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

Marco Bertoni[**]
PhD Candidate in Economics
University of Padova and CEP-LSE

Giorgio Brunello
Professor of Economics
University of Padova, IZA and CESifo

Lorenzo Rocco
Assistant Professor of Economics
University of Padova

Abstract

We use a natural experiment to shows that, in the context of standardized educational tests, the presence of an external examiner has both a direct and an indirect negative effect on the performance of tested classes. 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 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 distraction from having a stranger in the class.

[*] The authors are grateful to Thomas Breda, Daniele Checchi, Ifty Hussain, Edwin Leuven, Marco Manacorda, Guy Michaels, Hessel Oosterbeek, Steve Pischke, Olmo Silva and the audiences at the 2012 LSE-CEP Annual Conference and at the HECER Economics of Education Summer Meeting for comments and suggestions. We also thank Patrizia Falzetti and Paolo Sestito (INVALSI) for help with data collection. Financial support by the Ministry of Italian Universities (PRIN contract n. 2009MAATFS_002) is gratefully acknowledged. All errors are our own.

[**] Corresponding author. Address: Centre for Economic Performance – London School of Economics and Political Science. Houghton Street, London WC2A 2AE. E-mail: m.bertoni@lse.ac.uk.

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. 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 observed students’ outcomes.

In this paper, we take a different approach and start from the observation that strategic manipulation by teachers, students and administrators can be substantially reduced if an external examiner is actively engaged in monitoring entirely or in part the test process. We use a 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 argue that the gap is a measure of cheating in un-monitored classes.

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-

[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. Yet 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.

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%).

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 distraction from having a stranger in the class. 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.

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 cheating by looking at the direct and indirect effects of having external examiners monitor teachers and students during the test. The presence of indirect treatment effects has been already uncovered in 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 maths were introduced in Italian primary schools in 2008[4] to evaluate school productivity (in terms of value added). 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 both because of the widespread expectation that they might be used at some point to evaluate teachers and because the school reputation was at stake. While results of the tests are not publicly available, schools and their principals can access the results of their students and decide to make them public, which creates 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, 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 2000 randomly selected classes - out of a population of about 30000 - the test was administered in the presence of an external examiner[5], 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 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

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

[5] External examiners were selected by INVALSI and the Regional Schooling Authorities mainly among retired teachers and active teachers employed in non-primary schools.

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 a two-stages sampling scheme, stratified by region[6]. 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. In the second stage, one or two classes within each treated school were selected by simple random sampling[7]. 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.

Table 1. Total and Sampled Number of Schools, Classes and Students. INVALSI SNV Test 2009/2010

	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

We have access to the data containing the individual answers to the questions of the test given by the second and fifth grade primary school students who took the INVALSI tests in 2009/2010. For these students, we also have data on individual marks in Italian and maths during the semester before the test was taken and on parental background, both provided by school offices. Exclusively for fifth graders, INVALSI used a student questionnaire to collect additional data on parental background and the feelings and motivation during the tests. We obtained from INVALSI

[6] 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 the following analysis. Other data management operations are described in the Appendix.

[7] The precise number of sampled classes depends on school size.

additional information on school and class characteristics, including the number of students enrolled in each class and in each school for each tested grade, whether the school is public or private, the proportion of tenured teachers in each school and, only for fifth grade students, an index of individual economic, social and cultural status (ESCS) [8].

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)[9]. 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:

(1)

[8] Available variables for all students include also: 1) at the school level: whether the school offers 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 individual level: gender, place of birth, citizenship, attendance of pre-primary school, age, employment, education and nationality of 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.

[9] Potential outcomes are stochastic variables, with realization $(Y_{00},Y_{01},Y_{11})_{ij}$ for class i in school j.

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 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.

This procedure implies that we have conditional randomization, meaning 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 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.

Let *RD* be a vector of regional dummies, *RS$_j$* a vector of regional dummies interacted with the size of school *j*, *RC$_j$* a vector of regional dummies interacted with the number of classes in a given grade in school *j* and define the vector *R* as *R = [RD, RS, RC]*. Conditional randomization in each grade implies that

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

When (2) holds, the effects of external monitoring are given by

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

(4)

$$E[Y_{i\,j}| C_i = 1, S_j = 1, R] - E[Y_{i\,j}| C_i = 0, S_j = 0, R] = E[Y_{1\,1} - Y_{0\,0}| R] \tag{5}$$

Let X be a vector of covariates at the school, class and individual level. Table 2 shows the means and standard deviations of these covariates (Panel A) as well as of other covariates used in Section 7 (Panel B) for the sample of fifth graders attending the maths test. We test for successful randomization by checking whether the variables in vector X are balanced in the treatment and control sub-samples. Although we have data for second and fifth graders, we focus hereinafter on the latter to save space. Some results for second graders are shown in the Appendix. To test for balancing we consider both differences between treated and untreated schools and differences between treated and untreated classes within treated schools. For each covariate in vector X we run

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

for between-schools and

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

for within-school randomization, 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 significance level of the test of the hypothesis H_0: β=0 for each covariate in X. 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 virtually 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.

We evaluate the effects of external monitoring on class performance in the (maths) test by estimating

(8)

where the dependent variable is the percentage of correct answers in the class and the standard errors are robust and weighted with the number of students in the class. The direct, indirect and overall effect of external monitoring are given by

a) direct effect:

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

b) indirect effect:

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

c) total effect:

(11)

Table 2. Mean and Standard Deviation of Covariates - Maths 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
Maths 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

	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

(continues)

(continued)

	Mean	St Dev
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
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

	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	0.03	0.02
Average turnout at referenda (%)	80.28	8.37
Provincial unempl. rate (2009)	7.95	3.69
Provincial per capita GDP (2009)	23.84	5.60

Notes: The table reports the mean and standard deviation of the covariates included in the regressions. These statistics are based on individual, school and class level data. Except for the number of students enrolled in each class, all Panel A variables have been categorized as dummy variables. Class size refers to the number of students attending the test. Blood donations are the number of blood bags per million of inhabitants in the province. Per capita GDP is measured in thousands of euro. See the Appendix for further details.

Table 3 - Balancing Tests. First (between schools) and second stage (within school) randomization. - Maths tests - V graders.

Panel A

	Between schools	Within school
Private school (%)	0.003	
Full time schedule (%)	0.015	0.011
Number of students enrolled in class	0.039	0.425***
Gender		
Missing (%)	0.005***	0.021***
Male (%)	-0.003	-0.007*
Place of birth		
Missing (%)	-0.014***	-0.025***
Italy (%)	0.014***	0.026***
Citizenship		
Missing (%)	-0.008***	-0.011***
Italian (%)	0.008**	0.010***
First generation foreigner (%)	-0.001	0.000
Second generation foreigner (%)	0.002	0.002
Pre-primary school		
Missing (%)	-0.019**	-0.008
Yes (%)	0.018**	0.008
Age		
Missing (%)	0.006***	0.019***
Older than regular (%)	0.000	0.001
Regular (%)	-0.008***	-0.016***
Younger than regular (%)	0.001	-0.003
Maths grade		
Missing (%)	-0.023***	-0.008
1-4 (%)	0.000	-0.001**
5 (%)	0.001	0.000
6-7 (%)	0.010**	0.007
8-10 (%)	0.012**	0.002
Italian grade		
Missing (%)	-0.023***	-0.006
1-4 (%)	0.000	0.000
5 (%)	0.000	0.001
6-7 (%)	0.007	0.004
8-10 (%)	0.015**	0.002

	Between schools	Within school
Mother occupation		
Missing (%)	-0.010	-0.023***
Unemployed or retired (%)	0.005	0.010*
Employee (%)	0.002	0.004
Entrepreneur (%)	0.001	0.005**
Middle manager (%)	0.002	0.003
Father occupation		
Missing (%)	-0.010	-0.023***
Unemployed or retired (%)	-0.001	0.000
Employee (%)	0.000	0.014**
Entrepreneur (%)	0.010**	0.007
Middle manager (%)	0.002	0.003
Mother education		
Missing (%)	-0.011	-0.029***
Primary (%)	0.006	0.018***
Secondary (%)	0.002	0.010**
Tertiary (%)	0.002	0.001
Father education		
Missing (%)	-0.011	-0.026***
Primary (%)	0.011	0.016**
Secondary (%)	-0.001	0.008*
Tertiary (%)	0.001	0.002
Mother nationality		
Missing (%)	-0.015**	-0.012**
Italian (%)	0.013**	0.011*
Father nationality		
Missing (%)	-0.015**	-0.011**
Italian (%)	0.013**	0.008

Panel B

	Between schools	Within school
Has own bedroom		
Missing (%)	-0.007***	-0.008***
Yes (%)	-0.002	0.004
Has internet access		
Missing (%)	-0.006**	-0.007***
Yes (%)	0.006*	0.007**
Has an encyclopedia		
Missing (%)	-0.006**	-0.006***
Yes (%)	0.008*	0.015***
Has own desk		
Missing (%)	-0.006**	-0.006***
Yes (%)	0.004	0.007**
Has a PC		
Missing (%)	-0.006**	-0.006***
Yes (%)	0.008***	0.010***
Has a place for homework		
Missing (%)	-0.006**	-0.006***
Yes (%)	0.007***	0.006*
Number of books at home		
Missing (%)	-0.008***	-0.007***
0-10 (%)	-0.001	0.000
11-25 (%)	-0.003	0.000
26-100 (%)	0.001	0.005
101-200 (%)	0.005**	0.003
>200 (%)	0.006**	0.000

	Between schools	Within school
Number of siblings		
Missing (%)	-0.007***	-0.007***
0 (%)	0.000	0.000
1 (%)	0.004	0.006*
2 (%)	0.002	0.001
3 (%)	0.001	0.000
4 or more (%)	0.000	0.000
Lives with		
Missing (%)	-0.008***	-0.008***
Both parents (%)	0.008***	0.006*
One parent only (%)	0.000	0.001
Both parents alternatively (%)	0.000	0.002
Others (%)	0.000	0.000
Language spoken at home		
Missing (%)	-0.008***	-0.008***
Italian (%)	0.003	0.006
Dialect (%)	0.004	0.002
Other (%)	0.001	0.000
Help with homework		
Missing	-0.007***	-0.005*
No homework (%)	-0.001**	-0.001***
No help needed (%)	-0.002	0.006*
Parents (%)	0.005*	0.000
Siblings (%)	0.003**	-0.001
Private teacher (%)	-0.001	0.001
Other (%)	0.002**	-0.001
No one (%)	0.000	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 (5) and (6). 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. The variables in Panel A are available for students in both grades. The variables in Panel B are available only for fifth grade students. All regressions are weighted by the number of students in each school or class. Robust standard errors. One, two and three stars for statistical significance at the 10, 5 and 1 percent level.

5. Results

Table 4 presents our estimates of (8) for fifth graders and the maths test[10]. The first column in the table considers all Italian regions, and the remaining columns show 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 8.5% 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%).

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

	(1)	(2)	(3)	(4)
	Italy	North	Centre	South
Direct Effect x 100	-2.79***	-0.99***	-2.27***	-4.92***
	(0. 27)	(0.30)	(0.52)	(0.56)
Indirect Effect x 100	-0. 81***	-0.70***	-0.73**	-1.04***
	(0.19)	(0.20)	(0.34)	(0.40)
Overall Effect x 100	-3.59***	-1.69***	-2.99***	-5.96***
	(0.250)	(0.276)	(0.484)	(0.501)
Observations	27,325	11,541	4,886	10,898
R-squared	0.15	0.20	0.15	0.14
Covariates	Yes	Yes	Yes	Yes
Mean - Untreated Schools x 100	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. Estimates are weighted by class size. Robust standard errors in parentheses. One, two and three stars for statistical significance at the 10, 5 and 1 percent level of confidence.

Why are test results worse in classes with the external examiner? 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

[10] Results for Italian and second graders are qualitatively similar and are shown in the Appendix.

examiner engage in outright cheating[11]. We believe that the second one is the explanation, for the following three 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 students during the test, which included agreeing or not 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 equation (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 or nervousness. Quite the opposite, there is some mild evidence that students in these classes were less nervous and calmer during the test.

[11] We implicitly assume that external examiners have no incentive to cheat and to 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).

Table 5. Weighted OLS estimates of the Effects of External Monitoring on Student Psychological Conditions. Maths tests – 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	0.64
	(0.43)	(0.29)	(0.41)	(0.41)
Indirect Effect x 100	0.25	0.01	0.36	-0.01
	(0.28)	(0.19)	(0.26)	(0.26)
Overall Effect x 100	0.50	-0.90***	0.28	0.63*
	(0.39)	(0.272)	(0.38)	(0.37)
Observations	27,141	27,142	27,141	27,140
R-squared	0.08	0.11	0.10	0.08
Covariates	Yes	Yes	Yes	Yes
Mean - Untreated Schools x 100	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. Students with missing answers have been dropped from the estimation sample (about 2 percent of the total). The estimates refer to the entire country.

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 can reduce average performance, it is not obvious that it reduces its variability. Table 6 shows 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.

The table focuses on the results of the maths test taken by fifth graders in Southern Italy, where the gap in the percentage of correct answers between monitored and un-monitored classes is largest. We find that in classes with the external examiner the standard deviation and the coefficient of variation of results are about 10% and 20% 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. When compared with students in untreated schools, having an external examiner reduces the score of these students by a 12.9% (-7.62/59.0). This effect is stronger for second grade students, where it reaches a striking 18.7%.

Table 6. Weighted OLS estimates of 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. Southern Italy - Maths tests – V grade.

	(1) Standard Deviation	(2) Coefficient of Variation	(3) First quartile	(4) Second quartile	(5) Third quartile
Direct Effect x 100			-	-	-
	1.34***	3.82***	6.72***	5.57***	4.18***
	(0.17)	(0.39)	(0.66)	(0.62)	(0.56)
Indirect Effect x 100	0.13	0.47*	-0.90*	-0.91**	-0.93**
	(0.12)	(0.26)	(0.48)	(0.44)	(0.39)
Overall Effect x 100			-	-	-
	1.47***	4.30***	7.62***	6.48***	5.10***
	(0.16)	(0.37)	(0.59)	(0.56)	(0.51)
Observations	10,898	10,898	10,898	10,898	10,898
R-squared	0.10	0.11	0.10	0.10	0.10
Covariates	Yes	Yes	Yes	Yes	Yes
Mean - Untreated Schools x 100	12.2	20.1	59.0	67.6	75.4

Notes: see Table 4.

Last but not least, 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[12]

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

[12] See INVALSI (2010b).

where s_a is the within-class share of students who chose answer "a" in the set A of possible answers[13]. The variable 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 heterogeneity of answers in the class 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 estimates of equation (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)).

Table 7. Weighted OLS estimates of the Effects of External Monitoring on the Heterogeneity of Answers in each Class. Maths tests – V grade. Dependent Variable: Average Herfindhal Index in Each Class x 100.

	(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)
Overall Effect x 100	5.43***	2.32***	3.94***	9.50***
	(0.34)	(0.35)	(0.63)	(0.69)
Observations	27,325	11,541	4,886	10,898
R-squared	0.19	0.18	0.13	0.14
Covariates	Yes	Yes	Yes	Yes
Mean - Untreated Schools x 100	57.3	61.8	60.1	51.4

Notes: see Table 4.

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

[13] We treat missing values as a separate category. Answers to open questions with a univocally correct answer were coded as correct, incorrect or missing.

activities. This interpretation relies on irrational 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 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 the Appendix 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 test in each class. Again, differences in behaviour across the three groups are minimal (see Table A.7 in the Appendix).

7. Extensions

In this final section we ask whether the effects of having an external examiner vary with a) class size; b) whether the school is public or private; c) the percentage of tenured teachers in the school; d) an indicator of the average parental background of the students in the class; e) measures of social capital in the province where the school is located. Descriptive statistics for these variables are shown in Table 2.

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[14]. 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 the school type – public or private – does not influence in a statistically significant way the overall effect of external supervision. In contrast, column (3) shows that 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 to standardized tests and may therefore be more likely to engage in cheating and sabotaging.

Column (4) looks at 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

[14] In this and in the following regressions we include the interacted variable as an additional control.

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[15].

Table 8. Heterogeneous Effects of External Monitoring. Maths tests – V grade. Dependent variable: Percentage of Correct Answers in the Class. Interactions of direct, indirect and overall effects with class size, school type, % tenured teachers and average ESCS.

	(1)	(2)	(3)	(4)
	Class size	Private School	% Tenured Teachers	ESCS
Direct Effect x 100	-3.41***	-2.79***	-1.34***	-2.65***
	(0.43)	(0.275)	(0.33)	(0.35)
Direct Effect – Interaction x 100	0.98*	0.132	-2.98***	-0.15
	(0.55)	(2.70)	(0.55)	(0.55)
Indirect Effect x 100	-0.94***	-0.815***	-0.66***	-0.67***
	(0.28)	(0.190)	(0.22)	(0.23)
Indirect Effect – Interaction x 100	0.22	0.498	-0.19	-0.30
	(0.35)	(2.17)	(0.37)	(0.36)
Overall Effect x 100	-4.35***	-3.60***	-2.00***	-3.32***
	(0.003)	(0.30)	(0.29)	(0.32)
Overall Effect – Interaction x 100	1.20**	0.60	-3.17***	-0.45
	(0.48)	(1.80)	(0.49)	(0.48)
Observations	27,325	27,325	26,313	27,323
R-squared	0.15	0.15	0.15	0.15
Covariates	Yes	Yes	Yes	Yes
Mean - Untreated Schools x 100	65.1	65.1	64.9	65.1

Notes: In each regression the interacted variable enters also separately. Class size, proportion of tenured teachers and class ESCS are coded as dummy variables taking value one when above the median and zero when below (for class ESCS, the dummy takes 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. Robust standard errors in parentheses. One, two and three stars for statistical significance at the 10, 5 and 1 percent level of confidence.

Finally, 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 each province is endowed with[16].

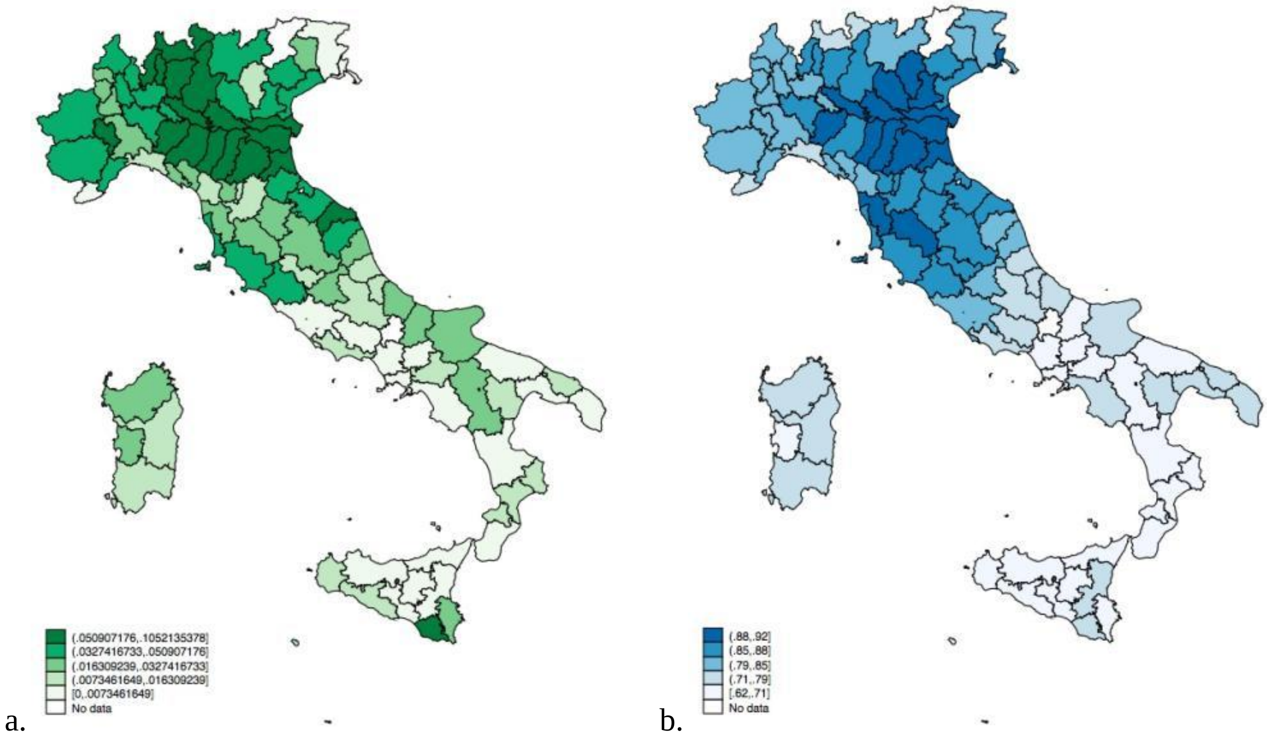
[15] One possible explanation is that not only teachers, but also external examiners may be induced to engage in compensatory behaviour.

[16] 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 determinant of financial development, and Nannicini et al. (2012) study the impact of social capital on political accountability. Finally, Ichino and Maggi

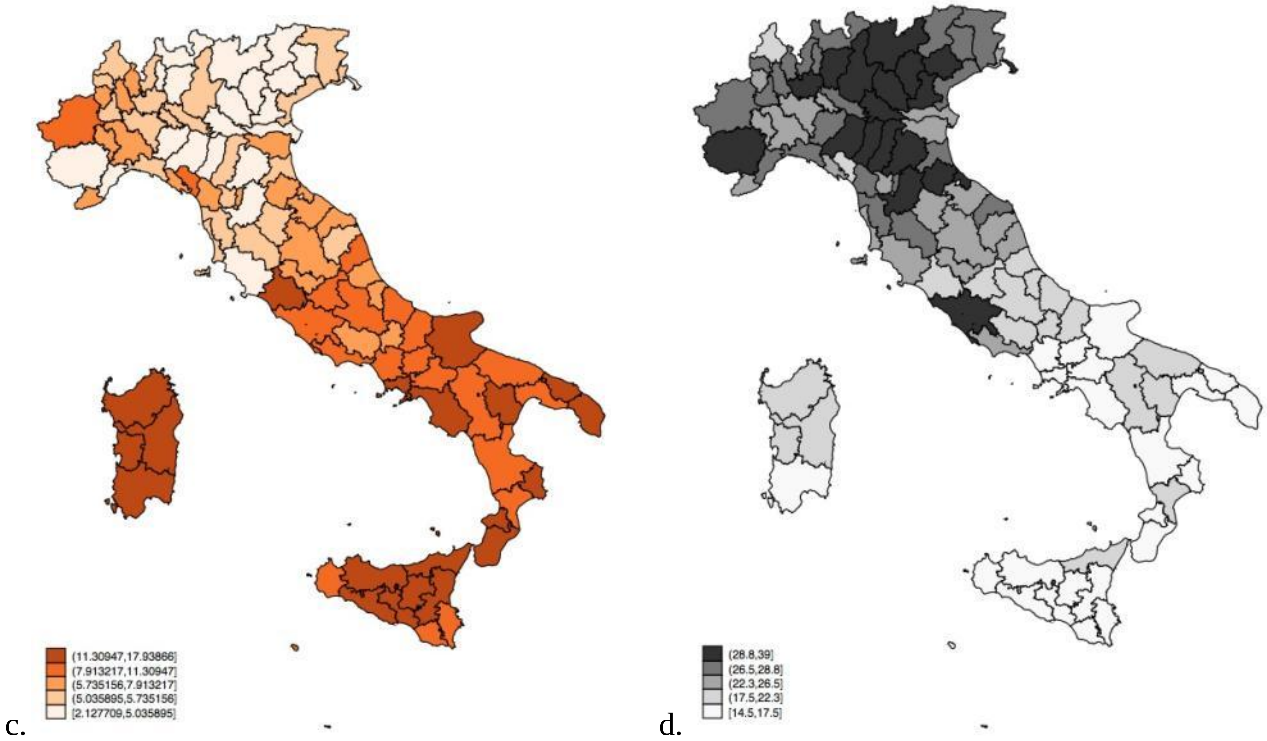
Guiso, Sapienza and Zingales (2010) define social capital as civic capital, or as "...those persistent and shared beliefs and values that help a group overcome free rider outcomes..."(p.8). They report higher levels of social capital in Northern and Central Italy compared to the South.

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 million 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.

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



(2000) measure civicness in terms of shirking behaviour in the workplace and document large shirking differentials between Northern and Southern Italy.



Notes: Panel a) number of blood donations per million of inhabitants in capita 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.

Results are shown in Table 9. In all regressions, both social capital and the macroeconomic variables are re-scaled to vary between 0 and 1. Column (1) of the table reports the estimates of the baseline model in the sub-sample of provinces for which data on social capital are available. Results 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). We find that both the direct and the overall effect are smaller in schools located in provinces with a higher social capital. This qualitative result remains when we add to the regressions the interactions with provincial unemployment and GDP per capita(columns (3) and (5)), although the effect of social capital is smaller.

Table 9. Weighted OLS estimates of the Effects of External Monitoring, with Interactions between Social Capital and External Monitoring. Maths tests – V grade. Dependent variable: Percentage of Correct Answers in the Class

	(1)	(2)	(3)	(4)	(5)
	Baseline	Interacted with Blood Donations	Interacted with Blood Donations and Macro Variables	Interacted with Turnover at Referenda	Interacted with Turnover at Referenda & Macro Variables
Direct Effect x 100	-2.78***	-4.76***	-3.47**	-7.13***	-4.73**
	(0.28)	(0.48)	(1.47)	(0.84)	(1.99)
Interacted Direct Effect x 100		8.54***	4.33***	7.39***	4.20**
		(1.36)	(1.46)	(1.17)	(2.00)
Indirect Effect x 100	-0.82***	-0.68**	-1.34	-1.05*	-1.27
	(0.19)	(0.33)	(1.04)	(0.59)	(1.44)
Interacted Indirect Effect x 100		-0.65	-1.35	0.42	-0.55
		(0.90)	(1.00)	(0.81)	(1.41)
Overall effect x 100	-3.60***	-5.44***	-4.82***	-8.19***	-5.99***
	(0.25)	(0.43)	(1.30)	(0.73)	(1.78)
Interacted Overall Effect x 100		7.89***	2.99**	7.81***	3.65**
		(1.19)	(1.31)	(1.01)	(1.81)
Observations	27,178	27,178	27,178	27,178	27,178
R-squared	0.15	0.15	0.15	0.15	0.15
Covariates	Yes	Yes	Yes	Yes	Yes
Mean - Untreated Schools x 100	65.1	65.1	65.1	65.1	65.1

Notes: Interacted effects refer to the interactions between direct, indirect and overall effects and the measure of social capital listed at the top of each column. Each measure enters also as an independent covariate in each regression. Social capital measures are not available for the provinces of Belluno and Isernia (147 classes). Per capita GDP and unemployment rates in the province enter in columns (3) and (5) both as independent covariates and interacted with each effect. 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. Robust standard errors in parentheses. One, two and three stars for statistical significance at the 10, 5 and 1 percent level of confidence.

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 negative effects on the measured performance of tested classes and schools. While the direct effect is on the monitored class, the indirect effect is on the un-monitored classes of the same school.

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 ones.

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.

Appendix

1) Tables

Table A.1. Weighted OLS estimates of the Effects of External Monitoring. Italian tests – 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)
Overall 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	0.19	0.28	0.22	0.17
Covariates	Yes	Yes	Yes	Yes
Mean - Untreated Schools x 100	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 tests – 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)
Overall 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	0.11	0.08	0.09	0.08
Covariates	Yes	Yes	Yes	Yes
Mean - Untreated Schools x 100	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 tests – 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)
Overall 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	0.13	0.20	0.16	0.10
Covariates	Yes	Yes	Yes	Yes
Mean - Untreated Schools x 100	65.9	65.0	66.2	66.7

Notes: see Table 4.

Table A.4. GLM estimates of the Effects of External Monitoring. Maths tests – 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***	-0.97***	-2.25***	-4.73***
	(0.27)	(0.30)	(0.51)	(0.53)
Indirect Effect x 100	-0.80***	-0.70***	-0.72**	-1.04**
	(0.19)	(0.20)	(0.34)	(0.4)
Overall Effect x 100	-3.54***	-1.67***	-2.97***	-5.77***
	(0.242)	(0.271)	(0.472)	(0.472)
Observations	27,325	11,541	4,886	10,898
Covariates	Yes	Yes	Yes	Yes
Mean - Untreated Schools x 100	65.1	63.9	64.0	66.8

Notes: see Table 4.

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

	(1)	(2)	(3)	(4)
	Italy	North	Centre	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)
Overall Effect x 100	-3.72***	-1.79***	-3.05***	-6.11***
	(0.0789)	(0.0880)	(0.155)	(0.155)
Observations	27,325	11,541	4,886	10,898
R-squared	0.15	0.19	0.15	0.15
Covariates	Yes	Yes	Yes	Yes
Mean - Untreated Schools x 100	65.1	63.9	64.0	66.8

Notes: Population size: 30.310 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. Robust standard errors in parentheses. One, two and three stars for statistical significance at the 10, 5 and 1 percent level of confidence.

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

	(1)	(2)	(3)	(4)
	Italy	North	Centre	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)
Overall 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	0.03	0.01	0.01	0.03
Covariates	No	No	No	No
Mean - Untreated Schools x 100	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 tests – 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)
Overall Effect x 100	-0.63***	-0.06	-0.91**	-1.06***
	(0.22)	(0.36)	(0.42)	(0.36)
Observations	27,325	11,541	4,886	10,898
R-squared	0.03	0.02	0.03	0.03
Covariates	No	No	No	No
Mean - Untreated Schools x 100	11.0	10.4	11.7	11.4

Notes: see Table 4.

2) From the initial dataset to the final sample

Our data are drawn from the 2009/2010 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 data handling.

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 maths 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.

3) 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

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.

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: Priceton University Press.