiner's mandate is 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. To illustrate, consider a school where a single class is supervised by an external examiner. If teachers administering the test in other classes cheat freely, their classes will look much better than the supervised class, where cheating is not possible. This may generate conflicts with the teacher in charge of the supervised class. To reduce these conflicts, teachers in un-monitored classes reduce their cheating.

Conclusions

Test-based accountability systems in education may be gamed by 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 both direct and indirect effects on the measured performance of tested classes and schools. The direct effect is on the monitored class, the indirect effect is on the un-monitored classes of the same school.

We have used a natural experiment designed by the Italian national test administrator (INVALSI) to monitor test procedures in Italian primary schools. The design of the experiment

is such that we observe test results in the universe of Italian primary schools. However, only some randomly selected classes of randomly selected schools were assigned to external monitoring. We have used random assignment to treatment to estimate both the direct and indirect effects of monitoring, by comparing monitored and unmonitored classes within the same schools in the former case and unmonitored classes in school where there is an external examiner and unmonitored classes in schools without the external examiner.

We have found that the overall effects (direct plus indirect) of having an external examiner are statistically significant and sizeable: depending on the grade, the presence of the examiner reduces the average percent of correct answers in the class by 5.5 to 8.5 percent with respect to classes in schools with no external monitor. There is also evidence that indirect effects are present, although their size is rather small (about 20 percent of the total effect). Using supporting evidence on the motivation of students and on the effect of external monitoring on the distribution of results within classes, we have argued that these effects are consistent with teachers and/or students manipulating results in the classes without the external examiner.

Perhaps unsurprisingly, the size of the overall effect of external supervision is highest in Southern Italy, where one may suspect that poor social capital leads to extensive cheating. Finally, we have argued that the presence of an indirect effect, albeit small, can be explained either with fear of supervision or with a behavioural model where teachers administering the tests dislike excessive dispersion of test results within the school.

Our results speak clearly against using raw average test results to rank schools. Such ranking would be grossly misleading, especially in Southern Italy, where the overall effect of monitoring is close to 10 percent. Given the unfeasibility of having an external examiner in all the classes of the universe of Italian primary schools, a more feasible alternative is to use estimates such as ours to properly correct the average scores of un-monitored schools[12].

---

[12] INVALSI (2010b) proposes an alternative methodology to detect and correct for cheating that is not based on the comparison between monitored an un-monitored classes.

FIGURES AND TABLES

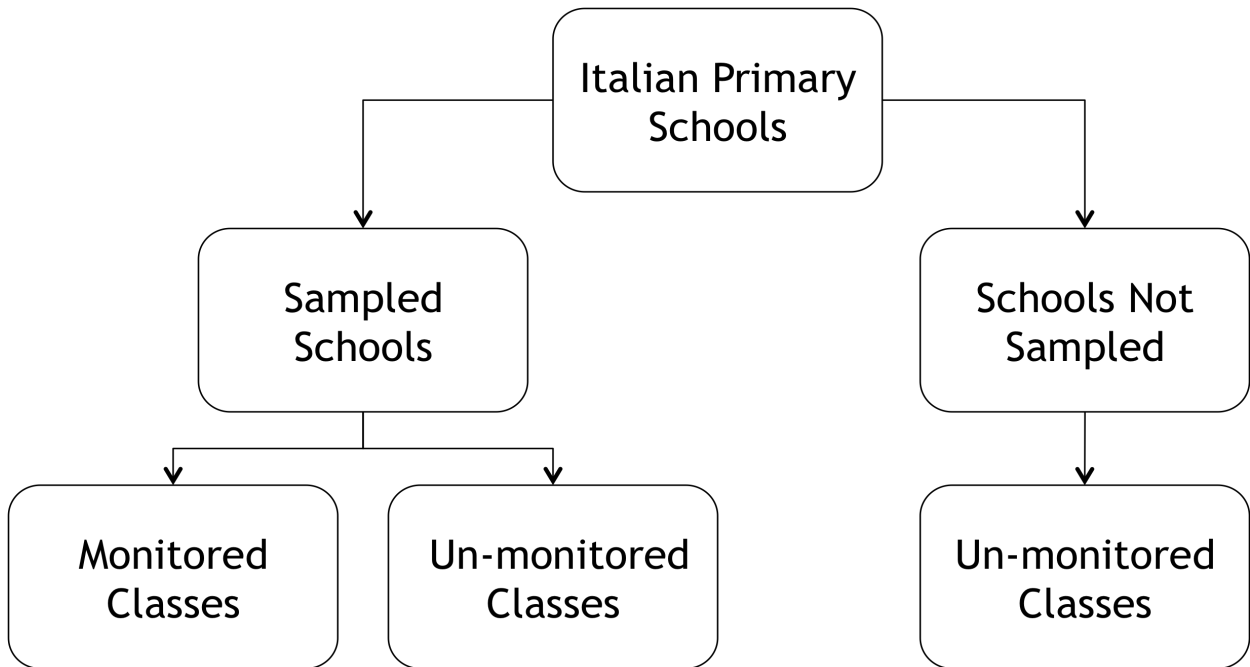Figure 1. Sampling Scheme – INVALSI SNV Test

Table 1. National Samples – INVALSI SNV Test

	Total schools	Total classes	Total students	Sampled schools	Sampled classes	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. Covariates - Maths - V grade
Panel a.

	Mean	St Dev		Mean	St Dev
Gender			Mother occupation		
Missing (%)	0.01	0.10	Missing (%)	0.20	0.40
Male (%)	0.50	0.50	Unemployed or retired (%)	0.35	0.48
Place of birth			Employee (%)	0.31	0.46
Missing (%)	0.04	0.20	Entrepreneur (%)	0.08	0.28
Italy (%)	0.89	0.31	Middle manager (%)	0.06	0.23
Citizenship			Father occupation		
Missing (%)	0.02	0.15	Missing (%)	0.22	0.41
Italian (%)	0.89	0.32	Unemployed or retired (%)	0.04	0.19
First generation foreigner (%)	0.05	0.22	Employee (%)	0.39	0.49
Second generation foreigner (%)	0.04	0.20	Entrepreneur (%)	0.25	0.43
Pre-primary school			Middle manager (%)	0.11	0.31
Missing (%)	0.15	0.35	Mother education		
Yes (%)	0.83	0.37	Missing (%)	0.21	0.41
Age			Primary (%)	0.39	0.49
Missing (%)	0.01	0.10	Secondary (%)	0.29	0.45
Older than regular (%)	0.03	0.16	Tertiary (%)	0.11	0.32
Regular (%)	0.87	0.33	Father education		
Younger than regular (%)	0.09	0.29	Missing (%)	0.22	0.42
Maths grade			Primary (%)	0.43	0.49
Missing (%)	0.07	0.26	Secondary (%)	0.25	0.43
1-4 (%)	0.00	0.04	Tertiary (%)	0.10	0.30
5 (%)	0.04	0.20	Mother nationality		
6-7 (%)	0.38	0.48	Missing (%)	0.09	0.28
8-10 (%)	0.51	0.50	Italian (%)	0.80	0.40
Italian grade			Father nationality		
Missing (%)	0.07	0.25	Missing(%)	0.09	0.29
1-4 (%)	0.00	0.04	Italian (%)	0.82	0.39
5 (%)	0.04	0.19	Private school	0.05	0.23
6-7 (%)	0.41	0.49	Full time schedule class	0.23	0.42
8-10 (%)	0.48	0.50	Class size	19.00	4.65

Panel b.

	Mean	St Dev		Mean	St Dev
Has own bedroom			Number of siblings		
Missing (%)	0.03	0.17	Missing (%)	0.02	0.15
Yes (%)	0.55	0.50	0 (%)	0.15	0.36
Has internet access			1 (%)	0.55	0.50
Missing (%)	0.03	0.16	2 (%)	0.20	0.40
Yes (%)	0.76	0.43	3 (%)	0.05	0.21
Has an encyclopedia			4 or more (%)	0.03	0.17
Missing (%)	0.03	0.16	Lives with		
Missing (%)	0.71	0.46	Missing (%)	0.02	0.15
Has own desk			Both parents (%)	0.86	0.35
Missing (%)	0.02	0.15	One parent only (%)	0.06	0.24
Yes (%)	0.85	0.36	Both parents alternatively (%)	0.05	0.22
Has a PC			Others (%)	0.01	0.08
Missing (%)	0.03	0.16	Language spoken at home		
Yes (%)	0.75	0.43	Missing (%)	0.04	0.21
Has a place for homework			Italian (%)	0.73	0.44
Missing (%)	0.03	0.16	Dialect (%)	0.15	0.36
Yes (%)	0.84	0.37	Other (%)	0.07	0.25
Number of books at home			Help with homework		
Missing (%)	0.04	0.20	Missing (%)	0.07	0.26
0-10 (%)	0.12	0.33	No homework (%)	0.01	0.07
11-25 (%)	0.25	0.43	No help needed (%)	0.20	0.40
26-100 (%)	0.31	0.46	Parents (%)	0.45	0.50
101-200 (%)	0.15	0.36	Siblings (%)	0.12	0.32
>200 (%)	0.12	0.33	Private teacher (%)	0.03	0.17
			Other (%)	0.04	0.20
			No one (%)	0.09	0.28

Notes: The table reports the mean and standard deviation of each covariate included in the regressions. These statistics are based on individual, school and class level data. All covariates have been categorized and reported as dummy variables. While variables in Panel a are available for students in both grades, those in Panel b are available only for fifth graders.

Table 3 - Balancing Tests - Maths - V grade

Panel a.

	First Stage	Second Stage		First Stage	Second Stage
Private school (%)	0.003	.	**Mother occupation**		
	(0.637)	.	Missing (%)	-0.010	-0.023
Full time schedule (%)	0.015	0.011		(0.353)	(0.008)
	(0.315)	(0.327)	Unemployed or retired (%)	0.005	0.010
Class size (%)	0.039	0.425		(0.394)	(0.065)
	(0.665)	(0.000)	Employee (%)	0.002	0.004
Gender				(0.682)	(0.400)
Missing (%)	0.005	0.021	Entrepreneur (%)	0.001	0.005
	(0.002)	(0.000)		(0.772)	(0.020)
Male (%)	-0.003	-0.007	Middle manager (%)	0.002	0.003
	(0.191)	(0.068)		(0.291)	(0.137)
Place of birth			**Father occupation**		
Missing (%)	-0.014	-0.025	Missing (%)	-0.010	-0.023
	(0.000)	(0.000)		(0.330)	(0.007)
Italy (%)	0.014	0.026	Unemployed or retired (%)	-0.001	0.000
	(0.000)	(0.000)		(0.558)	(0.973)
Citizenship			Employee (%)	0.000	0.014
Missing (%)	-0.008	-0.011		(0.970)	(0.015)
	(0.006)	(0.000)	Entrepreneur (%)	0.010	0.007
Italian (%)	0.008	0.010		(0.035)	(0.136)
	(0.032)	(0.003)	Middle manager (%)	0.002	0.003
First generation foreigner (%)	-0.001	0.000		(0.610)	(0.407)
	(0.438)	(0.863)	**Mother education**		
Second generation foreigner (%)	0.002	0.002	Missing (%)	-0.011	-0.029
	(0.215)	(0.238)		(0.350)	(0.001)
Pre-primary school			Primary (%)	0.006	0.018
Missing (%)	-0.019	-0.008		(0.410)	(0.005)
	(0.015)	(0.236)	Secondary (%)	0.002	0.010
Yes (%)	0.018	0.008		(0.634)	(0.043)
	(0.027)	(0.211)	Tertiary (%)	0.002	0.001
Age				(0.593)	(0.822)
Missing (%)	0.006	0.019	**Father education**		
	(0.000)	(0.000)	Missing (%)	-0.011	-0.026
Older than regular (%)	0.000	0.001		(0.306)	(0.004)
	(0.733)	(0.664)	Primary (%)	0.011	0.016
Regular (%)	-0.008	-0.016		(0.171)	(0.020)
	(0.002)	(0.000)	Secondary (%)	-0.001	0.008
Younger than regular (%)	0.001	-0.003		(0.860)	(0.091)
	(0.510)	(0.171)	Tertiary (%)	0.001	0.002
Maths grade				(0.708)	(0.507)
Missing (%)	-0.023	-0.008	**Mother nationality**		
	(0.000)	(0.126)	Missing (%)	-0.015	-0.012
1-4 (%)	0.000	-0.001		(0.015)	(0.029)
	(0.480)	(0.039)	Italian (%)	0.013	0.011
5 (%)	0.001	0.000		(0.038)	(0.052)
	(0.643)	(0.888)	**Father nationality**		
6-7 (%)	0.010	0.007	Missing (%)	-0.015	-0.011
	(0.025)	(0.181)		(0.018)	(0.031)
8-10 (%)	0.012	0.002	Italian (%)	0.013	0.008
	(0.032)	(0.778)		(0.033)	(0.150)
Italian grade					
Missing (%)	-0.023	-0.006			
	(0.001)	(0.209)			
1-4 (%)	0.000	0.000			
	(0.736)	(0.277)			
5 (%)	0.000	0.001			
	(0.721)	(0.727)			
6-7 (%)	0.007	0.004			
	(0.129)	(0.501)			
8-10 (%)	0.015	0.002			
	(0.010)	(0.683)			

Panel b.

	First Stage	Second Stage
Has own bedroom		
Missing (%)	-0.007	-0.008
	(0.007)	(0.000)
Yes (%)	-0.002	0.004
	(0.570)	(0.343)
Has internet access		
Missing (%)	-0.006	-0.007
	(0.011)	(0.001)
Yes (%)	0.006	0.007
	(0.066)	(0.042)
Has an encyclopedia		
Missing (%)	-0.006	-0.006
	(0.014)	(0.005)
Yes (%)	0.008	0.015
	(0.010)	(0.000)
Has own desk		
Missing (%)	-0.006	-0.006
	(0.013)	(0.003)
Yes (%)	0.004	0.007
	(0.103)	(0.020)
Has a PC		
Missing (%)	-0.006	-0.006
	(0.020)	(0.004)
Yes (%)	0.008	0.010
	(0.007)	(0.006)
Has a place for homework		
Missing (%)	-0.006	-0.006
	(0.011)	(0.003)
Yes (%)	0.007	0.006
	(0.006)	(0.051)
Number of books at home		
Missing (%)	-0.008	-0.007
	(0.001)	(0.001)
0-10 (%)	-0.001	0.000
	(0.788)	(0.975)
11-25 (%)	-0.003	0.000
	(0.194)	(0.973)
26-100 (%)	0.001	0.005
	(0.542)	(0.140)
101-200 (%)	0.005	0.003
	(0.022)	(0.346)
>200 (%)	0.006	0.000
	(0.020)	(0.875)

	First Stage	Second Stage
Number of siblings		
Missing (%)	-0.007	-0.007
	(0.002)	(0.000)
0 (%)	0.000	0.000
	(0.980)	(0.934)
1 (%)	0.004	0.006
	(0.104)	(0.091)
2 (%)	0.002	0.001
	(0.365)	(0.811)
3 (%)	0.001	0.000
	(0.345)	(0.887)
4 or more (%)	0.000	0.000
	(0.655)	(0.925)
Lives with		
Missing (%)	-0.008	-0.008
	(0.001)	(0.000
Both parents (%)	0.008	0.006)
	(0.003)	(0.050
One parent only (%)	0.000	0.001)
	(0.966)	(0.693
Both parents alternatively (%)	0.000	0.002)
	(0.711)	(0.318
Others (%)	0.000	0.000)
	(0.372)	(0.917
Language spoken at home		
Missing (%)	-0.008	-0.008
	(0.001)	(0.001)
Italian (%)	0.003	0.006
	(0.415)	(0.195)
Dialect (%)	0.004	0.002
	(0.237)	(0.595)
Other (%)	0.001	0.000
	(0.498)	(0.951)
Help with homework		
Missing (%)	-0.007	-0.005
	(0.004)	(0.073)
No homework (%)	-0.001	-0.001
	(0.019)	(0.009)
No help needed (%)	-0.002	0.006
	(0.462)	(0.097)
Parents (%)	0.005	0.000
	(0.087)	(0.989)
Siblings (%)	0.003	-0.001
	(0.046)	(0.684)
Private teacher (%)	-0.001	0.001
	(0.349)	(0.339)
Other (%)	0.002	-0.001
	(0.030)	(0.276)
No one (%)	0.000	0.002

Notes: the table shows the point estimates and p-values (in parentheses) of the balancing tests at the school level (first stage) and at class level within treated schools (second stage). For individual variables, we compute school or class averages and test for balancing using regressions (5) and (6) for the first and second stage, respectively. Full time schedule refers to schools offering this option in the first stage and to the schedule of the single class in the second stage. While variables in Panel a are available for students in both grades, those in Panel b are available for fifth grade students only. All regressions are weighted by the number of students in each school or class and use robust standard errors. *, p<.1; **, p<.05; ***p<.01.

Table 4. Weighted OLS estimates of the Effects of External Monitoring. Maths – V grade.
Dependent variable: Percentage of Correct Answers in the Class.

	(1) Italy	(2) North	(3) Centre	(4) South
Direct Effect x 100	-2.79*** (0. 27)	-0.99*** (0.30)	-2.27*** (0.52)	-4.92*** (0.56)
Indirect Effect x 100	-0. 81*** (0.19)	-0.70*** (0.20)	-0.73** (0.34)	-1.04*** (0.40)
Total Effect x 100	-3.59*** (0.250)	-1.69*** (0.276)	-2.99*** (0.484)	-5.96*** (0.501)
Observations	27,325	11,541	4,886	10,898
R-squared	0.97	0.99	0.98	0.96
Covariates in Table 2	Yes	Yes	Yes	Yes
Mean - Untreated Schools	65.1	63.9	64.0	66.8

Notes: column (1) is for the full country and columns (2) to (4) are for Northern, Central and Southern Italy, respectively. All models include regional dummies and regional dummies interacted with school size, with the number of classes in fifth grade in the school and with number of students in the class. Estimates are weighted by class size. Robust standard errors in parentheses. *** $p<0.01$, ** $p<0.05$, * $p<0.1$.

Table 5. Weighted OLS estimates of the Effects of External Monitoring on Student Motivation. Maths – V grade. Dependent variable: Percentage of Positive Answers in the Class.

	(1)	(2)	(3)	(4)
Direct Effect x 100	0.25	-0.92***	-0.08	0.635
	(0.43)	(0.29)	(0.408)	(0.407)
Indirect Effect x 100	0.25	0.01	0.356	-0.007
	(0.28)	(0.19)	(0.261)	(0.264)
Total Effect x 100	0.50	-0.90***	0.275	0.628*
	(0.39)	(0.272)	(0.375)	(0.373)
Observations	27,141	27,142	27,141	27,140
R-squared	0.94	0.75	0.92	0.93
Covariates in Table 2	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 agree with the following sentences: 1) I was already anxious before starting the test; 2) I was so nervous I couldn't find the answers; 3) while answering, I felt like I was doing badly; 4) while answering, I was calm. The estimates refer to the entire country.

Table 6. Weighted OLS estimates of the Effects of External Monitoring on the Coefficient of Variation, the Standard Deviation and the Quartiles of the Distribution of Correct Answers within the Class. Southern Italy - Maths – V grade.

	(1) Coefficient of Variation	(2) Standard Deviation	(3) p_25	(4) p_50	(5) p_75
Direct Effect x 100	3.82***	1.34***	-6.72***	-5.57***	-4.18***
	(0.39)	(0.17)	(0.66)	(0.62)	(0.56)
Indirect Effect x 100	0.47*	0.13	-0.90*	-0.91**	-0.93**
	(0.26)	(0.12)	(0.48)	(0.44)	(0.39)
Total Effect x 100	4.30***	1.47***	-7.62***	-6.48***	-5.10***
	(0.368)	(0.157)	(0.591)	(0.561)	(0.508)
Observations	10,898	10,898	10,898	10,898	10,898
R-squared	0.83	0.88	0.92	0.95	0.97
Covariates	Yes	Yes	Yes	Yes	Yes
Mean - Untreated Schools	22.8	14.1	55.6	65.6	75.2

Notes: see Table 4.

Table 7. Weighted OLS estimates of the Effects of External Monitoring on the Heterogeneity of Answers within each Class. Maths – V grade. Dependent Variable: Average Herfindhal Index within Each Class.

	(1) Italy	(2) North	(3) Centre	(4) South
Direct Effect x 100	4.35***	1.46***	2.99***	8.00***
	(0.37)	(0.38)	(0.68)	(0.77)
Indirect Effect x 100	1.08***	0.86***	0.95**	1.50***
	(0.26)	(0.27)	(0.48)	(0.57)
Total Effect x 100	5.43***	2.32***	3.94***	9.50***
	(0.339)	(0.350)	(0.630)	(0.692)
Observations	27,325	11,541	4,886	10,898
R-squared	0.94	0.98	0.96	0.88
Covariates	Yes	Yes	Yes	Yes
Mean - Untreated Schools	57.3	61.8	60.1	51.4

Notes: see Table 4.

Appendix

1) Tables

Table A.1. Weighted OLS estimates of the Effects of External Monitoring. Italian – V grade.
Dependent variable: Percentage of Correct Answers in the Class.

	(1) Italy	(2) North	(3) Centre	(4) South
Direct Effect x 100	-2.61***	-1.03***	-2.17***	-4.39***
	(0.21)	(0.23)	(0.43)	(0.42)
Indirect Effect x 100	-0.67***	-0.38**	-0.81***	-0.99***
	(0.15)	(0.17)	(0.28)	(0.31)
Total Effect x 100	-3.28***	-1.41***	-2.98***	-5.37***
	(0.194)	(0.205)	(0.393)	(0.381)
Observations	27,369	11,557	4,894	10,918
R-squared	0.98	0.99	0.99	0.97
Covariates	Yes	Yes	Yes	Yes
Mean - Untreated Schools	70.0	70.2	70.1	69.7

Notes: see Table 4.

Table A.2. Weighted OLS estimates of the Effects of External Monitoring. Maths – II grade.
Dependent variable: Percentage of Correct Answers in the Class.

	(1) Italy	(2) North	(3) Centre	(4) South
Direct Effect x 100	-4.20***	-1.57***	-3.09***	-7.50***
	(0.30)	(0.34)	(0.55)	(0.61)
Indirect Effect x 100	-1.22***	-0.91***	-1.37***	-1.53***
	(0.22)	(0.25)	(0.42)	(0.47)
Total Effect x 100	-5.42***	-2.48***	-4.47***	-9.03***
	(0.274)	(0.312)	(0.501)	(0.547)
Observations	27,012	11,724	4,905	10,383
R-squared	0.96	0.98	0.97	0.94
Covariates	Yes	Yes	Yes	Yes
Mean - Untreated Schools	62.9	59.9	61.8	66.7

Notes: see Table 4.

Table A.3. Weighted OLS estimates of the Effects of External Monitoring. Italian – II grade. Dependent variable: Percentage of Correct Answers in the Class.

	(1) Italy	(2) North	(3) Centre	(4) South
Direct Effect x 100	-3.40***	-1.36***	-2.17***	-6.21***
	(0.28)	(0.34)	(0.55)	(0.54)
Indirect Effect x 100	-1.04***	-0.71***	-1.25***	-1.33***
	(0.20)	(0.24)	(0.39)	(0.41)
Total Effect x 100	-4.44***	-2.07***	-3.42***	-7.54***
	(0.250)	(0.301)	(0.501)	(0.481)
Observations	27,025	11,721	4,911	10,393
R-squared	0.971	0.98	0.98	0.96
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. Maths – V grade. Dependent variable: Percentage of Correct Answers in the Class.

	(1) Italy	(2) North	(3) Centre	(4) South
Direct Effect x 100	-2.74***	-0.95***	-2.27***	-4.74***
	(0.27)	(0.30)	(0.51)	(0.53)
Indirect Effect x 100	-0.79***	-0.76***	-0.73**	-0.96**
	(0.19)	(0.20)	(0.34)	(0.39)
Total Effect x 100	-3.53***	-1.71***	-3.00***	-5.70***
	(0.242)	(0.271)	(0.472)	(0.469)
Observations	27,325	11,541	4,886	10,898
Covariates	Yes	Yes	Yes	Yes
Mean - Untreated Schools	65.1	63.9	64.0	66.8

Notes: see Table 4.

Table A.5. Weighted OLS estimates of the Effects of External Monitoring. Maths – V grade.
Dependent variable: Percentage of Correct Answers in the Class. Finite Population Correction.

	(1) Italy	(2) North	(3) Centre	(4) South
Direct Effect x 100	-2.89***	-1.08***	-2.35***	-5.05***
	(0.09)	(0.10)	(0.17)	(0.17)
Indirect Effect x 100	-0.83***	-0.71***	-0.70***	-1.06***
	(0.06)	(0.07)	(0.11)	(0.12)
Total Effect x 100	-3.72***	-1.79***	-3.05***	-6.11***
	(0.0789)	(0.0880)	(0.155)	(0.155)
Observations	27325	11541	4886	10898
R-squared	0.97	0.99	0.98	0.96
Covariates	Yes	Yes	Yes	Yes
Mean - Untreated Schools	65.1	63.9	64.0	66.8

Notes: see Table 4.

Table A.6. Weighted OLS estimates of the Effects of External Monitoring. Maths – V grade.
Dependent variable: Percentage of Correct Answers in the Class. Without Covariates.

	(1) Italy	(2) North	(3) Centre	(4) South
Direct Effect x 100	-2.82***	-0.85**	-2.04***	-5.29***
	(0.29)	(0.33)	(0.55)	(0.58)
Indirect Effect x 100	-0.70***	-0.82***	-0.46	-0.70*
	(0.20)	(0.23)	(0.37)	(0.42)
Total Effect x 100	-3.52***	-1.68***	-2.50***	-5.99***
	(0.264)	(0.302)	(0.499)	(0.527)
Observations	27,325	11,541	4,886	10,898
R-squared	0.97	0.98	0.98	0.95
Covariates	No	No	No	No
Mean - Untreated Schools	65.1	63.9	64.0	66.8

Notes: see Table 4.

Table A.7. Weighted OLS estimates of the Effects of External Monitoring. Maths – V grade.
Dependent variable: Percentage Absent from the Test

	(1) Italy	(2) North	(3) Centre	(4) South
Direct Effect x 100	-0.53**	-0.50	-0.47	-0.55
	(0.25)	(0.41)	(0.47)	(0.41)
Indirect Effect x 100	-0.10	0.44	-0.44	-0.51
	(0.18)	(0.28)	(0.35)	(0.32)
Total Effect x 100	-0.626***	-0.0586	-0.913**	-1.06***
	(0.219)	(0.356)	(0.424)	(0.359)
Observations	27,325	11,541	4,886	10,898
R-squared	0.5634	0.5756	0.6127	0.5331
Covariates	No	No	No	No
Mean - Untreated Schools	11.0	10.4	11.7	11.4

Notes: see Table 4

References

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.

Bokhari, F. A. S. and Schneider, H., 2011. School Accountability Laws and the Consumption of Psychostimulants. *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.

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

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.

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.