Does Monitoring Deter Future Cheating?
The Case of External Examiners in Italian Schools^

Marco Bertoni (Padova and IZA)*

Giorgio Brunello (Padova and IZA)

Marco Alberto De Benedetto (Messina)

Maria De Paola (Calabria and IZA)

Abstract

We use the repeated random assignment of external examiners to school institutes in Italy to investigate whether the effect of external monitoring on test score manipulation persists over time. We find that this effect is still present in the tests taken one year after exposure to the examiners, and is stronger for open-ended questions, for small school institutes, and for institutes located in the northern and central regions of the country. In the second year after exposure, however, this effect disappears, suggesting that persistence is short lived. We discuss teacher learning, reputational concerns, peer pressure and compliance with local identity and conclude that the last is the most convincing mechanism explaining our results.

Keywords: education; testing; external monitoring; short-run effects.

JEL codes: H52; I2.

---

^ The authors are grateful to Lorenzo Rocco and to the audience in Padova for comments and suggestions. Bertoni and Brunello acknowledge funding from a CARIPARO foundation "Starting Grant". The usual disclaimer applies.

* Marco Bertoni (corresponding author): Department of Economics and Management "Marco Fanno" - University of Padova. Via del Santo 33, 35123 Padova – Italy. Email: marco.bertoni@unipd.it. Telephone: +39-049-8274002. Giorgio Brunello: giorgio.brunello@unipd.it. Marco Alberto De Benedetto: mdebenedetto@unime.it. Maria De Paola: m.depaola@unical.it.

Introduction

According to Becker, 1968, individuals compare expected costs and benefits when choosing whether to abide by the law. Appropriate incentives for law abidance require that individuals be monitored and sanctioned when they deviate from the prescribed behavior. The choice of the optimal level of enforcement involves the comparison of monitoring and punishment costs with the benefits deriving from incapacitation, deterrence and re-education.[1] While the incapacitation effect is temporary – and disappears when individuals stop being monitored – deterrence and re-education have longer-lasting consequences.

Whether individuals who have been monitored (and eventually sanctioned) are more or less inclined to offend than other individuals is a central topic in the economics of law enforcement. Theoretically, agents can either reduce or increase their offending propensity, depending on whether they adjust upwards or downwards the probability of future punishment.

Empirical investigations have mainly focused on tax compliance, with mixed empirical results. Consistent with the idea that taxpayers believe that they are unlikely to be audited immediately following audits, DeBacker *et al.*, 2015, show that tax audits reduce tax compliance immediately after the audit.[2] Conversely, Beer *et al.*, 2019, find that taxpayers receiving an additional tax assessment report higher taxable income in the first year after the audit, giving support to the assumption that taxpayers update their perceived audit probability in a Bayesian manner. Interestingly, a positive impact is also found by DeBacker *et al.*, 2018, who exploit a setting where audits are random and used for research

[1] The economic literature has attempted to empirically assess these benefits by considering different types of crimes (see for instance Levitt, 1996; Buonanno and Raphael, 2013; Barbarino and Mastrobuoni, 2013; Drago *et al.*, 2009).

[2] Similar effects are found by Mittone, 2006, and Mittone *et al.*, 2017, using laboratory experiments.

purposes, and speculate that the uncovered increase in reported taxable income might be due to the fact that individuals misunderstood the nature of the audit.

In addition to affecting the perceived probability of being punished, monitoring can also change the perceived costs and benefits of illicit behaviour. For instance, the enhanced engagement with the enforcement authority involved in monitoring may operate as a nudge, making the value of honesty and good behavior more salient – at least temporarily.

In this paper, we focus on the impact of past monitoring on current score manipulation in the standardized tests carried out by Italian schools and show that rule enforcement and monitoring can affect subsequent compliance behavior even in a context where monitoring is not accompanied by credible sanctioning.[3]

The manipulation of test outcomes by school principals, teachers or students – by copying or by changing answers in the transcription process – may be considered as a minor violation of existing rules. Yet it could seriously hamper the attractiveness of standardized assessment systems, which are designed to compare the performance of students in different schools and geographical areas, invalidating the entire accountability system[4] and leading to uncorrect decisions both at the individual and at the aggregate level. For instance, students may fail to receive the remedial instruction they need, or could be awarded grants that they do not deserve, and governments may overlook the necessity to intervene and improve school quality.

[3] Lucifora and Tonello, 2016, evaluate the effect of the INVALSI sanctioning policy based on a "fame and shame" mechanism -consisting in a re-assessment of the test scores - on a cheating propensity indicator built at the class level, and find no significant effect. The ineffectiveness of the sanctioning scheme is due to the fact that sanctions are not embedded in a proper school accountability system. In fact, schools are not obliged to make their scores public once detected as potential cheaters.

[4] See Battistin *et al.*, 2017, for evidence on the effects of manipulation on regional score rankings in Italy.

Although incentives to cheat are clearly in place in high stakes accountability systems, in which test results have important consequences for schools, teachers and students (Ahn and Vigdor, 2014), illicit actions that result in score manipulation are widespread also in low stakes systems. In fact, cheating scandals have emerged both in the US and Sweden, where tests are high stakes, and in Italy, where they are mostly low stakes (see Diamond and Persson, 2016, and Bertoni *et al.*, 2013, as well as Battistin, 2016, for a recent review).

A possible remedy to score manipulation is strict monitoring of the whole testing process by external examiners. Since monitoring is costly, it is important to quantify its benefits in terms of lower score manipulation. Focusing on the Italian experience, Lucifora and Tonello, 2016, and Angrist *et al.*, 2017, document that external monitors reduce score manipulation in classes where they are present. Bertoni *et al.*, 2013, show that external monitoring not only negatively affects cheating in directly monitored classes, but has also positive spillover effects on other classes (not directly monitored) in the same school.

Because of the incapacitation effect, monitoring should induce a contemporaneous reduction of score manipulation, without necessarily affecting future behavior. Consistent with this view, researchers have focused so far only on the impact of external monitors on current test performance, ignoring the possibility that future test performance can also be affected, with potential implications for the expected benefits of monitoring.

Current monitoring could have persistent effects for several reasons. For instance, teachers may learn from examiners how to correct with diligence open-ended questions and fill in the answer sheets appropriately. Alternatively, large fluctuations over time in test scores may foster suspicion that manipulation is afoot (see Jacob and Levitt, 2003), and increase the likelihood of being sanctioned by stakeholders or the central authority (see Lucifora and Tonello, 2016, for Italian evidence). Teachers may also dislike sharp variations in test

scores from one year to another, as these could make some of them look bad and possibly receive poorer evaluations from the principal.

Last but not least, as argued above, the interaction with the monitoring authority may act as a nudge and enhance the salience of honesty and good behavior. However, in accordance with the work by Akerlof and Kranton (2000) on identity and economic behaviour, we may expect that, if this mechanism is in place, its effects should persist over time only in those areas where individual identity values honesty and law abidance.

In this paper, we investigate the effects of external monitoring in school *i* and years *t-2*, *t-1* and *t* on test scores in the same school in year *t*, using data on standardized math and literacy tests for the universe of 5^{th} graders in Italian primary schools. Our research design exploits the fact that, every year, external examiners in Italy are randomly allocated to groups of schools (called school institutes).

These examiners have the task of vigilating the entire test administration process, both by monitoring students taking the test and by supporting school staff in transcribing and transmitting the scores to the government agency in charge of test management. While our focus is on the low stakes tests taken by 5^{th} grade pupils, we also consider the tests taken by 8^{th} graders (lower secondary or middle schools), for whom tests are high stakes.

We find evidence of short-term persistency: external monitoring in the school reduces the average percentage of correct answers and an index of cheating propensity both in the current and in the following year. While the effect of lagged monitoring on the average percentage of correct answers is relatively small (-0.7 percent for math and -0.5 percent for literacy), the effect on cheating propensity is sizeable (-11.7 percent for math and -8.5 percent for literacy). After two years, however, the effects of having had an external examiner fade away completely.

We show that the effect of external monitoring in year *t-1* on test scores in year *t* is much larger in small school institutes, which typically have one single class in the grade, than in institutes with many classes in the grade. Even in these institutes, however, there is no statistically significant effect of external monitoring in year *t-2* on test scores in year *t*. We also investigate whether the impact of external monitoring varies across macro areas according to the level of social capital, measured by both voter turnout at referenda where voting is not mandatory and blood donation, and show that short-lived persistency is present in local areas with high social capital – in the Northern and Central regions of Italy – and absent in the South, where social capital is much lower.

We discuss four candidate mechanisms: a) teacher learning; b) reputational concerns; c) peer pressure; d) compliance with local identity, which varies across areas in the value given to honesty and law abidance, and conclude that the last does the best job in explaining our results.[5]

Our findings have implications for the design of policies using external monitors to deter cheating in school tests. In the areas where an "honesty salience" effect is present and the effects of external monitoring persist into the next year (the North and Centre of Italy in our study), the frequency of interventions could be reduced (for instance every two years) – especially in smaller schools – freeing up scarce resources to intensify yearly monitoring in the areas where social capital is low (the South of Italy). By so doing, the reduction in overall manipulation would be higher.

The paper is organized as follows. Section 1 describes the institutional background, Section 2 looks at the data and Section 3 introduces the empirical approach. The baseline results and findings by school institute size and by area

[5] Although we focus mainly on short-term persistency, our research is related to the literature focusing on the persistency of public policies. Some examples include Baird *et al.*, 2016, on de-worming policies in Kenya, Bloom *et al.*, 2018, on management interventions in India, Chetty *et al.*, 2016, on the US Moving to Opportunity Project, and Heckman and Karapakula, 2019, on the Perry Preschool Project in Michigan. Another related study is Duflo *et al.*, 2013, that looks at third-party auditing and compliance with pollution regulation in India.

are presented in Sections 4 and 5. We discuss mechanisms in Section 6 and report our estimates when tests are high stakes in Section 7. Conclusions follow.

1. Institutional background

Education in Italy is compulsory from ages 6 to 16, and consists of four main stages: primary school (grades 1 to 5); lower secondary or middle (grades 6 to 8), upper secondary school (grades 9 to 12 or 13) and university.[6] In the compulsory stages, schools are generally grouped in school institutes sharing the principal and several administrative services.[7]

Since school year 2009/2010, all students attending the 2nd, 5th, 8th and 10th grade have to take standardized tests assessing literacy and math skills.[8] These tests, managed by the National Institute for the Evaluation of the Education System (INVALSI), a government agency placed under the control of the Ministry of Education, University and Research (MIUR), are generally low stakes, with the exception of the test taken in grade 8th, which contributes to the final exit grade and is therefore high stakes.

The results of these tests can be disclosed in aggregate form by school principals, who can share them with stakeholders.[9] Although schools cannot be closed down and principals and teachers cannot be fired as a consequence of low test performance, these results can be used by principals to bolster the school reputation and attract new and "better" students.

As shown by Quintano *et al*., 2009,[10] there is pervasive evidence of score manipulation, especially in Southern Italy. In an effort to obtain a snapshot of

[6] Before entering primary schools, pupils can attend daycare (age 0-2) and kindergarten (age 3-5), but these stages are not mandatory.
[7] A school institute can group together schools located in different municipalities and belonging to different stages of education.
[8] 6th graders were also involved until the school year 2013/14.
[9] Although school principals have also access to the individual data, only aggregate data (at the school or class level) can be made public.
[10] For further details see Bertoni *et al.*, 2013, Angrist *et al.*, 2017, Pereda-Fernandez, 2018, and the references therein.

the evolution of educational achievement across Italy that is not contaminated by score manipulation, INVALSI selects every year a sample of school institutes and classes where it enforces a strict protocol of monitoring. In those classes, the tests take place in the presence of an external examiner, usually chosen among retired teachers, who not only must check that the students do not cheat during the test, but also supervises teachers in the correction of open-ended questions (see Angrist *et al.*, 2017), reports the answers of students on machine-readable answer sheets and sends them to INVALSI.[11] In non-sampled classes, on the other hand, the tests are managed by teachers, selected among those not belonging to the class being tested and to the subject being assessed.

The sample is selected using a two-stage stratified sampling. In a first stage, school institutes are randomly sampled within regions with the probability of sampling proportional to the number of students enrolled at the beginning of the school year. In a second step, one or two entire classes are randomly selected for monitoring. In institutes with less than 100 pupils in the grade, external monitors observe a single class. In larger institutes, they observe at most two.

2. The data

Our data refer to the universe of Italian primary and middle schools. Although these data are available from academic year 2009/10 until 2016/17, we only use the waves from 2013/14, the initial year when the INVALSI index of cheating propensity by class – one of our outcomes – becomes available. Since we wish to compare results across low and high stake tests, we select 5th graders for the former type of test and 8th graders for the latter type.[12]

We select our final sample as follows: first, we exclude schools located in Valle d'Aosta and Trentino Alto Adige, two smallish Northern regions which decided

[11] See http://banner.orizzontescuola.it/Manuale_osservatore_esterno_2014.pdf.
[12] We do not consider 2nd graders because of the limited available background information, and 6th graders because the test is only available until 2013.

to have all their classes assigned to an external invigilator.[13] Second, we drop: a) classes with less than 10 enrolled students, which are often multi-grade classes; b) school institutes with less than 10 enrolled students in the grade in any year, which are excluded from the sampling of external examiners; c) a handful of classes with missing data on cheating propensity. As the math and literacy tests are taken in different days, sample selection criteria are specific to each test. Finally, we only keep school institutes which are present in the data in years *t, t-1* and *t*-2, because we want to assess the effect of the presence of examiners in the institute in those years on test scores in year *t*.

Because of these selection criteria, we start from 26,875 primary school institute-by-year observations in a population of 7,288 primary school institutes, and end up with 22,984 observations in 6,790 school institutes. For middle schools, we start from 23,232 school institute-by-year observations in a population of 6,181 school institutes, and end up with 20,205 observations in 5,734 school institutes.

In their analysis of the INVALSI monitoring program, Angrist *et al.*, 2017, show that the protocol for the randomization of external examiners is valid across institutes. They also show that the assignment of monitors to classes within institutes is suspect of deviations from randomness. Because of this, we use school institutes as the unit of analysis, and define as treatment variables the presence of an examiner in the institute in year *t, t-1* and *t-2*. For the sake of brevity, and with a slight abuse of language, we shall use the words "school institutes" and "schools" as synonymous hereafter.

As discussed in the previous section, every year INVALSI randomly selects a sample of schools that are subject to external monitoring. The sampling of schools happens within region, and the probability of being sampled is

[13] In our analysis of 8th graders, we also drop 248 schools from Umbria, because for that region and grade we detect significant positive serial correlation in the probability of assignment to external monitors across years.

proportional to the number of students enrolled. Samples are drawn independently every year. Therefore, to guarantee conditional randomization, all our regressions include as randomization controls region-by-year dummies and the interactions of enrollment at *t, t-1* and *t-2* with region-by-year dummies.

For both math and literacy tests, we investigate the dynamic impact of examiners on the following outcomes, computed at the school-by-year level: the average percentage of correct answers (or score) given by each student;[14] the 25th, 50th and 75th quartile and the standard deviation of the score.[15] As argued by Bertoni *et al.*, 2013, the standard deviation of the score is likely to be reduced by outright cheating, as results look more alike across students within schools. In addition, manipulation usually helps low performers more than top students, who would do well in any case.

We also consider as outcome the cheating propensity index computed by INVALSI, a class-level probability of manipulation similar to the one estimated in Angrist *et al.*, 2017, and computed by using information both on the percentage of correct answers and on the patterns of wrong answers.[16] To assess whether the presence of examiners affected the selection of the pool of tested students, we look also at the share of students who were absent in the day of the test. Finally, to dig into the mechanisms behind our uncovered effects, we use item-level data and compute the share of correct answers by school, distinguishing between open-answer and closed-answer (multiple choice) questions. We do so following Angrist *et al.*, 2017, who argue that manipulation in the INVALSI tests arises mainly as a consequence of shirking by internal teachers who devote low effort in correcting open-answer questions (which typically require careful interpretation).

[14] In a robustness test, we also use the mean scores computed by INVALSI by applying the IRT Rasch model to the test answers to account for the fact that items vary in their difficulty.
[15] Given that tests are managed and scores are marked at the class level, we first compute these outcomes by class and then average them by school using class-size weights.
[16] For a detailed description of the method see Quintano *et al.*, 2009.

Our controls include the characteristics of students and schools in year *t, t-1* and *t-2*, that we obtain by matching standardized test scores to information either contained in the student background questionnaires or provided by school staff when scores are submitted to INVALSI. We compute the number of students enrolled in each grade at the beginning of the school year and the school-by-year share of: i) students who attended pre-primary schools; ii) males; iii) immigrants; iv) pupils with parents having elementary, middle, high-school or college education; v) irregular students (i.e. grade-repeaters or early-starters); vi) students in a full-time (8am-4pm) vs. part-time (8am-1pm) schedule; vii) missing values for each control. The descriptive statistics of the outcome and control variables (including the treatment) are shown in Tables 1a and 1b, respectively.

3. The empirical approach

We examine the causal impact of external monitoring on average math and literacy test scores using school-by-year data and the following empirical specification:

$$y_{irt} = \alpha + \beta_1 Monitored_{irt} + \beta_2 Monitored_{irt-1} + \beta_3 Monitored_{irt-2} +$$
$$+\delta_{1rt} Size_{irt} + \delta_{2rt} Size_{irt-1} + \delta_{3rt} Size_{irt-2} + \mu_{rt} +$$
$$+\gamma_1 X_{irt} + \gamma_2 W_{irt-1} + \gamma_3 Z_{irt-2} + \varepsilon_{irt} \quad (1)$$

In equation (1), the indices *i, r* and *t* are for school, region and year; *y* is the outcome variable – measured in year *t*; $Monitored_{irt}$, $Monitored_{irt-1}$ and $Monitored_{irt-2}$ are binary variables equal to 1 if external examiners proctored the test in school *i* in years *t*, *t-1* and *t-2*, and to 0 otherwise.

If the current assignment of an external monitor reduces score manipulation (or cheating), coefficient β_1 should be negative for all our outcomes except the

standard deviation of test scores (for which it should be positive). On the other hand, if the assignment of an external monitor in year *t-1* or *t-2* has no persistent effect on current outcomes, coefficients β_2 and/or β_3 should be equal to zero.

We take into account the INVALSI randomization protocol, which independently samples schools every year at the regional level with a probability of being selected that is proportional to the number of students enrolled at the beginning of the school year, by including in the specification both region-by-year dummies (μ_{rt}) and their interactions with school size in year *t, t-1* and *t-2* ($\delta_{1rt}Size_{irt}$, $\delta_{2rt}Size_{irt-1}$and $\delta_{3rt}Size_{irt-2}$).

In addition, X_{irt} is a vector of control variables which includes the share of male and immigrant students; the share of mothers and fathers with an elementary, middle, high-school diploma and a degree; the share of students who attended pre-primary schools; the share of students following a full-day schedule and the share of irregular students. We further include in vector X_{irt} the share of missing values for each of the covariates described above. The vectors $W_{irt-1}$and Z_{irt-2} contain the same variables included in the vector X_{irt}, but measured in year *t-1* and *t-2* respectively. Finally, ε_{irt} is an error term that we allow to be clustered by school.

If the allocation of schools to external monitors is as good as random, the observables included in vectors X_{irt}, W_{irt-1} and Z_{irt-2} should be balanced across schools with and without randomly assigned external examiners (see Angrist *et al.*, 2017; Bertoni *et al.*, 2013), and their inclusion in the model would be superfluous for identification but useful to increase precision.

We investigate whether this is the case in Tables 2a and 2b (for primary and middle schools), which report the point estimates (with the corresponding level of significance) obtained from regressing each observable on current and lagged monitoring (in year *t-1* and *t-2*). In all regressions, we add randomization controls and cluster standard errors by schools.

For a few covariates, we find that the differences between sampled and non-sampled schools are statistically significant, but that the point estimates are small and close to zero. Since the balancing of covariates is not perfect, and to increase precision, we add all covariates to the vector of controls in our regressions. Nevertheless, our results hold irrespective of whether we include or exclude covariates, lending further support to the internal validity of our research design.[17]

A potential concern when estimating equation (1) is that the effect of previous monitoring on school performance in year *t* might be affected by the spurious correlation between $Monitored_{irt}$, $Monitored_{irt-1}$ and $Monitored_{irt-2}$: although school *i* is randomly sampled by INVALSI and every sample is independently drawn from the population of schools in each year, the probability of being selected in year *t* might be affected by having already had an external invigilator at *t-1* and *t-2* for reasons that go beyond the formal assignment procedures. For instance, this might happen if principals bargained with INVALSI to avoid being monitored for two years in a row.

We verify whether this is a problem by regressing current on lagged monitoring, always controlling for the randomization variables. Tables 3a and 3b report our reassuring results (for primary and middle schools), both without (see column 1) and adding covariates (see column 2), showing that the correlation between lagged monitoring (in year *t-1* and *t-2*) and current monitoring (in year *t)* is small and not statistically significant.

4. Main results

[17] Angrist et al, 2017, also present evidence of balancing across school institutions with and without randomly assigned monitors. They document that administrative data (such as school size, grade enrolment, participation to the test) are well balanced. Demographic variables such as parental education, however, show evidence of imbalance. They argue that "...this seems likely to reflect the influence of monitoring on data quality, rather than a problem with the experimental design or implementation…" (p.14).

Our baseline results for math and literacy are reported in Tables 4a and 4b, respectively.[18] Consistent with the previous literature, we find that the percentage of correct answers in schools where an external examiner was present at the test taken in year *t* is 4.2 percent lower for literacy and 5.4 percent lower for math than in schools that did not have an external examiner – see column (1) in the tables.[19]

We can use these results to derive a rough estimate of the percent reduction in the share of correct answers in the class where the external examiner was in fact present. Let μ and μ_i be the average score in the school and class, and let α_i be the share of pupils in class *i*. Suppose that the external examiner in class *i* reduces the average score in the class from μ_i to $\pi\mu_i$, where $\pi < 1$. In the absence of spillover effects from one class to another, the average score in the school declines to $\hat{\mu} = \mu - (1-\pi)\alpha_i\mu_i$. Therefore, the percent change in the school mean score due to the presence of an examiner in class *i* is $\frac{\hat{\mu}-\mu}{\mu} = \frac{(\pi-1)\alpha_i\mu_i}{\mu}$. If $\mu_i \cong \mu$ and α_i is equal to the average share of pupils in each class of the grade – 0.257 in our data – the percent change in class *i* is 5.4 percent / 0.257=21.1 percent for math and 4.1 percent / 0.257 = 15.9 percent for literacy.[20]

If the presence of an external examiner had only a temporary effect of average school test scores, having had an examiner in year *t-1* or *t-2* should have no effect on test scores in year *t*. Yet we find that schools which had an external examiner during the test taken at *t-1* experience a statistically significant[21] reduction in the percent of correct answers in the test taken in year *t*, ranging from 0.5 percent for literacy to 0.7 percent for math. This effect is roughly 1/7

[18] In Table A1 and A2 in the Appendix we report the same results without the controls in vectors X_{irt}, W_{irt-1} and Z_{irt-2}.
[19] Percent changes are computed by dividing the treatment effect by the mean outcome for the control group.
[20] This change is larger the lower the share of pupils in the monitored class.
[21] At the 5 or 10 percent level of confidence.

of the current examiner's effect.[22] Persistency, however, is short-lived: the binary treatment in year *t-2* does not produce a statistically significant effect on test scores at time *t*, implying that over time schools revert to their original behavior.[23]

We also evaluate the dynamic impact of the external examiner on the bottom, median and top quartile of the school-specific distribution of test scores (see columns (2), (3) and (4) of Tables 4a and 4b).[24] If external examiners persistently affect the propensity to manipulate test scores, we expect a higher impact on the bottom part of the distribution of scores, because low-performing students are likely to benefit more from manipulation than top performers, who would have scored well in any case. This conjecture is in line with our findings for monitoring the math test in both year *t* and *t-1*. Results for year *t-1* are however less clear-cut when we consider the literacy test.

Next, we consider the effect of external monitoring on the within-school standard deviation of scores. As discussed by Bertoni *et al.*, 2013, manipulation is expected to reduce the variability of test results. Therefore, if the presence of an external examiner reduces manipulation, we expect an increased dispersion in the performance distribution within schools. As shown in column (5) of both tables, there is a significant and positive effect of monitoring in year *t* on the standard deviation of scores in year *t.* However, the impact of having had an external examiner in the school in year *t-1* is only significant for the math test, and the effect of monitoring in year *t-2* is always very close to zero.

[22] Our estimates do not change qualitatively when we replace the percentage of correct answers as dependent variable with the score computed by INVALSI using the IRT Rasch model to account for the fact that questions vary in their difficulty. Results are reported in Table A3 in the Appendix.

[23] We have explored whether having had the external monitor in the previous year improves the ability of the current monitor to prevent cheating (by interacting *Monitored in year t* with *Monitored in year t-1*), but have found no evidence that this is the case.

[24] As discussed in footnote 15, we compute these outcomes by class and then average by school after weighting each class by its size.

In column (6) we turn our attention to the INVALSI cheating index. We find that schools being monitored in year *t* show a large reduction in the cheating index – ranging between 43 and 50 percent. Having been monitored in year *t-1* also reduces the index by 8.5 percent for literacy and by 11.7 percent for math. No effect is found instead for monitoring in year *t-2*.

In column (7) we investigate whether external monitoring affects test scores by limiting the opportunistic behavior of both teachers and principals who have an incentive to manipulate the pool of test takers and induce poorly performing students not to show up at the test – see Bertoni *et al.* 2013, and Lucifora and Tonello, 2016. We find that, while monitoring in year *t* reduces absences, monitoring in year *t-1* and *t-2* has no effect.

Finally, in columns (8) and (9) we consider as outcome variables the percentage of correct answers in open-ended and close-ended questions, respectively. As argued by Angrist *et al.*, 2017,[25] evaluating the first type of questions requires more effort and is more discretionary since teachers have to interpret and transcribe students' answers into the machine-readable sheet called "scheda risposta". Because of this, the answers to these questions are more likely to be manipulated by dishonest or lazy teachers. This conjecture finds support in our estimates, showing that the negative effect of monitoring in year *t* and *t-1* on the percentage of correct answers is larger in absolute value for open-ended than for close-ended questions. Monitoring in year *t-2*, however, has no effect on either type of questions.

5. Results by school size and geographic area

5.1. School size

The results presented above are based on data collapsed by school to bypass the threat to randomization induced by the fact that school principals have some

[25] See also Dee *et al.*, 2019.

discretion in allocating the external inspector to classes.[26] We have seen above that, when there are no spill-over effects from the treated class to the other classes and pre-treatment means are similar across classes in the same grade, the percent change in the school mean score caused by the presence of the external examiner in one class can be written as $\frac{\hat{\mu}-\mu}{\mu} = (\pi - 1)\alpha_i$, where α_i is the share of pupils in the treated class (for a given grade).

Since this share is inversely related to the number of classes, we expect the average effect of the external examiner in year *t* on the contemporaneous average score to be mechanically smaller in schools with many classes in the same grade. This does not hold, however, for the effect of the external monitor at time *t-1* or *t-2*, which applies to the whole school rather than to the single treated class.

We investigate whether the relationship between lagged monitoring and test scores varies with the size of school in Tables 5a, 5b and 5c (math) and 6a, 6b and 6c (literacy). We classify schools in three groups: i) small, with 10 to 35 pupils in the grade, corresponding to a median number of classes in the grade equal to 1; ii) medium, with 36 to 75 pupils in the grade, corresponding to a median number of classes in the grade equal to 3; iii) large, with more than 75 pupils and close to 6 classes in the grade (median value).[27]

Our estimates show that the effect of having been monitored in year *t-1* on current test outcomes is negative, often statistically significant and declining in absolute value with the size of schools. Monitoring in year *t-2* instead often attracts a positive coefficient which is almost never statistically significant at

[26] School principals might adopt opportunistic behavior in choosing classes monitored by the external invigilator in order to select those that usually perform better than others within the same school (Angrist *et al.*, 2017). In fact, the incentives of principals to select better classes are very strong, since they might be interested in achieving high scores in INVALSI tests to attract in the following years better stakeholders, such as high-skilled students or students whose parents have a stronger socio-economic background. In this case, principals' behavior would invalidate the randomization protocol of classes within the same school used by INVALSI.
[27] We require these conditions to be met in each year (*t, t-1* and *t-2*).

the 5 or 10 percent level of confidence. The presence of an external examiner in year *t-1* in schools with less than 35 pupils in the 5th grade generates a 4.4 percent reduction in current test performance for math and a 5.5 percent reduction for literacy. This effect declines in absolute value to 1.8 and 1.2 percent in medium–sized schools and is close to zero in larger schools.

When looking at the other outcomes of interest we also find heterogeneous effects by school size. Supporting the view that low performing students are those benefitting the most from manipulation, we find that in schools with less than 35 pupils in the grade the effect of both current and past monitoring (*t-1*) is larger in absolute value for the bottom quartile of the math and literacy score distribution (columns 2, 3 and 4 of Tables 5a and 6a) than for the median and top quartile. This difference is much less pronounced or even absent in medium and large size schools (columns 2, 3 and 4 of Tables 5b and 5c and Tables 6b and 6c).

We also find that in small schools both current and past monitoring (in year *t-1*) increase the within-school standard deviation of math and literacy test scores (column 4 of Tables 5a and 6a) and reduce the cheating test index (column 6 of Tables 5a and 6a), suggesting that past monitoring affects current cheating. The latter effect is sizeable in small schools (-61.5 percent for math and -53 percent for literacy), almost half as big in medium schools (-32 percent for math and -24 percent for literacy) and lower than 10 percent in large schools. Yet, and independently of school size, monitoring in year *t-2* never affects current test scores.[28]

5.2. Geographic area

[28] We check the robustness of these results by splitting our sample according to the number of classes in the 5th grade and by running separate regressions for schools with no more than one class in the grade, 2 to 3 classes and more than 3 classes in the grade. As shown in the appendix of the paper (Tables A4a, A4b, A4c, A5a, A5b, and A5c) we find results that are qualitatively very similar to those discussed above.

Italy is very heterogeneous both in terms of economic conditions and of social capital, with the North and Centre being richer and endowed with higher social capital than the South. We investigate whether the effect of previous external monitoring on current test scores varies by macro area and find that it has a statistically significant effect on the current test score only in the Northern and Central regions, in spite of the fact that current monitoring has a much larger effect on current test scores in the South (see Tables 7a, 7b, 8a and 8b).

Although the main difference in the endowment of social capital in Italy is between the Centre-North and the South, there are also differences within each area that we capture by considering the following province-specific measures of the propensity to cooperate and to create collective goods: i) voter turnout in referenda where voting is not mandatory and ii) blood donation (already used by Guiso *et al.*, 2004).[29] Voter turnout refers to all the referenda that occurred in Italy between 1946 and 1989 (data from the World Value Social Survey),[30] and blood donation is measured by the number of blood bags (each bag containing 16 ounces of blood) per million inhabitants collected by AVIS, the national agency. We find (see Tables 9a-9h) that the effect of past monitoring on current scores is statistically significant only for the schools located in areas endowed with high social capital.[31]

6. Mechanisms

[29] See also Ferrer-Esteban, 2013.

[30] These referenda cover a very broad set of issues, including the choice between republic and monarchy (1946), divorce (1974), abortion (1981), hunting regulations (1987), use of nuclear power (1987) and public order measures (1978, 1981).

[31] We define areas with high (low) social capital as the areas where voter turnout at referenda and blood donation are above (below) the median. Notice, however, that differences between schools located in areas endowed with varying levels of social capital are not statistically significant. Similar results (not reported but available upon request) hold when we use voter turnout in European elections, which is available at municipal level. Our results are in line with the findings by Paccagnella and Sestito, 2014, who document the negative correlation between school cheating and measures of social capital at the local level.

In low stakes tests, the benefits from score manipulation include: i) lower teacher effort in the transcription of results into machine-ready answer sheets, for instance by copying all or part of an answer key, especially when questions are open-ended; ii) helping students and partially or entirely offsetting poor results that could be attributed to teaching deficiencies. The costs of manipulation include: i) reputational loss in the event of detection. Starting from 2013, INVALSI has implemented a sanctioning policy based on a "fame and shame" mechanism, consisting of two measures: deflation of class test scores and non-return of test scores to the class and school when the computed cheating index is above a threshold (see Lucifora and Tonello, 2016); ii) potential conflict with honest teachers or with teachers whose classes were proctored by the external examiner.

We have shown that: i) the negative effects of external monitoring on average test scores extend beyond the current test and involve also the tests taken in the following year; ii) these effects are stronger in smaller schools and in areas with higher social capital; iii) they fade away after two years. In this section, we discuss mechanisms that could explain short-term persistency.

We start by noticing that the finding that external monitoring reduces tests scores points to teachers as the main source of manipulation. According to Angrist *et al.*, 2017, "…honest teacher-proctors should have the same deterrent effect as external monitors on cheating students: both are likely to catch cheaters, teachers even more so if they recognize cheating more readily. External monitoring should therefore have little effect on student cheating unless cheating is accomplished with the collaboration or at least assent of school staff..." (p.11).

A candidate mechanism driving the negative effect of having had an external examiner in year *t-1* on test scores in year *t* is teacher learning. In the treated classes, monitors supervise sheet transcription, a task completed by local school staff by the end of the test day. In non-treated classes, this task is not supervised.

The teachers who interact with the examiner may learn how to code correctly the answers to open-ended questions and eventually pass on this skill to other teachers.

Since the probability that affected teachers are involved in the proctoring of tests in the next year is higher in smaller schools, teacher learning is consistent with the larger negative effect found both in these schools and for open-ended questions, but does not explain why the effect of the external examiner disappears after two years. One might think of teacher turnover, yet less than 10% of teachers state that their tenure in the school is lower than 2 years in the self-reported data from the teacher questionnaire administered by INVALSI.

An important drawback of learning as the main story is that one would expect that literacy results are more affected than math results, simply because math is generally more straightforward to grade and less ambiguous, thus providing less scope for the learning effect. Yet we find that the effect of lagged monitoring is about as large for math and literacy open-ended questions and larger for math than for literacy close-ended questions (see Tables 4a and 4b).

Another candidate mechanism is reputational concerns: teachers and school administrators might not revert in year *t* to the level of cheating they would have had without external monitoring in year *t-1* because they are afraid that either INVALSI or other school stakeholders may identify them as cheaters. The higher the reduction induced by the external monitor the more difficult it is to revert to previous results in the following year without running the risk of being identified. This is particularly true for small schools, where most if not all classes in the grade have been monitored in year *t-1*. For these schools, returning in year *t* to the pre-monitoring levels of cheating would produce a larger and therefore more noticeable swing in test scores.

A drawback of this mechanism is that these concerns may not be credible when explicit sanctions for misbehavior are either absent or ineffective, as shown by

Lucifora and Tonello, 2016, in their study of the impact of the "fame or shame" sanctions introduced by INVALSI in 2013.

A third candidate mechanism is teacher peer pressure. Let us assume that cheating is widespread in the absence of the external examiner. When the latter walks in a class, the teacher of that class cannot engage in cheating and the expected class-specific score is lower. In order not to look bad, he/she may exert pressure on fellow teachers so that in the following year, when no external monitor is present in the school, there is no full reversion to the original manipulation. Since teachers involved in the 5^{th} grade within a school typically rotate, they may dislike sharp variations in test scores from one year to another in order to attain similar evaluations from the principal.

Peer pressure, however, does not explain why the presence of an external monitor in year *t-2* has no effect on current scores, unless there is a presumption of myopia and comparison of outcomes only across neighboring years. It also does not explain why short-term persistence is confined to the Northern and Central areas of the country.

The last mechanism that we consider is compliance with local identity, where identity is a person's sense of self, as in Akerlof and Kranton, 2000. When identity varies across areas, individual payoffs from the same action – in this case score manipulation – may differ depending on whether this action conforms or not with the behavioral prescriptions of one's identity. Under the realistic assumption that the areas of Italy endowed with high social capital also share a self-image prescribing not to cheat, the presence of monitors makes this identity more salient, promoting conformism even in the year after the examiners visited the school. Conversely, in areas with low social capital external monitoring has no persistent effects because the local identity is more open to cheating and the policy does not induce individuals to conform to local values.

Given the institutional setting considered in our analysis, where no explicit sanctions for misbehavior are in place, the effect of past monitoring on cheating can be assimilated to the one produced by nudging policies that induce changes in individual choices without altering the option set or incentives. As for nudges relying on the salience bias, the contact with external monitors can lead individuals to behave honestly (even when monitoring is no more in place) by making honesty a more salient virtue.

Adherence with local identity might also contribute to explain why the impact of external monitoring at time *t-1* on current test scores varies with the size of the school. Compared to medium-large schools, a higher share of teachers in small institutes is exposed to external examiners and involved in the invigilation of tests in the current and following year. This higher exposure increases current compliance with local identity, with effects that could spill-over in future tests as well. Yet, and consistent with the results of the literature on nudging (see for instance Manoli and Turner, 2018), the effect of past monitoring is short lived – even in small institutes – and has no effects when subjects have strong contrary preferences (Sunstein, 2017).

In summary, we exclude teacher learning because it is not consistent with our results for literacy and peer pressure because it does not explain why the effects of lagged monitoring are absent in the South. We are also skeptical about reputational concerns because of the absence of credible sanctions to misbehave. Therefore, we consider compliance with local identity in an environment characterized by heterogeneous identities as the most convincing mechanism explaining our results.

8. The impact of past monitoring on high stake tests

We expect the effect of the external examiner to be lower in absolute value when stakes are high, either because the incentives for students and teachers to cheat are much high, which makes it more difficult for the external monitor to deter

it, or because cheating is inherently more difficult due to more stringent controls on test procedures carried out in control classes by the involved stakeholders (the principal, other teachers, parents).

To investigate whether this is the case, we estimate equation (1) using data for the 8th grade. The scores in the math and literacy tests taken in that grade are part of the final exit exam that students need to pass in order to enroll in upper secondary education and eventually college.[32] Using the same specifications as in Table 4, we report our findings in Tables 10a and 10b for math and literacy, respectively. These findings confirm our expectations. First, the impact of the current external examiner on the average percent of correct answers is much smaller than in low stakes tests (-0.9 percent in math versus -5.4 percent in the 5th grade and -0.4 percent in literacy versus -4.1 percent in the 5th grade). Second, we find no evidence that having had an external examiner in year *t-1* or *t-2* affects current test scores.

Conclusions

Standardized tests that measure and compare students' cognitive skills have become common in many countries. While in some countries these assessments are used mainly to provide external comparisons with no formal consequences on schools or students, in other countries they are employed either to evaluate teachers or to select students applying to different educational tracks.

A well-known problem with testing is score manipulation, which happens both in low and high stakes tests, and undermines both the reliability of results and the possibility of using them to compare schools and countries and support accountability policies. The existing empirical evidence points out that external monitoring is effective in reducing manipulation problems. The relevant

[32] As a consequence, students have an immediate interest in trying to exchange information with their peers and to use prohibited materials and technologies. Similarly, teachers and school administrators, might face pressure from students' families to lower the monitoring standards or to teach to the test.

literature, however, has focused exclusively on the immediate incapacitation effects of monitoring, with the implicit assumption that these effects vanish once external invigilators leave the school.

In this paper, we have questioned this assumption by investigating whether the presence of external examiners can also impact future test scores. Using the repeated random assignment of external examiners to Italian primary schools, we have found that external monitoring reduces average test scores and cheating not only currently but also in the year after its implementation. After two years, however, the effect vanishes. We have discussed potential mechanisms that could explain short-term persistence, including learning, reputational concerns, peer pressure and compliance with local identity and argued that the last provides the most convincing story.

Our results provide useful input for the correct assessment of the costs and benefits of policies that try to reduce score manipulation and show that considering only the current impact of external invigilators under-estimates the benefits, especially in small schools and in areas endowed with high social capital.

The finding that rules enforcement produces effects on individual behavior also when the risk of punishment is low has implications also for the recent literature documenting the effects of audits on taxpayers' future behavior. While this literature has explained short-term persistence with the expected probability of being re-audited, our paper suggests that an alternative mechanism could be that being in contact with the enforcement authority makes honesty a more salient value. This mechanism would explain the results by DeBacker *et al.*, 2018, showing that audits increase future reported taxable income even in the situation where audit probability is random and audits are conducted exclusively for research purposes.

References

Ahn, T., and Vigdor, J., 2014. *The impact of No Child Left Behind's accountability sanctions on school performance: Regression discontinuity evidence from North Carolina.* NBER working paper n. 20511.

Akerlof, G.A., and Kranton, R.E., 2000. Economics and identity. *The Quarterly Journal of Economics*, *115*(3), 715-753.

Alm, J., 2019. What motivates tax compliance? *Journal of Economic Surveys,* 33(2), 353-388.

Angrist, J.D., Battistin, E., and Vuri, D., 2017. In a small moment: Class size and moral hazard in the Italian Mezzogiorno. *American Economic Journal: Applied Economics*, 9(4), 216-249.

Ayers, B., C., Seidman, J.K., and Towery, E.M., 2019. Tax reporting behavior under audit certainty. *Contemporary Accounting Research*, 36(1), 326-358.

Baird, S., Hicks, J.H., Kremer, M., and Miguel, E., 2016. Worms at work: Long-run impacts of a child health investment. *Quarterly Journal of Economics*, 131(4), 1637-1680.

Barbarino, A., and Mastrobuoni, G., 2013. The Incapacitation effect of incarceration: Evidence from several Italian collective pardons. *American Economic Journal: Economic Policy*, 6(1), 1-37.

Battistin, E., 2016. How manipulating test scores affects school accountability and student achievement. *IZA World of Labor*.

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

Becker, G.S., 1968. Crime and punishment: An economic approach. *Journal of Political Economy*, 76(2), 169-217.

Beer, S., Kasper, M., Kirchler, E., and Erard, B., 2016, *Do audits deter future noncompliance? Evidence on self-employed taxpayers*. IMF working papers 223.

Bertoni, M., Brunello, G., and Rocco, L., 2013. When the cat is near, the mice won't play: The effect of external examiners in Italian schools. *Journal of Public Economics*, 104, 65-77.

Bloom, N., Mahajan, A., McKenzie, D., and Roberts, J., 2018. *Do management interventions last? Evidence from India*. NBER working paper n. 24249.

Buonanno, P., and Raphael, S., 2013. Incarceration and incapacitation: Evidence from the 2006 Italian collective pardon. *American Economic Review*, 103(6), 2437-2465.

Campodifiori, E., Figura, E., Papini, M., and Ricci, R., 2010. *Un indicatore di status socio-economico-culturale degli allievi della quinta primaria in Italia*. INVALSI Working paper 27.

Chetty, R., Hendren, N., and Katz, L.F., 2016. The effects of exposure to better neighborhoods on children: New evidence from the Moving to Opportunity experiment. *American Economic Review*, 106(4), 855-902.

DeBacker J., Heim, B., Tran, A., and Yuskavage, A., 2015, Legal enforcement and corporate behavior: An analysis of tax aggressiveness after an audit. *The Journal of Law and Economics*, 58(2), 291-324.

DeBacker, J., Heim, B.T., Tran, A., and Yuskavage, A., 2018, Once bitten, twice shy? The lasting impact of enforcement on tax compliance. *The Journal of Law and Economics*, 61(1), 1-35.

Dee, T. S., Dobbie, W., Jacob, B.A., and Rockoff, J., 2019. The causes and consequences of test score manipulation: Evidence from the New York Regents examinations. *American Economic Journal: Applied Economic*s, 11(3), 382-423.

Diamond, R. and Persson, P., 2016. *The long-term consequences of teacher discretion in grading of high-stakes tests*. NBER working paper n. 22207.

Drago, F., Galbiati, R., and Vertova, P., 2009. The deterrent effects of prison: Evidence from a natural experiment. *Journal of Political Economy*, 117(2): 254-280.

Duflo, E., Greenstone, M. and Ryan, N., 2013. Truth-telling by third-party auditors and the response of polluting firms: Experimental evidence from India. *Quarterly Journal of Economics*, 128(4), 1499-1545.

Ferrer-Esteban, G., 2013. *Rationale and incentives for cheating in the standardised tests of the Italian assessment system*. Working Paper.

Guiso, L., Sapienza, P., and Zingales, L., 2004. The role of social capital in financial development. *American Economic Review*, 94(3), 526-556.

Heckman, J.J., and Karapakula, G., 2019. *The Perry preschoolers at late midlife: A study in design-specific inference.* NBER working paper n. 25888.

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

Levitt, S.D., 1996. The effect of prison population size on crime rates: Evidence from prison overcrowding litigation. *The Quarterly Journal of Economics*, 111(2), 319-351.

Lucifora, C., and Tonello, M., 2016. *Monitoring and sanctioning cheating at school: What works? Evidence from a National Evaluation Program*. Working Paper Series, Dipartimento di Economia e Finanza, Università Cattolica del Sacro Cuore.

Manoli, D. and Turner, N., 2018. *Nudges and learning: Evidence from informational interventions for low-income taxpayers*. NBER working papers 20718.

Mittone, L., 2006. Dynamic behaviour in tax evasion: An experimental approach. *Journal of Socio-Economics*, 35(5), 813-35.

Mittone, L., Panebianco, F., and Santoro, A., 2017. The Bomb-Crater effect of tax audits: Beyond the misperception of chance. *Journal of Economic Psychology*, 61, 225-243.

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

Quintano, C., Castellano, R., and Longobardi, S., 2009. A fuzzy clustering approach to improve the accuracy of Italian student data. An experimental procedure to correct the impact of outliers on assessment test scores. *Statistica & Applicazioni*, 7, 149-171.

Pereda-Fernandez, S., 2018. *Teachers and cheaters. Just an anagram?* Mimeo Banca d'Italia.

Sunstein, C., 2017. Nudges that fail, behavioral public policies, Oxford University Press.

Tables

Table 1a. Descriptive Statistics – outcome variables

	(1) Primary schools N = 22,984				(2) Middle schools N = 20,205			
	Math		Literacy		Math		Literacy	
	Mean	Std. dev	Mean	Std. dev	Mean	Std. dev	Mean	Std. dev
Within-school score distribution:								
Mean	61.73	10.95	63.97	8.83	57.26	8.19	65.87	6.26
Std. dev	15.25	3.54	15.16	3.24	16.51	2.99	15.35	2.45
25^{th} percentile	51.17	13.35	53.77	11.16	45.31	9.51	55.59	7.82
50^{th} percentile	62.31	11.79	65.22	9.48	57.07	9.08	67.24	6.86
75^{th} percentile	72.86	9.87	76.18	7.74	69.19	8.51	77.36	5.89
Cheating index	0.04	0.09	0.04	0.09	0.04	0.06	0.04	0.07
% absent students	14.41	11.77	14.96	12.58	0.09	0.07	0.09	0.07
% correct open-ended questions	61.75	12.34	64.03	11.58	53.26	8.75	58.68	7.99
% correct close-ended question	60.41	12.09	63.88	8.60	59.36	8.38	68.31	6.36

Note: INVALSI SNV data.

Table 1b. Descriptive Statistics – controls (in year *t*)

	(1) 5th graders N = 22,984		(2) 8th graders N = 20,205	
	Mean	Std.dev	Mean	Std.dev
Panel A. School and area				
Monitored in year *t*	0.10	0.30	0.23	0.42
Monitored in year *t-1*	0.10	0.30	0.23	0.42
Monitored in year *t-2*	0.11	0.32	0.22	0.41
# students enrolled in year *t*	78.84	43.12	97.58	56.29
# students enrolled in year *t-1*	77.82	42.78	97.84	56.66
# students enrolled in year *t-2*	76.61	42.63	97.47	56.85
South	0.37	0.48	0.37	0.48
Panel B. Student in year t				
% Male students	0.50	0.07	0.51	0.08
% Fathers with a middle school diploma	0.28	0.17	0.32	0.17
% Fathers with a high school diploma	0.38	0.18	0.36	0.17
% Fathers with a degree	0.13	0.14	0.10	0.11
% Mothers with a middle school diploma	0.23	0.16	0.27	0.16
% Mothers with a high school diploma	0.43	0.19	0.41	0.19
% Mothers with a degree	0.16	0.15	0.12	0.11
% Regular	0.95	0.08	0.89	0.08
% Immigrants	0.09	0.10	0.09	0.09
% Kindergarten	0.78	0.35	0.75	0.36
% Daycare	0.24	0.21	0.19	0.17
% Full-time	0.07	0.22	0.08	0.22
Panel C. % Missing in year t				
% Male students missing	0.00	0.06	0.00	0.04
% Fathers' education missing	0.18	0.29	0.19	0.28
% Mothers' education missing	0.16	0.29	0.17	0.28
% Regular missing	0.01	0.06	0.00	0.05
% Immigrants missing	0.01	0.08	0.01	0.06
% Kindergarten missing	0.12	0.28	0.11	0.26
% Daycare missing	0.27	0.38	0.22	0.36
% Full-time missing	0.01	0.08	0.22	0.41

Notes: to save space we only report descriptive statistics for student characteristics in year *t*. Descriptive statistics for covariates in year *t-1* and *t-2* are available from the authors. The omitted category for parental education is primary education.

Table 2a. Balancing tests – 5th graders.

Covariate in year	(1) *t*			(2) *t-1*			(3) *t-2*		
	(a)	(b)	(c)	(a)	(b)	(c)	(a)	(b)	(c)
Monitored in year	*t*	*t-1*	*t-2*	*t*	*t-1*	*t-2*	*t*	*t-1*	*t-2*
% Male students	0.001	0.001	-0.000	-0.000	0.001	-0.001	0.002	0.001	0.001
% Fathers with a middle school diploma	-0.004	-0.005	-0.001	-0.005	-0.003	-0.003	-0.006	-0.005	0.003
% Fathers with a high school diploma	0.003	-0.002	0.001	-0.008*	0.006	-0.005	0.000	-0.007*	0.011***
% Fathers with a degree	0.002	-0.002	-0.001	0.001	0.001	-0.001	0.001	-0.002	0.002
% Mothers with a middle school diploma	-0.001	-0.004	0.000	-0.004	-0.001	-0.003	-0.002	-0.004	0.003
% Mothers with a high school diploma	0.002	-0.002	-0.002	-0.009**	0.005	-0.006	-0.004	-0.009**	0.011***
% Mothers with a degree	0.002	-0.002	-0.000	0.001	0.001	-0.001	-0.000	-0.001	0.003
% Regular	0.002*	0.000	-0.000	0.001	0.003***	-0.001	0.002*	-0.000	0.001
% Immigrants	-0.002	0.000	0.000	-0.000	0.000	0.002	0.001	0.001	0.003
% Kindergarten	-0.003	-0.020**	-0.010	-0.013	-0.009	-0.013	-0.004	-0.016*	0.004
% Day care	-0.002	-0.006	-0.002	-0.004	-0.004	-0.002	0.002	-0.009**	0.001
% Full-time	0.004	0.002	-0.006	0.004	0.002	-0.005	-0.001	0.007	0.000
% Male students missing	-0.002**	-0.000	0.000	-0.001	-0.002***	-0.000	-0.002***	-0.001*	-0.001*
% Fathers' education missing	-0.001	0.009	0.001	0.012*	-0.004	0.009	0.006	0.014*	-0.018**
% Mothers' education missing	-0.002	0.009	0.001	0.013*	-0.004	0.009	0.007	0.015*	-0.018**
% Regular missing	-0.002**	0.000	0.000	-0.001	-0.002***	0.000	-0.002**	-0.002**	-0.002**
% Immigrants missing	0.001	0.000	0.001	-0.001*	-0.002**	-0.000	0.002	-0.005**	-0.010***
% Kindergarten missing	0.005	0.011*	0.003	0.015**	0.010	0.010*	0.003	0.017**	-0.002
% Day care missing	0.007	0.022**	0.004	0.018*	0.015*	0.011	0.009	0.028***	0.004
% Full-time missing	-0.001	0.001	-0.000	-0.002	-0.001	0.000	0.002	-0.005**	-0.009***

Note: The table reports the coefficients of balancing regressions of each covariate on monitored in year *t, t-1* and *t-2* and randomization controls (region-by-year dummies and their interactions with current and lagged enrolment). Separate regressions are run for each covariate measured in year *t, t-1* or *t-2*. As a result, for each row and column, sub-columns a, b, and c report coefficients from the same regression. Different rows and columns refer instead to different regressions. Standard errors clustered by school are omitted to save space. ***, **, * for statistical significance at the 1, 5 and 10 percent level of confidence. The number of observations is 22,984.

Table 2b. Balancing tests – 8th graders.

Covariate in year	(1) *t*			(2) *t-1*			(3) *t-2*		
	(a)	(b)	(c)	(a)	(b)	(c)	(a)	(b)	(c)
Monitored in year	*t*	*t-1*	*t-2*	*t*	*t-1*	*t-2*	*t*	*t-1*	*t-2*
% Male students	-0.003**	0.001	-0.001	0.000	-0.000	0.001	-0.001	0.001	0.001
% Fathers with a middle school diploma	0.001	-0.003	0.000	-0.002	-0.001	-0.002	0.001	-0.001	-0.002
% Fathers with a high school diploma	0.004	0.002	-0.004	0.004	0.000	-0.004	0.003	0.001	-0.001
% Fathers with a degree	0.004*	0.002	-0.001	0.003*	0.002	0.002	0.002	0.002	0.001
% Mothers with a middle school diploma	0.001	-0.000	-0.001	-0.001	-0.001	0.001	0.000	-0.001	-0.002
% Mothers with a high school diploma	0.004	-0.000	-0.004	0.004	0.000	-0.007**	0.004	0.002	-0.001
% Mothers with a degree	0.004**	0.002	-0.001	0.003*	0.002	0.002	0.002	0.002	0.001
% Regular	0.001	0.001	-0.001	-0.000	0.001	0.000	0.002	-0.001	-0.001
% Immigrants	-0.003*	-0.002*	0.000	0.000	-0.003**	-0.003*	-0.004***	-0.001	-0.004***
% Kindergarten	0.003	0.004	-0.002	-0.003	-0.001	-0.001	-0.001	-0.002	-0.007
% Day care	0.005	0.001	-0.002	0.004	-0.002	-0.002	0.001	0.001	-0.004
% Full-time	-0.002	-0.004	-0.005	-0.003	0.002	-0.006**	-0.001	-0.002	0.000
% Male students missing	-0.001	-0.001**	-0.000	-0.001	-0.001	-0.001	0.000	-0.000	-0.001
% Fathers' education missing	-0.008	-0.001	0.005	-0.005	-0.001	0.005	-0.007	-0.002	0.002
% Mothers' education missing	-0.008*	-0.001	0.005	-0.006	-0.000	0.005	-0.006	-0.002	0.002
% Regular missing	-0.001	-0.001	0.000	-0.001	-0.001	-0.001	0.000	-0.000	-0.001
% Immigrants missing	-0.001	-0.000	-0.001	-0.001	-0.001	-0.001*	0.000	-0.001	-0.001
% Kindergarten missing	-0.003	-0.006	0.002	-0.004	-0.004	-0.004	0.000	-0.007	0.001
% Day care missing	-0.000	0.001	0.003	-0.006	0.006	0.007	0.000	-0.002	0.010
% Full-time missing	-0.001*	-0.001*	-0.001*	-0.000	-0.000	-0.001	0.001	-0.000	-0.001

Note: The table reports the coefficients of balancing regressions of each covariate on monitored in year *t*, *t-1* and *t-2* and randomization controls (region-by-year dummies and their interactions with current and lagged enrolment). Separate regressions are run for each covariate measured in year*t*, *t-1* or *t-2*. As a result, for each row and column, sub-columns a, b, and c report coefficients from the same regression. Different rows and columns refer instead to different regressions. Standard errors clustered by school are omitted to save space. ***, **, * for statistical significance at the 1, 5 and 10 percent level of confidence. The number of observations is 20,782.

Table 3a. Correlation between current and lagged monitoring. 5th graders. With and without additional controls.

	(1)	(2)
Outcome variable	Monitored in year *t*	Monitored in year *t*
Monitored in year *t-1*	0.007	0.008
	(0.008)	(0.008)
Monitored in year *t-2*	-0.002	-0.003
	(0.007)	(0.007)
Observations	22,984	22,984
Other controls	No	Yes
Randomization controls	Yes	Yes

Note: Each regression includes randomization controls (region by wave dummies and their interactions with current and lagged enrolment). Column (2) also includes the additional controls in vectors X, W and Z. Standard errors clustered by school within parentheses. ***, **, * for statistical significance at the 1, 5 and 10 percent level of confidence.

Table 3b. Correlation between current and lagged monitoring. 8th graders. With and without additional controls.

	(1)	(2)
Outcome variable	Monitored in year *t*	Monitored in year *t*
Monitored in year *t-1*	0.011	0.012
	(0.008)	(0.008)
Monitored in year *t-2*	0.005	0.005
	(0.008)	(0.008)
Observations	20,205	20,205
Other controls	No	Yes
Randomization controls	Yes	Yes

Note: Each regression includes randomization controls (region-by-year dummies and their interactions with current and lagged enrolment). Column (2) also includes the additional controls in vectors X, W and Z. Standard errors clustered by school within parentheses. ***, **, * for statistical significance at the 1, 5 and 10 percent level of confidence.

Table 4a. The effects of external monitoring on test scores. Math - 5th graders.

Outcome variable	(1) Mean	(2) Bottom quartile	(3) Median	(4) Top quartile	(5) Standard deviation	(6) Cheating index	(7) Percent absent in the test	(8) Mean open-ended questions	(9) Mean close-ended questions
Monitored in year *t*	-3.372***	-4.057***	-3.606***	-2.938***	0.626***	-0.020***	-0.004**	-4.282***	-2.567***
	(0.195)	(0.231)	(0.214)	(0.187)	(0.054)	(0.001)	(0.002)	(0.225)	(0.185)
Monitored in year *t-1*	-0.440**	-0.560**	-0.546**	-0.307	0.130**	-0.005***	0.001	-0.565**	-0.348*
	(0.204)	(0.244)	(0.223)	(0.190)	(0.059)	(0.002)	(0.002)	(0.235)	(0.195)
Monitored in year *t-2*	0.172	0.160	0.131	0.170	-0.001	0.001	0.001	0.088	0.207
	(0.196)	(0.236)	(0.214)	(0.182)	(0.059)	(0.002)	(0.002)	(0.233)	(0.184)
Observations	22,984	22,984	22,984	22,984	22,984	22,984	22,984	22,984	22,984
Other controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Randomization controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Mean for control group at *t*	62.11	51.62	62.70	73.17	15.17	0.046	0.144	62.20	60.70
Mean for control group at *t-1*	61.85	51.31	62.43	72.93	15.22	0.045	0.143	61.87	60.51
Mean for control group at *t-2*	61.69	51.13	62.27	72.81	15.25	0.045	0.143	61.87	60.27
% change for monitored at *t*	-0.054	-0.078	-0.057	-0.040	0.041	-0.426	-0.031	-0.068	-0.042
% change for monitored at *t-1*	-0.007	-0.010	-0.008	-0.004	0.008	-0.117	0.006	-0.009	-0.005
% change for monitored at *t-2*	0.002	0.003	0.002	0.002	-0.001	0.028	0.003	0.001	0.003

Note: each regression includes randomization controls (region-by-year dummies and their interactions with current and lagged enrolment) and the other controls in vectors X, W and Z. Standard errors clustered by school within parentheses. Percent changes for monitored at *t, t-1* and *t-2* are obtained by dividing the treatment effect by the mean outcome for the control group. ***, **, * for statistical significance at the 1, 5 and 10 percent level of confidence.

Table 4b. The effects of external monitoring on test scores. Literacy – 5th graders.

Outcome variable	(1) Mean	(2) Bottom quartile	(3) Median	(4) Top quartile	(5) Standard deviation	(6) Cheating index	(7) Percent absent in the test	(8) Mean open-ended questions	(9) Mean close-ended questions
Monitored in year *t*	-2.690***	-3.381***	-2.768***	-2.145***	0.718***	-0.020***	-0.005**	-4.376***	-2.127***
	(0.155)	(0.196)	(0.171)	(0.139)	(0.054)	(0.001)	(0.002)	(0.208)	(0.146)
Monitored in year *t-1*	-0.312*	-0.297	-0.387**	-0.304**	0.024	-0.003**	0.001	-0.626***	-0.210
	(0.163)	(0.208)	(0.177)	(0.144)	(0.057)	(0.001)	(0.003)	(0.219)	(0.155)
Monitored in year *t-2*	0.135	0.192	0.096	0.126	0.009	0.001	-0.002	-0.036	0.180
	(0.158)	(0.201)	(0.172)	(0.137)	(0.058)	(0.001)	(0.003)	(0.214)	(0.150)
Observations	22,984	22,984	22,984	22,984	22,984	22,984	22,984	22,984	22,984
Other controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Randomization controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Mean for control group at *t*	64.27	54.15	65.53	75.41	15.07	0.040	0.149	64.49	64.13
Mean for control group at *t-1*	64.07	53.89	65.33	75.26	15.13	0.038	0.149	64.17	63.97
Mean for control group at *t-2*	63.99	53.80	65.25	75.19	15.14	0.038	0.149	64.12	63.89
% change for monitored at *t*	-0.041	-0.062	-0.042	-0.028	0.047	-0.499	-0.035	-0.067	-0.033
% change for monitored at *t-1*	-0.004	-0.005	-0.005	-0.004	0.001	-0.084	0.004	-0.009	-0.003
% change for monitored at *t-2*	0.002	0.003	0.001	0.001	0.001	0.019	-0.010	-0.001	0.002

Note: each regression includes randomization controls (region-by-year dummies and their interactions with current and lagged enrolment) and the other controls in vectors X, W and Z. Standard errors clustered by school within parentheses. Percent changes for monitored at *t, t-1* and *t-2* are obtained by dividing the treatment effect by the mean outcome for the control group. ***, **, * for statistical significance at the 1, 5 and 10 percent level of confidence.

Table 5a. The effects of external monitoring on test scores. Primary schools with 10 to 35 pupils in the grade. Math – 5th graders.

Outcome variable	(1) Mean	(2) Bottom quartile	(3) Median	(4) Top quartile	(5) Standard deviation	(6) Cheating index	(7) Percent absent in the test	(8) Mean open-ended questions	(9) Mean close-ended questions
Monitored in year *t*	-7.300***	-8.849***	-7.671***	-6.946***	1.046**	-0.063***	0.005	-8.520***	-6.459***
	(1.595)	(1.843)	(1.698)	(1.537)	(0.445)	(0.015)	(0.013)	(1.883)	(1.482)
Monitored in year *t-1*	-2.924*	-3.801**	-3.555**	-1.963	1.031**	-0.055***	-0.010	-3.119*	-2.635*
	(1.497)	(1.774)	(1.631)	(1.371)	(0.441)	(0.013)	(0.013)	(1.672)	(1.484)
Monitored in year *t-2*	1.357	1.937	1.191	0.821	-0.585	0.020	-0.008	1.408	1.205
	(1.192)	(1.491)	(1.282)	(1.041)	(0.458)	(0.019)	(0.010)	(1.488)	(1.169)
Observations	3,752	3,752	3,752	3,752	3,752	3,752	3,752	3,752	3,752
Other controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Randomization controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Mean for control group at *t*	66.01	56.88	66.64	75.80	13.57	0.090	0.098	66.72	63.85
Mean for control group at *t-1*	65.90	56.74	66.53	75.69	13.58	0.090	0.099	66.63	63.71
Mean for control group at *t-2*	65.78	56.59	66.42	75.61	13.62	0.088	0.098	66.55	63.58
% change for monitored at *t*	-0.111	-0.156	-0.115	-0.091	0.077	-0.694	0.055	-0.128	-0.101
% change for monitored at *t-1*	-0.044	-0.067	-0.053	-0.025	0.075	-0.615	-0.096	-0.046	-0.041
% change for monitored at *t-2*	0.020	0.034	0.017	0.010	-0.042	0.220	-0.085	0.021	0.019

Note: each regression includes randomization controls (region-by-year dummies and their interactions with current and lagged enrolment) and the other controls in vectors X, W and Z. Standard errors clustered by school within parentheses. Percent changes for monitored at *t, t-1* and *t-2* are obtained by dividing the treatment effect for the mean outcome for the control group for monitored at *t, t-1* and *t-2*, respectively. ***, **, * for statistical significance at the 1, 5 and 10 percent level of confidence.

Table 5b. The effects of external monitoring on test scores. Primary schools with 36 to 75 pupils in the grade. Math – 5th graders.

Outcome variable	(1) Mean	(2) Bottom quartile	(3) Median	(4) Top quartile	(5) Standard deviation	(6) Cheating index	(7) Percent absent in the test	(8) Mean open-ended questions	(9) Mean close-ended questions
Monitored in year *t*	-4.407***	-5.467***	-4.643***	-3.768***	0.980***	-0.026***	-0.017***	-5.685***	-3.272***
	(0.506)	(0.595)	(0.563)	(0.499)	(0.142)	(0.003)	(0.006)	(0.572)	(0.487)
Monitored in year *t-1*	-1.130**	-1.395**	-1.298**	-0.904*	0.200	-0.012***	0.001	-1.246**	-1.108**
	(0.498)	(0.598)	(0.545)	(0.477)	(0.147)	(0.004)	(0.007)	(0.587)	(0.468)
Monitored in year *t-2*	0.491	0.600	0.499	0.473	-0.086	0.005	0.003	0.510	0.488
	(0.493)	(0.589)	(0.539)	(0.474)	(0.148)	(0.004)	(0.007)	(0.600)	(0.452)
Observations	4,842	4,842	4,842	4,842	4,842	4,842	4,842	4,842	4,842
Other controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Randomization controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Mean for control group at *t*	61.76	51.28	62.34	72.81	15.17	0.040	0.164	61.60	60.50
Mean for control group at *t-1*	61.50	50.96	62.07	72.58	15.22	0.039	0.162	61.24	60.33
Mean for control group at *t-2*	61.25	50.64	61.80	72.37	15.29	0.037	0.162	61.13	59.98
% change for monitored at *t*	-0.071	-0.107	-0.074	-0.051	0.064	-0.655	-0.101	-0.092	-0.054
% change for monitored at *t-1*	-0.018	-0.027	-0.020	-0.012	0.013	-0.317	0.006	-0.020	-0.018
% change for monitored at *t-2*	0.008	0.011	0.008	0.006	-0.005	0.136	0.016	0.008	0.008

Note: each regression includes randomization controls (region-by-year dummies and their interactions with current and lagged enrolment) and the other controls in vectors X, W and Z. Standard errors clustered by school within parentheses. Percent changes for monitored at *t, t-1* and *t-2* are obtained by dividing the treatment effect for the mean outcome for the control group for monitored at *t, t-1* and *t-2*, respectively. ***, **, * for statistical significance at the 1, 5 and 10 percent level of confidence.

Table 5c. The effects of external monitoring on test scores. Primary schools with more than 75 pupils in the grade. Math – 5th graders.

Outcome variable	(1) Mean	(2) Bottom quartile	(3) Median	(4) Top quartile	(5) Standard deviation	(6) Cheating index	(7) Percent absent in the test	(8) Mean open-ended questions	(9) Mean close-ended questions
Monitored in year *t*	-2.570***	-3.064***	-2.778***	-2.206***	0.464***	-0.014***	-0.002	-3.239***	-1.955***
	(0.223)	(0.263)	(0.244)	(0.212)	(0.061)	(0.001)	(0.002)	(0.254)	(0.212)
Monitored in year *t-1*	-0.064	-0.111	-0.087	0.026	0.083	-0.002	-0.003	-0.140	-0.015
	(0.241)	(0.289)	(0.263)	(0.224)	(0.067)	(0.002)	(0.002)	(0.272)	(0.231)
Monitored in year *t-2*	0.354	0.359	0.384	0.340	-0.054	0.002	0.000	0.262	0.349
	(0.238)	(0.286)	(0.257)	(0.218)	(0.065)	(0.002)	(0.002)	(0.274)	(0.228)
Observations	9,096	9,096	9,096	9,096	9,096	9,096	9,096	9,096	9,096
Other controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Randomization controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Mean for control group at *t*	60.65	49.47	61.21	72.39	16.00	0.030	0.135	60.58	59.62
Mean for control group at *t-1*	60.35	49.11	60.90	72.13	16.05	0.029	0.135	60.19	59.41
Mean for control group at *t-2*	60.10	48.82	60.64	71.93	16.12	0.028	0.134	60.20	59.02
% change for monitored at *t*	-0.042	-0.061	-0.045	-0.030	0.029	-0.452	-0.016	-0.053	-0.032
% change for monitored at *t-1*	-0.001	-0.002	-0.001	0.001	0.005	-0.071	-0.020	-0.002	-0.001
% change for monitored at *t-2*	0.005	0.007	0.006	0.004	-0.003	0.063	0.001	0.004	0.005

Note: each regression includes randomization controls (region-by-year dummies and their interactions with current and lagged enrolment) and the other controls in vectors X, W and Z. Standard errors clustered by school within parentheses. Percent changes for monitored at *t, t-1* and *t-2* are obtained by dividing the treatment effect for the mean outcome for the control group for monitored at *t, t-1* and *t-2*, respectively. ***, **, * for statistical significance at the 1, 5 and 10 percent level of confidence.

Table 6a. The effects of external monitoring on test scores. Primary schools with 10 to 35 pupils in the grade. Literacy – 5th graders.

Outcome variable	(1) Mean	(2) Bottom quartile	(3) Median	(4) Top quartile	(5) Standard deviation	(6) Cheating index	(7) Percent absent in the test	(8) Mean open-ended questions	(9) Mean close-ended questions
Monitored in year *t*	-6.410***	-7.330***	-6.684***	-5.522***	1.483***	-0.065***	0.008	-9.874***	-5.228***
	(1.347)	(1.677)	(1.443)	(1.238)	(0.451)	(0.010)	(0.013)	(1.841)	(1.258)
Monitored in year *t-1*	-3.673***	-4.570***	-3.957***	-2.628**	1.277***	-0.039***	-0.006	-4.343***	-3.482***
	(1.235)	(1.617)	(1.295)	(1.079)	(0.470)	(0.014)	(0.013)	(1.646)	(1.206)
Monitored in year *t-2*	1.055	1.185	0.889	1.084	-0.215	0.011	-0.013	1.212	0.987
	(1.030)	(1.312)	(1.086)	(0.847)	(0.424)	(0.015)	(0.011)	(1.325)	(1.061)
Observations	3,752	3,752	3,752	3,752	3,752	3,752	3,752	3,752	3,752
Other controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Randomization controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Mean for control group at *t*	67.53	58.57	68.79	77.52	13.64	0.074	0.098	69.21	66.90
Mean for control group at *t-1*	67.45	58.49	68.71	77.44	13.65	0.074	0.098	69.09	66.83
Mean for control group at *t-2*	67.36	58.38	68.62	77.37	13.67	0.073	0.098	69.00	66.74
% change for monitored at *t*	-0.094	-0.125	-0.097	-0.071	0.109	-0.867	0.076	-0.143	-0.078
% change for monitored at *t-1*	-0.054	-0.078	-0.057	-0.033	0.093	-0.529	-0.056	-0.062	-0.052
% change for monitored at *t-2*	0.015	0.020	0.013	0.014	-0.015	0.146	-0.134	0.017	0.014

Note: each regression includes randomization controls (region-by-year dummies and their interactions with current and lagged enrolment) and the other controls in vectors X, W and Z. Standard errors clustered by school within parentheses. Percent changes for monitored at *t, t-1* and *t-2* are obtained by dividing the treatment effect for the mean outcome for the control group for monitored at *t, t-1* and *t-2*, respectively. ***, **, * for statistical significance at the 1, 5 and 10 percent level of confidence.

Table 6b. The effects of external monitoring on test scores. Primary schools with 36 to 75 pupils in the grade. Literacy – 5th graders.

Outcome variable	(1) Mean	(2) Bottom quartile	(3) Median	(4) Top quartile	(5) Standard deviation	(6) Cheating index	(7) Percent absent in the test	(8) Mean open-ended questions	(9) Mean close-ended questions
Monitored in year *t*	-2.929***	-3.991***	-2.991***	-2.036***	0.985***	-0.024***	-0.014**	-5.353***	-2.154***
	(0.378)	(0.484)	(0.424)	(0.354)	(0.134)	(0.002)	(0.007)	(0.518)	(0.357)
Monitored in year *t-1*	-0.758*	-0.867*	-0.849*	-0.695*	0.087	-0.008**	-0.003	-1.255**	-0.591
	(0.395)	(0.502)	(0.434)	(0.356)	(0.140)	(0.003)	(0.007)	(0.534)	(0.377)
Monitored in year *t-2*	0.338	0.433	0.213	0.340	-0.109	0.004	0.005	0.340	0.319
	(0.392)	(0.498)	(0.432)	(0.351)	(0.149)	(0.004)	(0.007)	(0.551)	(0.367)
Observations	4,842	4,842	4,842	4,842	4,842	4,842	4,842	4,842	4,842
Other controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Randomization controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Mean for control group at *t*	63.74	53.55	64.94	74.89	15.10	0.035	0.167	63.83	63.63
Mean for control group at *t-1*	63.57	53.30	64.78	74.79	15.18	0.033	0.166	63.50	63.51
Mean for control group at *t-2*	63.40	53.11	64.61	74.64	15.21	0.032	0.165	63.34	63.34
% change for monitored at *t*	-0.046	-0.074	-0.046	-0.027	0.065	-0.691	-0.084	-0.083	-0.033
% change for monitored at *t-1*	-0.011	-0.016	-0.013	-0.009	0.005	-0.239	-0.018	-0.019	-0.009
% change for monitored at *t-2*	0.005	0.008	0.003	0.004	-0.007	0.120	0.027	0.005	0.005

Note: each regression includes randomization controls (region-by-year dummies and their interactions with current and lagged enrolment) and the other controls in vectors X, W and Z. Standard errors clustered by school within parentheses. Percent changes for monitored at *t, t-1* and *t-2* are obtained by dividing the treatment effect for the mean outcome for the control group for monitored at *t, t-1* and *t-2*, respectively. ***, **, * for statistical significance at the 1, 5 and 10 percent level of confidence.

Table 6c. The effects of external monitoring on test scores. Primary schools with more than 75 pupils in the grade. Literacy – 5th graders.

Outcome variable	(1) Mean	(2) Bottom quartile	(3) Median	(4) Top quartile	(5) Standard deviation	(6) Cheating index	(7) Percent absent in the test	(8) Mean open-ended questions	(9) Mean close-ended questions
Monitored in year *t*	-2.096***	-2.610***	-2.144***	-1.742***	0.515***	-0.013***	-0.004	-3.512***	-1.613***
	(0.178)	(0.227)	(0.195)	(0.159)	(0.063)	(0.001)	(0.003)	(0.245)	(0.167)
Monitored in year *t-1*	0.020	0.095	0.021	0.043	0.030	-0.001	-0.003	-0.319	0.136
	(0.194)	(0.248)	(0.209)	(0.171)	(0.066)	(0.001)	(0.003)	(0.261)	(0.182)
Monitored in year *t-2*	0.307*	0.365	0.339*	0.285*	-0.023	0.001	-0.004	0.230	0.310*
	(0.186)	(0.237)	(0.200)	(0.162)	(0.064)	(0.002)	(0.003)	(0.250)	(0.176)
Observations	9,096	9,096	9,096	9,096	9,096	9,096	9,096	9,096	9,096
Other controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Randomization controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Mean for control group at *t*	63.33	52.72	64.63	75.02	15.73	0.026	0.144	62.79	63.44
Mean for control group at *t-1*	63.08	52.39	64.38	74.82	15.79	0.025	0.144	62.40	63.24
Mean for control group at *t-2*	62.97	52.27	64.27	74.73	15.82	0.025	0.144	62.31	63.13
% change for monitored at *t*	-0.033	-0.049	-0.033	-0.023	0.032	-0.488	-0.028	-0.055	-0.025
% change for monitored at *t-1*	0.001	0.001	0.001	0.001	0.001	-0.037	-0.019	-0.005	0.002
% change for monitored at *t-2*	0.004	0.006	0.005	0.003	-0.001	0.039	-0.025	0.003	0.004

Note: each regression includes randomization controls (region-by-year dummies and their interactions with current and lagged enrolment) and the other controls in vectors X, W and Z. Standard errors clustered by school within parentheses. Percent changes for monitored at *t, t-1* and *t-2* are obtained by dividing the treatment effect for the mean outcome for the control group for monitored at *t, t-1* and *t-2*, respectively. ***, **, * for statistical significance at the 1, 5 and 10 percent level of confidence.

Table 7a. The effects of external monitoring on test scores. Primary schools in Northern and Central Italy. Math – 5th graders.

Outcome variable	(1) Mean	(2) Bottom quartile	(3) Median	(4) Top quartile	(5) Standard deviation	(6) Cheating index	(7) Percent absent in the test	(8) Mean open-ended questions	(9) Mean close-ended questions
Monitored in year *t*	-2.225***	-2.725***	-2.365***	-1.884***	0.467***	-0.011***	-0.005*	-2.965***	-1.578***
	(0.183)	(0.220)	(0.208)	(0.184)	(0.056)	(0.001)	(0.002)	(0.212)	(0.175)
Monitored in year *t-1*	-0.476**	-0.580**	-0.633***	-0.401**	0.107*	-0.003***	-0.001	-0.653***	-0.301
	(0.213)	(0.256)	(0.236)	(0.203)	(0.065)	(0.001)	(0.003)	(0.245)	(0.203)
Monitored in year *t-2*	0.055	-0.032	0.065	0.149	0.067	-0.001	-0.001	-0.041	0.106
	(0.185)	(0.225)	(0.207)	(0.179)	(0.058)	(0.001)	(0.003)	(0.224)	(0.176)
Observations	14,515	14,515	14,515	14,515	14,515	14,515	14,515	14,515	14,515
Other controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Randomization controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Mean for control group at *t*	61.11	49.84	61.66	72.91	16.06	0.024	0.133	60.38	60.54
Mean for control group at *t-1*	60.98	49.68	61.53	72.80	16.09	0.024	0.133	60.20	60.44
Mean for control group at *t-2*	60.80	49.47	61.34	72.64	16.13	0.023	0.132	60.19	60.18
% change for monitored at *t*	-0.036	-0.054	-0.038	-0.025	0.029	-0.456	-0.035	-0.049	-0.026
% change for monitored at *t-1*	-0.007	-0.011	-0.010	-0.005	0.006	-0.137	-0.010	-0.010	-0.004
% change for monitored at *t-2*	0.001	-0.001	0.001	0.002	0.004	-0.038	-0.004	-0.001	0.001

Note: each regression includes randomization controls (region-by-year dummies and their interactions with current and lagged enrolment) and the other controls in vectors X, W and Z. Standard errors clustered by school within parentheses. Percent changes for monitored at *t, t-1* and *t-2* are obtained by dividing the treatment effect for the mean outcome for the control group for monitored at *t, t-1* and *t-2*, respectively. ***, **, * for statistical significance at the 1, 5 and 10 percent level of confidence.

Table 7b. The effects of external monitoring on test scores. Primary schools in Southern Italy. Math – 5th graders.

Outcome variable	(1) Mean	(2) Bottom quartile	(3) Median	(4) Top quartile	(5) Standard deviation	(6) Cheating index	(7) Percent absent in the test	(8) Mean open-ended questions	(9) Mean close-ended questions
Monitored in year *t*	-4.856***	-5.783***	-5.218***	-4.304***	0.828***	-0.031***	-0.005	-5.978***	-3.854***
	(0.380)	(0.448)	(0.409)	(0.356)	(0.100)	(0.003)	(0.004)	(0.437)	(0.360)
Monitored in year *t-1*	-0.377	-0.510	-0.407	-0.166	0.154	-0.008**	0.004	-0.429	-0.398
	(0.386)	(0.460)	(0.418)	(0.354)	(0.108)	(0.003)	(0.005)	(0.442)	(0.368)
Monitored in year *t-2*	0.359	0.468	0.243	0.200	-0.120	0.005	0.002	0.292	0.369
	(0.404)	(0.484)	(0.434)	(0.366)	(0.118)	(0.004)	(0.005)	(0.474)	(0.379)
Observations	8,469	8,469	8,469	8,469	8,469	8,469	8,469	8,469	8,469
Other controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Randomization controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Mean for control group at *t*	63.88	54.79	64.54	73.63	13.59	0.086	0.163	65.45	61.00
Mean for control group at *t-1*	63.39	54.22	64.02	73.18	13.66	0.083	0.162	64.83	60.62
Mean for control group at *t-2*	63.27	54.05	63.91	73.11	13.71	0.082	0.162	64.82	60.43
% change for monitored at *t*	-0.076	-0.106	-0.080	-0.058	0.061	-0.362	-0.028	-0.091	-0.063
% change for monitored at *t-1*	-0.005	-0.009	-0.006	-0.002	0.011	-0.094	0.023	-0.006	-0.006
% change for monitored at *t-2*	0.005	0.008	0.003	0.002	-0.008	0.060	0.011	0.004	0.006

Note: each regression includes randomization controls (region-by-year dummies and their interactions with current and lagged enrolment) and the other controls in vectors X, W and Z. Standard errors clustered by school within parentheses. Percent changes for monitored at *t, t-1* and *t-2* are obtained by dividing the treatment effect for the mean outcome for the control group for monitored at *t, t-1* and *t-2*, respectively. ***, **, * for statistical significance at the 1, 5 and 10 percent level of confidence.

Table 8a. The effects of external monitoring on test scores. Primary schools in Northern and Central Italy. Literacy – 5th graders.

Outcome variable	(1) Mean	(2) Bottom quartile	(3) Median	(4) Top quartile	(5) Standard deviation	(6) Cheating index	(7) Percent absent in the test	(8) Mean open-ended questions	(9) Mean close-ended questions
Monitored in year *t*	-1.766***	-2.275***	-1.707***	-1.394***	0.520***	-0.012***	-0.005*	-3.405***	-1.223***
	(0.146)	(0.191)	(0.166)	(0.135)	(0.057)	(0.001)	(0.003)	(0.212)	(0.136)
Monitored in year *t-1*	-0.352**	-0.410*	-0.433**	-0.288*	0.087	-0.003***	-0.000	-0.908***	-0.169
	(0.164)	(0.213)	(0.183)	(0.149)	(0.063)	(0.001)	(0.003)	(0.231)	(0.154)
Monitored in year *t-2*	0.048	0.089	0.013	0.083	0.048	-0.000	-0.004	-0.166	0.104
	(0.154)	(0.201)	(0.172)	(0.136)	(0.060)	(0.001)	(0.003)	(0.222)	(0.144)
Observations	14,515	14,515	14,515	14,515	14,515	14,515	14,515	14,515	14,515
Other controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Randomization controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Mean for control group at *t*	64.03	53.49	65.36	75.64	15.62	0.021	0.141	62.96	64.31
Mean for control group at *t-1*	63.93	53.35	65.27	75.56	15.65	0.021	0.140	62.77	64.24
Mean for control group at *t-2*	63.86	53.26	65.20	75.51	15.67	0.020	0.140	62.71	64.17
% change for monitored at *t*	-0.027	-0.042	-0.026	-0.018	0.033	-0.541	-0.034	-0.054	-0.019
% change for monitored at *t-1*	-0.005	-0.007	-0.006	-0.003	0.005	-0.154	-0.003	-0.014	-0.002
% change for monitored at *t-2*	0.001	0.001	0.001	0.001	0.003	-0.019	-0.029	-0.002	0.001

Note: each regression includes randomization controls (region-by-year dummies and their interactions with current and lagged enrolment) and the other controls in vectors X, W and Z. Standard errors clustered by school within parentheses. Percent changes for monitored at *t, t-1* and *t-2* are obtained by dividing the treatment effect for the mean outcome for the control group for monitored at *t, t-1* and *t-2*, respectively. ***, **, * for statistical significance at the 1, 5 and 10 percent level of confidence.

Table 8b. The effects of external monitoring on test scores. Primary schools in Southern Italy. Literacy – 5th graders.

Outcome variable	(1) Mean	(2) Bottom quartile	(3) Median	(4) Top quartile	(5) Standard deviation	(6) Cheating index	(7) Percent absent in the test	(8) Mean open-ended questions	(9) Mean close-ended questions
Monitored in year *t*	-3.895***	-4.819***	-4.164***	-3.131***	0.960***	-0.031***	-0.006	-5.618***	-3.313***
	(0.299)	(0.373)	(0.324)	(0.267)	(0.097)	(0.002)	(0.004)	(0.392)	(0.283)
Monitored in year *t-1*	-0.246	-0.115	-0.306	-0.324	-0.064	-0.003	0.002	-0.234	-0.255
	(0.314)	(0.395)	(0.334)	(0.276)	(0.104)	(0.003)	(0.005)	(0.409)	(0.300)
Monitored in year *t-2*	0.277	0.364	0.230	0.203	-0.053	0.003	0.002	0.188	0.299
	(0.318)	(0.400)	(0.342)	(0.274)	(0.114)	(0.003)	(0.005)	(0.419)	(0.306)
Observations	8,469	8,469	8,469	8,469	8,469	8,469	8,469	8,469	8,469
Other controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Randomization controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Mean for control group at *t*	64.70	55.34	65.82	74.99	14.08	0.073	0.165	67.22	63.81
Mean for control group at *t-1*	64.34	54.86	65.43	74.72	14.20	0.070	0.163	66.65	63.51
Mean for control group at *t-2*	64.24	54.75	65.33	74.63	14.21	0.069	0.163	66.60	63.39
% change for monitored at *t*	-0.060	-0.087	-0.063	-0.041	0.068	-0.421	-0.039	-0.083	-0.051
% change for monitored at *t-1*	-0.003	-0.002	-0.004	-0.004	-0.004	-0.044	0.013	-0.003	-0.004
% change for monitored at *t-2*	0.004	0.006	0.003	0.002	-0.003	0.036	0.012	0.002	0.004

Note: each regression includes randomization controls (region-by-year dummies and their interactions with current and lagged enrolment) and the other controls in vectors X, W and Z. Standard errors clustered by school within parentheses. Percent changes for monitored at *t, t-1* and *t-2* are obtained by dividing the treatment effect for the mean outcome for the control group for monitored at *t, t-1* and *t-2*, respectively. ***, **, * for statistical significance at the 1, 5 and 10 percent level of confidence.

Table 9a. The effects of external monitoring on test scores. Primary schools with high social capital (blood donation). Math – 5th graders.

Outcome variable	(1) Mean	(2) Bottom quartile	(3) Median	(4) Top quartile	(5) Standard deviation	(6) Cheating index	(7) Percent absent in the test	(8) Mean open-ended questions	(9) Mean close-ended questions
Monitored in year *t*	-2.506***	-3.006***	-2.655***	-2.215***	0.450***	-0.013***	-0.004*	-3.188***	-1.894***
	(0.228)	(0.274)	(0.255)	(0.222)	(0.066)	(0.001)	(0.003)	(0.261)	(0.214)
Monitored in year *t-1*	-0.473*	-0.675**	-0.594**	-0.327	0.170**	-0.003**	0.004	-0.743***	-0.266
	(0.248)	(0.300)	(0.275)	(0.235)	(0.075)	(0.001)	(0.003)	(0.287)	(0.237)
Monitored in year *t-2*	-0.036	-0.162	-0.053	0.092	0.095	-0.001	-0.002	-0.245	0.096
	(0.225)	(0.274)	(0.251)	(0.217)	(0.071)	(0.001)	(0.003)	(0.272)	(0.211)
Observations	11,344	11,344	11,344	11,344	11,344	11,344	11,344	11,344	11,344
Other controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Randomization controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Mean for control group at *t*	61.05	49.72	61.60	72.93	16.12	0.024	0.130	60.26	60.53
Mean for control group at *t-1*	60.90	49.55	61.45	72.78	16.14	0.023	0.129	60.05	60.42
Mean for control group at *t-2*	60.72	49.33	61.26	72.64	16.19	0.023	0.129	60.06	60.15
% change for monitored at *t*	-0.041	-0.060	-0.043	-0.030	0.027	-0.523	-0.034	-0.052	-0.031
% change for monitored at *t-1*	-0.007	-0.013	-0.009	-0.004	0.010	-0.149	0.030	-0.012	-0.004
% change for monitored at *t-2*	-0.001	-0.003	-0.001	0.001	0.005	-0.052	-0.015	-0.004	0.001

Note: each regression includes randomization controls (region-by-year dummies and their interactions with current and lagged enrolment) and the other controls in vectors X, W and Z. Standard errors clustered by school within parentheses. Percent changes for monitored at *t, t-1* and *t-2* are obtained by dividing the treatment effect for the mean outcome for the control group for monitored at *t, t-1* and *t-2*, respectively. ***, **, * for statistical significance at the 1, 5 and 10 percent level of confidence.

Table 9b. The effects of external monitoring on test scores. Primary schools with low social capital (blood donation). Math – 5th graders.

Outcome variable	(1) Mean	(2) Bottom quartile	(3) Median	(4) Top quartile	(5) Standard deviation	(6) Cheating index	(7) Percent absent in the test	(8) Mean open-ended questions	(9) Mean close-ended questions
Monitored in year *t*	-4.185***	-5.070***	-4.483***	-3.607***	0.807***	-0.027***	-0.004	-5.259***	-3.242***
	(0.307)	(0.361)	(0.333)	(0.290)	(0.083)	(0.002)	(0.003)	(0.353)	(0.293)
Monitored in year *t-1*	-0.440	-0.510	-0.534	-0.328	0.103	-0.007***	-0.002	-0.439	-0.443
	(0.316)	(0.377)	(0.343)	(0.290)	(0.090)	(0.003)	(0.004)	(0.361)	(0.302)
Monitored in year *t-2*	0.375	0.466	0.328	0.244	-0.096	0.003	0.003	0.416	0.307
	(0.316)	(0.379)	(0.340)	(0.287)	(0.092)	(0.003)	(0.004)	(0.371)	(0.298)
Observations	11,509	11,509	11,509	11,509	11,509	11,509	11,509	11,509	11,509
Other controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Randomization controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Mean for control group at *t*	63.19	53.56	63.82	73.42	14.20	0.070	0.158	64.20	60.87
Mean for control group at *t-1*	62.82	53.12	63.43	73.09	14.27	0.068	0.157	63.73	60.59
Mean for control group at *t-2*	62.69	52.96	63.30	73.00	14.31	0.067	0.157	63.72	60.38
% change for monitored at *t*	-0.066	-0.094	-0.070	-0.049	0.056	-0.382	-0.026	-0.081	-0.053
% change for monitored at *t-1*	-0.007	-0.009	-0.008	-0.004	0.007	-0.101	-0.015	-0.006	-0.007
% change for monitored at *t-2*	0.005	0.008	0.005	0.003	-0.006	0.048	0.016	0.006	0.005

Note: each regression includes randomization controls (region-by-year dummies and their interactions with current and lagged enrolment) and the other controls in vectors X, W and Z. Standard errors clustered by school within parentheses. Percent changes for monitored at *t, t-1* and *t-2* are obtained by dividing the treatment effect for the mean outcome for the control group for monitored at *t, t-1* and *t-2*, respectively. ***, **, * for statistical significance at the 1, 5 and 10 percent level of confidence.

Table 9c. The effects of external monitoring on test scores. Primary schools with high social capital (blood donation). Literacy – 5th graders.

Outcome variable	(1) Mean	(2) Bottom quartile	(3) Median	(4) Top quartile	(5) Standard deviation	(6) Cheating index	(7) Percent absent in the test	(8) Mean open-ended questions	(9) Mean close-ended questions
Monitored in year *t*	-1.944***	-2.553***	-1.905***	-1.500***	0.559***	-0.012***	-0.006**	-3.452***	-1.451***
	(0.175)	(0.230)	(0.195)	(0.157)	(0.068)	(0.001)	(0.003)	(0.248)	(0.165)
Monitored in year *t-1*	-0.371*	-0.481*	-0.511**	-0.241	0.165**	-0.003**	0.004	-0.914***	-0.197
	(0.194)	(0.255)	(0.217)	(0.175)	(0.074)	(0.001)	(0.003)	(0.263)	(0.185)
Monitored in year *t-2*	0.036	0.069	0.049	0.015	0.001	0.001	-0.003	-0.232	0.108
	(0.180)	(0.232)	(0.198)	(0.159)	(0.069)	(0.001)	(0.003)	(0.259)	(0.168)
Observations	11,344	11,344	11,344	11,344	11,344	11,344	11,344	11,344	11,344
Other controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Randomization controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Mean for control group at *t*	63.88	53.29	65.21	75.52	15.66	0.021	0.136	62.74	64.18
Mean for control group at *t-1*	63.76	53.13	65.10	75.43	15.69	0.020	0.135	62.55	64.09
Mean for control group at *t-2*	63.69	53.03	65.02	75.38	15.72	0.020	0.136	62.48	64.01
% change for monitored at *t*	-0.030	-0.047	-0.029	-0.019	0.035	-0.556	-0.047	-0.055	-0.022
% change for monitored at *t-1*	-0.005	-0.009	-0.007	-0.003	0.010	-0.154	0.027	-0.014	-0.003
% change for monitored at *t-2*	0.001	0.001	0.001	0.001	0.001	0.029	-0.021	-0.003	0.001

Note: each regression includes randomization controls (region-by-year dummies and their interactions with current and lagged enrolment) and the other controls in vectors X, W and Z. Standard errors clustered by school within parentheses. Percent changes for monitored at *t, t-1* and *t-2* are obtained by dividing the treatment effect for the mean outcome for the control group for monitored at *t, t-1* and *t-2*, respectively. ***, **, * for statistical significance at the 1, 5 and 10 percent level of confidence.

Table 9d. The effects of external monitoring on test scores. Primary schools with low social capital (blood donation). Literacy – 5th graders.

Outcome variable	(1) Mean	(2) Bottom quartile	(3) Median	(4) Top quartile	(5) Standard deviation	(6) Cheating index	(7) Percent absent in the test	(8) Mean open-ended questions	(9) Mean close-ended questions
Monitored in year *t*	-3.373***	-4.136***	-3.549***	-2.728***	0.869***	-0.027***	-0.004	-5.204***	-2.751***
	(0.246)	(0.306)	(0.269)	(0.221)	(0.081)	(0.002)	(0.004)	(0.324)	(0.232)
Monitored in year *t-1*	-0.319	-0.204	-0.360	-0.411*	-0.082	-0.004	-0.002	-0.398	-0.292
	(0.253)	(0.318)	(0.270)	(0.223)	(0.086)	(0.002)	(0.004)	(0.336)	(0.241)
Monitored in year *t-2*	0.248	0.304	0.160	0.259	0.013	0.001	0.000	0.194	0.260
	(0.252)	(0.320)	(0.273)	(0.217)	(0.092)	(0.003)	(0.004)	(0.335)	(0.242)
Observations	11,509	11,509	11,509	11,509	11,509	11,509	11,509	11,509	11,509
Other controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Randomization controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Mean for control group at *t*	64.68	55.05	65.86	75.30	14.47	0.059	0.162	66.28	64.08
Mean for control group at *t-1*	64.41	54.69	65.58	75.10	14.56	0.057	0.162	65.84	63.87
Mean for control group at *t-2*	64.32	54.60	65.49	75.00	14.56	0.057	0.161	65.80	63.77
% change for monitored at *t*	-0.052	-0.075	-0.053	-0.036	0.060	-0.456	-0.023	-0.078	-0.042
% change for monitored at *t-1*	-0.004	-0.003	-0.005	-0.005	-0.005	-0.063	-0.015	-0.006	-0.004
% change for monitored at *t-2*	0.003	0.005	0.002	0.003	0.001	0.017	0.001	0.002	0.004

Note: each regression includes randomization controls (region-by-year dummies and their interactions with current and lagged enrolment) and the other controls in vectors X, W and Z. Standard errors clustered by school within parentheses. Percent changes for monitored at *t, t-1* and *t-2* are obtained by dividing the treatment effect for the mean outcome for the control group for monitored at *t, t-1* and *t-2*, respectively. ***, **, * for statistical significance at the 1, 5 and 10 percent level of confidence.

Table 9e. The effects of external monitoring on test scores. Primary schools with high social capital (referenda turnout). Math – 5th graders.

Outcome variable	(1) Mean	(2) Bottom quartile	(3) Median	(4) Top quartile	(5) Standard deviation	(6) Cheating index	(7) Percent absent in the test	(8) Mean open-ended questions	(9) Mean close-ended questions
Monitored in year *t*	-2.011***	-2.494***	-2.123***	-1.674***	0.422***	-0.009***	-0.005*	-2.688***	-1.383***
	(0.213)	(0.257)	(0.241)	(0.213)	(0.066)	(0.001)	(0.003)	(0.243)	(0.207)
Monitored in year *t-1*	-0.535**	-0.673**	-0.636**	-0.513**	0.127	-0.003**	-0.002	-0.665**	-0.426*
	(0.242)	(0.295)	(0.267)	(0.230)	(0.077)	(0.001)	(0.003)	(0.283)	(0.226)
Monitored in year *t-2*	0.105	0.066	0.114	0.175	0.019	-0.001	-0.004	0.029	0.105
	(0.211)	(0.257)	(0.239)	(0.208)	(0.069)	(0.001)	(0.003)	(0.257)	(0.201)
Observations	9,990	9,990	9,990	9,990	9,990	9,990	9,990	9,990	9,990
Other controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Randomization controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Mean for control group at *t*	60.90	49.41	61.46	72.92	16.31	0.019	0.124	59.85	60.62
Mean for control group at *t-1*	60.78	49.26	61.35	72.84	16.34	0.018	0.123	59.70	60.55
Mean for control group at *t-2*	60.59	49.03	61.15	72.67	16.39	0.018	0.123	59.68	60.27
% change for monitored at *t*	-0.033	-0.050	-0.034	-0.023	0.025	-0.472	-0.037	-0.044	-0.022
% change for monitored at *t-1*	-0.008	-0.013	-0.010	-0.007	0.007	-0.142	-0.016	-0.011	-0.007
% change for monitored at *t-2*	0.001	0.001	0.001	0.002	0.001	-0.056	-0.030	0.001	0.001

Note: each regression includes randomization controls (region-by-year dummies and their interactions with current and lagged enrolment) and the other controls in vectors X, W and Z. Standard errors clustered by school within parentheses. Percent changes for monitored at *t, t-1* and *t-2* are obtained by dividing the treatment effect for the mean outcome for the control group for monitored at *t, t-1* and *t-2*, respectively. ***, **, * for statistical significance at the 1, 5 and 10 percent level of confidence.

Table 9f. The effects of external monitoring on test scores. Primary schools with low social capital (referenda turnout). Math – 5th graders.

Outcome variable	(1) Mean	(2) Bottom quartile	(3) Median	(4) Top quartile	(5) Standard deviation	(6) Cheating index	(7) Percent absent in the test	(8) Mean open-ended questions	(9) Mean close-ended questions
Monitored in year *t*	-4.236***	-5.056***	-4.537***	-3.729***	0.758***	-0.027***	-0.004	-5.275***	-3.331***
	(0.288)	(0.340)	(0.313)	(0.272)	(0.077)	(0.002)	(0.003)	(0.332)	(0.272)
Monitored in year *t-1*	-0.379	-0.486	-0.483	-0.182	0.126	-0.007***	0.002	-0.506	-0.294
	(0.300)	(0.357)	(0.326)	(0.276)	(0.084)	(0.002)	(0.004)	(0.343)	(0.287)
Monitored in year *t-2*	0.256	0.271	0.189	0.198	-0.026	0.003	0.003	0.174	0.307
	(0.301)	(0.361)	(0.324)	(0.274)	(0.088)	(0.003)	(0.004)	(0.354)	(0.282)
Observations	12,863	12,863	12,863	12,863	12,863	12,863	12,863	12,863	12,863
Other controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Randomization controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Mean for control group at *t*	63.09	53.41	63.70	73.37	14.25	0.069	0.160	64.10	60.76
Mean for control group at *t-1*	62.71	52.96	63.30	73.02	14.31	0.067	0.159	63.62	60.47
Mean for control group at *t-2*	62.58	52.80	63.17	72.93	14.35	0.066	0.159	63.62	60.27
% change for monitored at *t*	-0.067	-0.094	-0.071	-0.050	0.053	-0.389	-0.024	-0.082	-0.054
% change for monitored at *t-1*	-0.006	-0.009	-0.007	-0.002	0.008	-0.105	0.014	-0.007	-0.004
% change for monitored at *t-2*	0.004	0.005	0.003	0.002	-0.001	0.043	0.021	0.002	0.005

Note: each regression includes randomization controls (region-by-year dummies and their interactions with current and lagged enrolment) and the other controls in vectors X, W and Z. Standard errors clustered by school within parentheses. Percent changes for monitored at *t, t-1* and *t-2* are obtained by dividing the treatment effect for the mean outcome for the control group for monitored at *t, t-1* and *t-2*, respectively. ***, **, * for statistical significance at the 1, 5 and 10 percent level of confidence.

Table 9g. The effects of external monitoring on test scores. Primary schools with high social capital (referenda turnout). Literacy – 5th graders.

Outcome variable	(1) Mean	(2) Bottom quartile	(3) Median	(4) Top quartile	(5) Standard deviation	(6) Cheating index	(7) Percent absent in the test	(8) Mean open-ended questions	(9) Mean close-ended questions
Monitored in year *t*	-1.529***	-1.978***	-1.451***	-1.186***	0.464***	-0.010***	-0.005*	-3.075***	-1.016***
	(0.169)	(0.225)	(0.192)	(0.155)	(0.068)	(0.001)	(0.003)	(0.245)	(0.158)
Monitored in year *t-1*	-0.405**	-0.498**	-0.522**	-0.306*	0.149**	-0.003***	-0.002	-0.935***	-0.226
	(0.184)	(0.242)	(0.208)	(0.168)	(0.074)	(0.001)	(0.003)	(0.263)	(0.172)
Monitored in year *t-2*	0.236	0.335	0.205	0.212	-0.013	0.000	-0.006*	0.123	0.266
	(0.175)	(0.232)	(0.195)	(0.152)	(0.070)	(0.001)	(0.003)	(0.252)	(0.163)
Observations	9,990	9,990	9,990	9,990	9,990	9,990	9,990	9,990	9,990
Other controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Randomization controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Mean for control group at *t*	63.80	53.09	65.16	75.60	15.84	0.017	0.130	62.34	64.21
Mean for control group at *t-1*	63.71	52.98	65.08	75.53	15.87	0.016	0.130	62.17	64.14
Mean for control group at *t-2*	63.63	52.87	65.01	75.47	15.90	0.016	0.130	62.08	64.07
% change for monitored at *t*	-0.024	-0.037	-0.022	-0.015	0.029	-0.551	-0.040	-0.049	-0.015
% change for monitored at *t-1*	-0.006	-0.009	-0.008	-0.004	0.009	-0.179	-0.018	-0.015	-0.003
% change for monitored at *t-2*	0.003	0.006	0.003	0.002	-0.001	0.027	-0.044	0.001	0.004

Note: each regression includes randomization controls (region-by-year dummies and their interactions with current and lagged enrolment) and the other controls in vectors X, W and Z. Standard errors clustered by school within parentheses. Percent changes for monitored at *t, t-1* and *t-2* are obtained by dividing the treatment effect for the mean outcome for the control group for monitored at *t, t-1* and *t-2*, respectively. ***, **, * for statistical significance at the 1, 5 and 10 percent level of confidence.

Table 9h. The effects of external monitoring on test scores. Primary schools with low social capital (referenda turnout). Literacy – 5th graders.

Outcome variable	(1) Mean	(2) Bottom quartile	(3) Median	(4) Top quartile	(5) Standard deviation	(6) Cheating index	(7) Percent absent in the test	(8) Mean open-ended questions	(9) Mean close-ended questions
Monitored in year *t*	-3.434***	-4.274***	-3.613***	-2.754***	0.878***	-0.027***	-0.005	-5.193***	-2.842***
	(0.228)	(0.285)	(0.249)	(0.205)	(0.075)	(0.002)	(0.003)	(0.302)	(0.215)
Monitored in year *t-1*	-0.255	-0.149	-0.319	-0.332	-0.075	-0.003	0.002	-0.403	-0.212
	(0.242)	(0.305)	(0.259)	(0.213)	(0.081)	(0.002)	(0.004)	(0.319)	(0.231)
Monitored in year *t-2*	0.106	0.119	0.066	0.111	0.026	0.001	0.002	-0.080	0.154
	(0.239)	(0.301)	(0.259)	(0.207)	(0.086)	(0.002)	(0.004)	(0.320)	(0.228)
Observations	12,863	12,863	12,863	12,863	12,863	12,863	12,863	12,863	12,863
Other controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Randomization controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Mean for control group at *t*	64.66	55.01	65.83	75.26	14.45	0.058	0.164	66.23	64.07
Mean for control group at *t-1*	64.38	54.64	65.55	75.06	14.54	0.056	0.163	65.79	63.85
Mean for control group at *t-2*	64.30	54.55	65.45	74.97	14.54	0.056	0.163	65.75	63.75
% change for monitored at *t*	-0.053	-0.077	-0.054	-0.036	0.060	-0.455	-0.028	-0.078	-0.044
% change for monitored at *t-1*	-0.003	-0.002	-0.004	-0.004	-0.005	-0.060	0.013	-0.006	-0.003
% change for monitored at *t-2*	0.001	0.002	0.001	0.001	0.001	0.015	0.010	-0.001	0.002

Note: each regression includes randomization controls (region-by-year dummies and their interactions with current and lagged enrolment) and the other controls in vectors X, W and Z. Standard errors clustered by school within parentheses. Percent changes for monitored at *t, t-1* and *t-2* are obtained by dividing the treatment effect for the mean outcome for the control group for monitored at *t, t-1* and *t-2*, respectively. ***, **, * for statistical significance at the 1, 5 and 10 percent level of confidence.

Table 10a. The effects of external monitoring on test scores. Middle schools. Math – 8th graders.

Outcome variable	(1) Mean	(2) Bottom quartile	(3) Median	(4) Top quartile	(5) Standard deviation	(6) Cheating index	(7) Percent absent in the test	(8) Mean open-ended questions	(9) Mean close-ended questions
Monitored in year *t*	-0.502***	-0.576***	-0.600***	-0.491***	0.074**	-0.000	0.002	-0.649***	-0.431***
	(0.113)	(0.129)	(0.128)	(0.120)	(0.036)	(0.001)	(0.001)	(0.123)	(0.116)
Monitored in year *t-1*	0.174	0.195	0.183	0.208*	-0.012	0.001	0.001	0.177	0.190
	(0.115)	(0.132)	(0.128)	(0.120)	(0.037)	(0.001)	(0.001)	(0.128)	(0.116)
Monitored in year *t-2*	0.033	0.051	0.047	0.046	0.004	0.001	-0.000	0.006	0.074
	(0.117)	(0.135)	(0.131)	(0.123)	(0.037)	(0.001)	(0.001)	(0.130)	(0.118)
Observations	20,205	20,205	20,205	20,205	20,205	20,205	20,205	20,205	20,205
Other controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Randomization controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Mean for control group at *t*	57.36	45.42	57.18	69.27	16.48	0.036	0.089	53.33	59.47
Mean for control group at *t-1*	57.24	45.29	57.04	69.14	16.49	0.035	0.089	53.19	59.36
Mean for control group at *t-2*	57.30	45.36	57.10	69.20	16.48	0.035	0.090	53.26	59.41
% change for monitored at *t*	-0.008	-0.012	-0.010	-0.007	0.004	-0.012	0.018	-0.012	-0.007
% change for monitored at *t-1*	0.003	0.004	0.003	0.003	-0.001	0.027	0.015	0.003	0.003
% change for monitored at *t-2*	0.001	0.001	0.001	0.001	0.001	0.030	-0.002	0.001	0.001

Note: each regression includes randomization controls (region-by-year dummies and their interactions with current and lagged enrolment) and the other controls in vectors X, W and Z. Standard errors clustered by school within parentheses. Percent changes for monitored at *t, t-1* and *t-2* are obtained by dividing the treatment effect for the mean outcome for the control group for monitored at *t, t-1* and *t-2*, respectively. ***, **, * for statistical significance at the 1, 5 and 10 percent level of confidence.

Table 10b. The effects of external monitoring on test scores. Middle schools. Literacy – 8th graders.

Outcome variable	(1) Mean	(2) Bottom quartile	(3) Median	(4) Top quartile	(5) Standard deviation	(6) Cheating index	(7) Percent absent in the test	(8) Mean open-ended questions	(9) Mean close-ended questions
Monitored in year *t*	-0.264***	-0.288***	-0.241***	-0.261***	0.091**	0.001	0.002	-0.329***	-0.250***
	(0.086)	(0.109)	(0.093)	(0.080)	(0.036)	(0.001)	(0.001)	(0.104)	(0.089)
Monitored in year *t-1*	0.144	0.167	0.128	0.151*	-0.057	0.001	0.001	0.074	0.178*
	(0.090)	(0.115)	(0.097)	(0.083)	(0.037)	(0.001)	(0.001)	(0.107)	(0.093)
Monitored in year *t-2*	-0.139	-0.193*	-0.139	-0.102	0.026	-0.001	-0.000	-0.214**	-0.115
	(0.088)	(0.113)	(0.095)	(0.082)	(0.037)	(0.001)	(0.001)	(0.107)	(0.090)
Observations	20,205	20,205	20,205	20,205	20,205	20,205	20,205	20,205	20,205
Other controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Randomization controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Mean for control group at *t*	65.86	55.58	67.22	77.35	15.33	0.040	0.089	58.59	68.33
Mean for control group at *t-1*	65.80	55.51	67.16	77.28	15.35	0.040	0.089	58.52	68.26
Mean for control group at *t-2*	65.88	55.61	67.23	77.35	15.33	0.041	0.090	58.58	68.34
% change for monitored at *t*	-0.004	-0.005	-0.003	-0.003	0.005	0.016	0.018	-0.005	-0.003
% change for monitored at *t-1*	0.002	0.003	0.001	0.001	-0.003	0.026	0.015	0.001	0.002
% change for monitored at *t-2*	-0.002	-0.003	-0.002	-0.001	0.001	-0.021	-0.002	-0.003	-0.001

Note: each regression includes randomization controls (region-by-year dummies and their interactions with current and lagged enrolment) and the other controls in vectors X, W and Z. Standard errors clustered by school within parentheses. Percent changes for monitored at *t, t-1* and *t-2* are obtained by dividing the treatment effect for the mean outcome for the control group for monitored at *t, t-1* and *t-2*, respectively. ***, **, * for statistical significance at the 1, 5 and 10 percent level of confidence.

Appendix

Table A1. The effects of external monitoring on test scores. Primary schools. Math – 5th graders. Without controls.

Outcome variable	(1) Mean	(2) Bottom quartile	(3) Median	(4) Top quartile	(5) Standard deviation	(6) Cheating index	(7) Percent absent in the test	(8) Mean open-ended questions	(9) Mean close-ended questions
Monitored in year *t*	-3.355***	-4.042***	-3.589***	-2.918***	0.630***	-0.020***	-0.005**	-4.268***	-2.545***
	(0.201)	(0.237)	(0.220)	(0.194)	(0.055)	(0.001)	(0.002)	(0.231)	(0.191)
Monitored in year *t-1*	-0.433**	-0.553**	-0.542**	-0.302	0.130**	-0.005***	0.001	-0.551**	-0.345*
	(0.209)	(0.250)	(0.228)	(0.194)	(0.060)	(0.002)	(0.003)	(0.238)	(0.200)
Monitored in year *t-2*	0.168	0.158	0.123	0.167	-0.000	0.002	0.001	0.078	0.206
	(0.201)	(0.241)	(0.219)	(0.186)	(0.059)	(0.002)	(0.002)	(0.236)	(0.190)
Observations	22,984	22,984	22,984	22,984	22,984	22,984	22,984	22,984	22,984
Other controls	No	No	No	No	No	No	No	No	No
Randomization controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Mean for control group at *t*	62.11	51.62	62.70	73.17	15.17	0.046	0.144	62.20	60.70
Mean for control group at *t-1*	61.85	51.31	62.43	72.93	15.22	0.045	0.143	61.87	60.51
Mean for control group at *t-2*	61.69	51.13	62.27	72.81	15.25	0.045	0.143	61.87	60.27
% change for monitored at *t*	-0.054	-0.078	-0.057	-0.039	0.041	-0.431	-0.034	-0.068	-0.041
% change for monitored at *t-1*	-0.007	-0.010	-0.008	-0.004	0.008	-0.115	0.006	-0.008	-0.005
% change for monitored at *t-2*	0.002	0.003	0.001	0.002	-0.001	0.033	0.004	0.001	0.003

Note: each regression includes randomization controls (region-by-year dummies and their interactions with current and lagged enrolment). Standard errors clustered by school within parentheses. Percent changes for monitored at *t*, *t-1* and *t-2* are obtained by dividing the treatment effect for the mean outcome for the control group for monitored at *t*, *t-1* and *t-2*, respectively. ***, **, * for statistical significance at the 1, 5 and 10 percent level of confidence.

Table A2. The effects of external monitoring on test scores. Primary schools. Literacy – 5th graders. Without controls.

Outcome variable	(1) Mean	(2) Bottom quartile	(3) Median	(4) Top quartile	(5) Standard deviation	(6) Cheating index	(7) Percent absent in the test	(8) Mean open-ended questions	(9) Mean close-ended questions
Monitored in year *t*	-2.653***	-3.334***	-2.726***	-2.116***	0.708***	-0.020***	-0.006**	-4.337***	-2.090***
	(0.165)	(0.206)	(0.182)	(0.149)	(0.054)	(0.001)	(0.002)	(0.216)	(0.157)
Monitored in year *t-1*	-0.303*	-0.289	-0.378**	-0.297*	0.024	-0.003**	0.001	-0.606***	-0.204
	(0.172)	(0.217)	(0.186)	(0.152)	(0.058)	(0.001)	(0.003)	(0.225)	(0.164)
Monitored in year *t-2*	0.132	0.189	0.095	0.125	0.012	0.001	-0.001	-0.046	0.178
	(0.165)	(0.209)	(0.180)	(0.144)	(0.059)	(0.001)	(0.003)	(0.220)	(0.158)
Observations	22,984	22,984	22,984	22,984	22,984	22,984	22,984	22,984	22,984
Other controls	No	No	No	No	No	No	No	No	No
Randomization controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Mean for control group at *t*	64.27	54.15	65.53	75.41	15.07	0.040	0.149	64.49	64.13
Mean for control group at *t-1*	64.07	53.89	65.33	75.26	15.13	0.038	0.149	64.17	63.97
Mean for control group at *t-2*	63.99	53.80	65.25	75.19	15.14	0.038	0.149	64.12	63.89
% change for monitored at *t*	-0.041	-0.061	-0.041	-0.028	0.047	-0.501	-0.039	-0.067	-0.032
% change for monitored at *t-1*	-0.004	-0.005	-0.005	-0.003	0.001	-0.088	0.004	-0.009	-0.003
% change for monitored at *t-2*	0.002	0.003	0.001	0.001	0.001	0.021	-0.009	-0.001	0.002

Note: each regression includes randomization controls (region-by-year dummies and their interactions with current and lagged enrolment). Standard errors clustered by school within parentheses. Percent changes for monitored at *t*, *t-1* and *t-2* are obtained by dividing the treatment effect for the mean outcome for the control group for monitored at *t*, *t-1* and *t-2*, respectively. ***, **, * for statistical significance at the 1, 5 and 10 percent level of confidence.

Table A3. The effects of external monitoring on test scores. Primary schools. Math and Literacy – 5th graders. Rasch scores.

Outcome variable	(1) Mean Math	(2) Bottom quartile Math	(3) Median Math	(4) Top quartile Math	(5) Standard deviation Math	(6) Mean Literacy	(7) Bottom quartile Literacy	(8) Median Literacy	(9) Top quartile Literacy	(10) Standard deviation Literacy
Monitored in year *t*	-0.211***	-0.221***	-0.211***	-0.203***	0.007**	-0.154***	-0.169***	-0.150***	-0.138***	0.017***
	(0.012)	(0.013)	(0.013)	(0.013)	(0.003)	(0.009)	(0.010)	(0.009)	(0.009)	(0.003)
Monitored in year *t-1*	-0.034***	-0.034**	-0.039***	-0.031**	0.001	-0.022**	-0.017	-0.024**	-0.025***	-0.002
	(0.013)	(0.013)	(0.013)	(0.013)	(0.003)	(0.009)	(0.010)	(0.010)	(0.009)	(0.003)
Monitored in year *t-2*	0.010	0.009	0.007	0.012	0.002	0.011	0.012	0.006	0.012	0.003
	(0.013)	(0.013)	(0.013)	(0.013)	(0.003)	(0.010)	(0.011)	(0.010)	(0.009)	(0.003)
Observations	22,984	22,984	22,984	22,984	22,984	22,984	22,984	22,984	22,984	22,984
Other controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Randomization controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Mean for control group at *t*	0.376	-0.228	0.359	0.962	0.898	0.271	-0.284	0.281	0.832	0.839
Mean for control group at *t-1*	0.360	-0.244	0.343	0.947	0.898	0.259	-0.298	0.270	0.823	0.841
Mean for control group at *t-2*	0.358	-0.247	0.341	0.945	0.898	0.258	-0.299	0.269	0.820	0.841
% change for monitored at *t*	-0.561	0.970	-0.588	-0.211	0.007	-0.567	0.593	-0.532	-0.166	0.020
% change for monitored at *t-1*	-0.094	0.139	-0.113	-0.032	0.001	-0.083	0.056	-0.089	-0.030	-0.002
% change for monitored at *t-2*	0.028	-0.034	0.020	0.012	0.002	0.041	-0.039	0.023	0.014	0.003

Note: each regression includes randomization controls (region-by-year dummies and their interactions with current and lagged enrolment) and the other controls in vectors X, W and Z. Standard errors clustered by schools within parentheses. ***, **, * for statistical significance at the 1, 5 and 10 percent level of confidence.

Table A4a. The effects of external monitoring on test scores. Primary schools with at most one class in the grade. Math – 5th graders.

Outcome variable	(1) Mean	(2) Bottom quartile	(3) Median	(4) Top quartile	(5) Standard deviation	(6) Cheating index	(7) Percent absent in the test	(8) Mean open-ended questions	(9) Mean close-ended questions
Monitored in year *t*	-8.622***	-10.602***	-8.933***	-7.851***	1.388***	-0.068***	0.006	-10.002***	-7.579***
	(1.940)	(2.232)	(2.076)	(1.829)	(0.521)	(0.017)	(0.011)	(2.254)	(1.821)
Monitored in year *t-1*	-3.145*	-3.990**	-4.025**	-2.125	0.999**	-0.053***	0.009	-3.254*	-3.050*
	(1.696)	(1.996)	(1.852)	(1.563)	(0.500)	(0.015)	(0.012)	(1.864)	(1.718)
Monitored in year *t-2*	1.588	2.349	1.155	0.747	-0.971*	0.034	-0.002	1.540	1.670
	(1.413)	(1.759)	(1.529)	(1.220)	(0.550)	(0.023)	(0.010)	(1.782)	(1.384)
Observations	3,114	3,114	3,114	3,114	3,114	3,114	3,114	3,114	3,114
Other controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Randomization controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Mean for control group at *t*	66.45	57.41	67.09	76.14	13.43	0.092	0.084	67.17	64.27
Mean for control group at *t-1*	66.31	57.25	66.95	76.01	13.45	0.092	0.084	67.05	64.11
Mean for control group at *t-2*	66.18	57.08	66.83	75.92	13.49	0.090	0.084	66.95	63.98
% change for monitored at *t*	-0.130	-0.185	-0.133	-0.103	0.103	-0.736	0.069	-0.149	-0.118
% change for monitored at *t-1*	-0.047	-0.069	-0.060	-0.028	0.074	-0.578	0.110	-0.048	-0.047
% change for monitored at *t-2*	0.024	0.041	0.017	0.009	-0.072	0.371	-0.019	0.023	0.026

Note: each regression includes randomization controls (region-by-year dummies and their interactions with current and lagged enrolment) and the other controls in vectors X, W and Z. Standard errors clustered by school within parentheses. Percent changes for monitored at *t, t-1* and *t-2* are obtained by dividing the treatment effect for the mean outcome for the control group for monitored at *t, t-1* and *t-2*, respectively. ***, **, * for statistical significance at the 1, 5 and 10 percent level of confidence.

Table A4b. The effects of external monitoring on test scores. Primary schools with two to three classes in the grade. Math – 5th graders.

Outcome variable	(1) Mean	(2) Bottom quartile	(3) Median	(4) Top quartile	(5) Standard deviation	(6) Cheating index	(7) Percent absent in the test	(8) Mean open-ended questions	(9) Mean close-ended questions
Monitored in year *t*	-4.904***	-6.026***	-5.068***	-4.420***	0.858***	-0.029***	0.001	-6.230***	-3.654***
	(0.699)	(0.821)	(0.763)	(0.686)	(0.202)	(0.005)	(0.007)	(0.800)	(0.669)
Monitored in year *t-1*	-0.987	-1.369	-1.144	-0.617	0.236	-0.006	0.001	-1.442*	-0.839
	(0.696)	(0.854)	(0.758)	(0.648)	(0.216)	(0.006)	(0.007)	(0.810)	(0.648)
Monitored in year *t-2*	0.419	0.709	0.484	0.425	-0.089	0.001	0.002	0.590	0.331
	(0.623)	(0.741)	(0.678)	(0.605)	(0.191)	(0.006)	(0.007)	(0.740)	(0.597)
Observations	3,816	3,816	3,816	3,816	3,816	3,816	3,816	3,816	3,816
Other controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Randomization controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Mean for control group at *t*	62.43	52.10	63.06	73.29	14.96	0.044	0.136	62.43	60.99
Mean for control group at *t-1*	62.19	51.82	62.83	73.06	15.00	0.043	0.136	62.14	60.81
Mean for control group at *t-2*	62.02	51.59	62.65	72.94	15.05	0.043	0.135	62.07	60.58
% change for monitored at *t*	-0.078	-0.116	-0.080	-0.060	0.057	-0.639	0.009	-0.099	-0.059
% change for monitored at *t-1*	-0.015	-0.026	-0.018	-0.008	0.015	-0.144	0.004	-0.023	-0.013
% change for monitored at *t-2*	0.006	0.013	0.007	0.005	-0.005	0.033	0.016	0.009	0.005

Note: each regression includes randomization controls (region-by-year dummies and their interactions with current and lagged enrolment) and the other controls in vectors X, W and Z. Standard errors clustered by school within parentheses. Percent changes for monitored at *t, t-1* and *t-2* are obtained by dividing the treatment effect for the mean outcome for the control group for monitored at *t, t-1* and *t-2*, respectively. ***, **, * for statistical significance at the 1, 5 and 10 percent level of confidence.

Table A4c. The effects of external monitoring on test scores. Primary schools with more than 3 classes in the grade. Math – 5th graders.

Outcome variable	(1) Mean	(2) Bottom quartile	(3) Median	(4) Top quartile	(5) Standard deviation	(6) Cheating index	(7) Percent absent in the test	(8) Mean open-ended questions	(9) Mean close-ended questions
Monitored in year *t*	-2.767***	-3.266***	-2.983***	-2.415***	0.493***	-0.015***	-0.006***	-3.469***	-2.128***
	(0.207)	(0.245)	(0.229)	(0.198)	(0.057)	(0.001)	(0.002)	(0.239)	(0.198)
Monitored in year *t-1*	-0.252	-0.301	-0.310	-0.179	0.105*	-0.003**	-0.001	-0.332	-0.174
	(0.224)	(0.267)	(0.244)	(0.208)	(0.063)	(0.002)	(0.003)	(0.255)	(0.214)
Monitored in year *t-2*	0.237	0.176	0.284	0.272	-0.013	0.000	-0.000	0.179	0.241
	(0.218)	(0.263)	(0.237)	(0.202)	(0.061)	(0.002)	(0.003)	(0.255)	(0.207)
Observations	11,821	11,821	11,821	11,821	11,821	11,821	11,821	11,821	11,821
Other controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Randomization controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Mean for control group at *t*	60.77	49.71	61.33	72.38	15.85	0.033	0.154	60.71	59.66
Mean for control group at *t-1*	60.49	49.38	61.03	72.14	15.90	0.031	0.153	60.33	59.46
Mean for control group at *t-2*	60.25	49.10	60.78	71.94	15.96	0.031	0.153	60.33	59.10
% change for monitored at *t*	-0.045	-0.065	-0.048	-0.033	0.031	-0.446	-0.041	-0.057	-0.035
% change for monitored at *t-1*	-0.004	-0.006	-0.005	-0.002	0.006	-0.110	-0.005	-0.005	-0.002
% change for monitored at *t-2*	0.003	0.003	0.004	0.003	-0.001	0.015	-0.001	0.002	0.004

Note: each regression includes randomization controls (region-by-year dummies and their interactions with current and lagged enrolment) and the other controls in vectors X, W and Z. Standard errors clustered by school within parentheses. Percent changes for monitored at *t, t-1* and *t-2* are obtained by dividing the treatment effect for the mean outcome for the control group for monitored at *t, t-1* and *t-2*, respectively. ***, **, * for statistical significance at the 1, 5 and 10 percent level of confidence.

Table A5a. The effects of external monitoring on test scores. Primary schools with at most one class in the grade. Literacy – 5th graders.

Outcome variable	(1) Mean	(2) Bottom quartile	(3) Median	(4) Top quartile	(5) Standard deviation	(6) Cheating index	(7) Percent absent in the test	(8) Mean open-ended questions	(9) Mean close-ended questions
Monitored in year *t*	-7.089***	-8.336***	-7.349***	-6.034***	1.544***	-0.075***	0.012	-10.831***	-5.777***
	(1.618)	(1.996)	(1.701)	(1.484)	(0.524)	(0.013)	(0.011)	(2.247)	(1.493)
Monitored in year *t-1*	-4.004***	-4.709***	-4.260***	-2.819**	1.543***	-0.039**	0.017	-4.522**	-3.844***
	(1.378)	(1.782)	(1.451)	(1.231)	(0.535)	(0.017)	(0.012)	(1.813)	(1.360)
Monitored in year *t-2*	1.625	1.998	1.688	1.252	-0.527	0.019	-0.005	2.188	1.458
	(1.188)	(1.536)	(1.242)	(0.975)	(0.471)	(0.017)	(0.010)	(1.569)	(1.230)
Observations	3,114	3,114	3,114	3,114	3,114	3,114	3,114	3,114	3,114
Other controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Randomization controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Mean for control group at *t*	68.04	59.19	69.33	77.98	13.55	0.076	0.083	69.78	67.39
Mean for control group at *t-1*	67.95	59.08	69.23	77.89	13.56	0.075	0.083	69.66	67.30
Mean for control group at *t-2*	67.85	58.96	69.13	77.82	13.59	0.075	0.083	69.54	67.22
% change for monitored at *t*	-0.104	-0.141	-0.106	-0.077	0.114	-0.980	0.146	-0.155	-0.085
% change for monitored at *t-1*	-0.058	-0.079	-0.061	-0.036	0.114	-0.513	0.199	-0.064	-0.057
% change for monitored at *t-2*	0.023	0.033	0.024	0.016	-0.038	0.254	-0.056	0.031	0.021

Note: each regression includes randomization controls (region-by-year dummies and their interactions with current and lagged enrolment) and the other controls in vectors X, W and Z. Standard errors clustered by school within parentheses. Percent changes for monitored at *t, t-1* and *t-2* are obtained by dividing the treatment effect for the mean outcome for the control group for monitored at *t, t-1* and *t-2*, respectively. ***, **, * for statistical significance at the 1, 5 and 10 percent level of confidence.

Table A5b. The effects of external monitoring on test scores. Primary schools with two to three classes in the grade. Literacy – 5th graders.

Outcome variable	(1) Mean	(2) Bottom quartile	(3) Median	(4) Top quartile	(5) Standard deviation	(6) Cheating index	(7) Percent absent in the test	(8) Mean open-ended questions	(9) Mean close-ended questions
Monitored in year *t*	-3.358***	-4.303***	-3.412***	-2.446***	1.044***	-0.031***	0.006	-5.387***	-2.706***
	(0.542)	(0.687)	(0.611)	(0.504)	(0.191)	(0.004)	(0.008)	(0.712)	(0.519)
Monitored in year *t-1*	-1.181**	-1.593**	-1.490**	-1.132**	0.227	-0.012***	0.001	-1.856**	-0.973*
	(0.567)	(0.724)	(0.615)	(0.511)	(0.201)	(0.004)	(0.008)	(0.759)	(0.547)
Monitored in year *t-2*	0.523	0.663	0.362	0.378	-0.220	0.007	0.005	0.387	0.546
	(0.499)	(0.639)	(0.540)	(0.446)	(0.201)	(0.005)	(0.007)	(0.689)	(0.477)
Observations	3,816	3,816	3,816	3,816	3,816	3,816	3,816	3,816	3,816
Other controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Randomization controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Mean for control group at *t*	64.47	54.57	65.68	75.42	14.83	0.039	0.140	64.72	64.31
Mean for control group at *t-1*	64.33	54.39	65.55	75.33	14.88	0.038	0.140	64.48	64.20
Mean for control group at *t-2*	64.21	54.24	65.43	75.23	14.92	0.037	0.139	64.37	64.08
% change for monitored at *t*	-0.052	-0.078	-0.052	-0.032	0.070	-0.769	0.045	-0.083	-0.042
% change for monitored at *t-1*	-0.018	-0.029	-0.022	-0.015	0.015	-0.319	0.005	-0.028	-0.015
% change for monitored at *t-2*	0.008	0.012	0.005	0.005	-0.014	0.175	0.038	0.006	0.008

Note: each regression includes randomization controls (region-by-year dummies and their interactions with current and lagged enrolment) and the other controls in vectors X, W and Z. Standard errors clustered by school within parentheses. Percent changes for monitored at *t, t-1* and *t-2* are obtained by dividing the treatment effect for the mean outcome for the control group for monitored at *t, t-1* and *t-2*, respectively. ***, **, * for statistical significance at the 1, 5 and 10 percent level of confidence.

Table A5c. The effects of external monitoring on test scores. Primary schools with more than 3 classes in the grade. Literacy – 5th graders.

Outcome variable	(1) Mean	(2) Bottom quartile	(3) Median	(4) Top quartile	(5) Standard deviation	(6) Cheating index	(7) Percent absent in the test	(8) Mean open-ended questions	(9) Mean close-ended questions
Monitored in year *t*	-2.228***	-2.797***	-2.280***	-1.822***	0.559***	-0.014***	-0.008***	-3.754***	-1.714***
	(0.164)	(0.209)	(0.181)	(0.147)	(0.057)	(0.001)	(0.003)	(0.224)	(0.155)
Monitored in year *t-1*	-0.074	0.014	-0.072	-0.082	-0.001	-0.002	-0.001	-0.305	0.009
	(0.180)	(0.229)	(0.193)	(0.160)	(0.062)	(0.001)	(0.003)	(0.244)	(0.168)
Monitored in year *t-2*	0.141	0.154	0.113	0.162	0.027	-0.000	-0.005*	-0.062	0.194
	(0.173)	(0.220)	(0.188)	(0.150)	(0.060)	(0.001)	(0.003)	(0.234)	(0.163)
Observations	11,821	11,821	11,821	11,821	11,821	11,821	11,821	11,821	11,821
Other controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Randomization controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Mean for control group at *t*	63.27	52.72	64.53	74.86	15.62	0.028	0.162	62.89	63.32
Mean for control group at *t-1*	63.02	52.39	64.29	74.67	15.69	0.027	0.161	62.48	63.14
Mean for control group at *t-2*	62.92	52.29	64.19	74.58	15.71	0.027	0.162	62.44	63.03
% change for monitored at *t*	-0.035	-0.053	-0.035	-0.024	0.035	-0.493	-0.048	-0.059	-0.027
% change for monitored at *t-1*	-0.001	0.001	-0.001	-0.001	-0.001	-0.059	-0.009	-0.004	0.001
% change for monitored at *t-2*	0.002	0.002	0.001	0.002	0.001	-0.011	-0.029	-0.001	0.003

Note: each regression includes randomization controls (region-by-year dummies and their interactions with current and lagged enrolment) and the other controls in vectors X, W and Z. Standard errors clustered by school within parentheses. Percent changes for monitored at *t, t-1* and *t-2* are obtained by dividing the treatment effect for the mean outcome for the control group for monitored at *t, t-1* and *t-2*, respectively. ***, **, * for statistical significance at the 1, 5 and 10 percent level of confidence.