ECONOMIC POLICY

70th Economic Policy Panel Meeting

10-11 October 2019
Helsinki

Hosted by Bank of Finland

Monitoring and Sanctioning Cheating at School: What Works? Evidence from a National Evaluation Program

Claudio Lucifora (Catholic University of Milan)
Marco Tonello (Bank of Italy and Catholic University of Milan)

The organisers would like to thank the Bank of Finland for their support.

Monitoring and sanctioning cheating at school: What works? Evidence from a national evaluation program

Claudio Lucifora* Marco Tonello♦

This version: September, 2019

Abstract

We investigate the effectiveness of different policies to reduce cheating and other forms of opportunistic behavior in school standardized testing. We exploit a randomized experiment in Italian schools to assess the causal effect of both an external monitoring and a sanctions program on cheating behavior and absence rates. We find, in line with previous studies, that external monitoring is effective in deterring cheating occurring during and after the test. We show evidence of a strategic response to monitoring in terms of higher absence rates, which alter the pool of students who sit the test. Sanctions are in general not effective in reducing cheating, while they have a discipline effect in decreasing absence rates. Both monitoring and sanctions programs work better in cultural and institutional settings that make the potential loss of reputation costlier to the school.

JEL Classification: I21, I28, K42

Keywords: cheating, sanctions, monitoring, incentives

---

* Università Cattolica del Sacro Cuore, CRELI and IZA (claudio.lucifora@unicatt.it). Address: Department of Economics and Finance, L.go Gemelli 1, 20123 Milano (Italy). **Corresponding author**.

♦Bank of Italy, Territorial Analysis and Economic Research Division, Firenze Branch, and CRELI, Università Cattolica del Sacro Cuore (marco.tonello@bancaditalia.it). Address: Via dell'Oriuolo 37/39, 50133 Firenze (Italy).

We are grateful to Patrizia Falzetti and Valeria Tortora (Invalsi) for making the data available and for their guidance and insights in using them; Santiago Pereda-Fernàndez and Sergio Longobardi kindly shared with us their statistical indicators of cheating. We thank for helpful comments and suggestions the Editor Andrea Ichino, the anonymous referees, and the EP Panel Members. We are also grateful to Mino Barone, Massimiliano Bratti, Lorenzo Cappellari, Mikael Lindhal, Federica Origo, Vincenzo Scoppa, seminar participants at CRELI (Università Cattolica), Invalsi, XXXI Italian Labour Economists Annual Meeting (University of Trento), Counterfactual Methods for Policy Evaluation Conference (COMPIE 2016) for their comments. The views expressed in this paper are those of the authors and do not necessarily reflect those of the Bank of Italy or Invalsi. The usual disclaimers apply.

1. Introduction

Standardized tests are increasingly used around the world to assess students' performance, to reward teachers' quality, or to allocate resources across schools. One of the most studied evaluation program is the No-Child-Left-Behind Act (NCLB) in the United States, which establishes standards for student achievement with rewards (sanctions) for high-performing (low-performing) schools. Similar programs with test-based accountability systems, with or without explicit rewards and sanctions based on performance indicators, have been introduced or are currently experimented in several other countries around the world (e.g. Germany, Italy, UK, Sweden, Mexico). Such evaluation programs have generated considerable controversy on whether they should be used simply to assess students' achievements, without further implications for teachers or schools (i.e. low-stakes system), or alternatively they should be part of a comprehensive accountability system with provisions for students' career, teachers' rewards and funding for schools (i.e. high-stakes system) (Neal 2013). Indeed, one major consequence associated with the diffusion of national evaluation programs has been the pervasiveness of opportunistic behaviors and cheating practices in schools (U.S. Department of Education 2009, Eurydice 2009, UK Standard & Testing Agency 2013). A survey conducted in several countries by 'The Wall Street Journal', reported that on average 28% of the respondents admitted to have ever cheated at school. This figure ranges from 15% in the UK, to 37% in Germany and Russia, up to 41% in France.[1]

While cheating practices are not confined to the education system[2], the occurrence of cheating in school can be particularly disruptive due to the long-run effects that the misallocation of resources generates (Mechtenberg 2009). Cheating in test score evaluation programs contaminates the information provided by the educational system about student achievement after instruction, it interferes with evaluators ability to assess students' performance and reduces the external validity of test results (Anderman and Murdock 2007). The long-term consequences of cheating in school can be even more severe in educational systems that rely on a strict tracking system (Brunello and Checchi 2007). Evaluation programs, particularly in high-stakes settings, also generate incentives for teachers and school administrators to manipulate the scores to improve their standing, increase schools' attractiveness and ratings (Ahn and Vigdor 2014). In low-stakes setting, due to the lack of significant consequences attached to testing outcomes for both students and schools, cheating is generally expected to be lower. However, extensive

[1] Figures obtained from the 'Survey on Deceit', The Wall Street Journal (2008). See Appendix A for further details.

[2] Examples in other fields are: untruthful tax declarations, free-riding on public goods, shirking on colleagues at work, cheating in sports or in games (Kleven et al. 2011; Card and Giuliano 2013).

cheating has been reported in many low-stakes settings too, suggesting that implicit incentives and reputational concerns may play a relevant role (Wall Street Journal 2008).

Starting from the seminal work of Jacob and Levitt (2003), a related and very recent literature has contributed to better understand the moral hazard problems arising in evaluation programs and empirically identify and quantify cheating practices in test score assessments (Behrman et al. 2015, Martinelli et al. 2018, Dee et al. 2019, Diamond and Persson 2016). A flourishing economic and psychological literature has documented the diffusion of cheating practices over the past years in all grades of the schooling system (Carrell et al. 2008, Anderman and Murdock, 2007, Davies et al. 2009, Dee and Jacob 2012, Dee et al. 2019). In this respect, the high incidence of cheating in Italian schools proves an interesting case study, as documented by several recent works (Bertoni et al. 2013, Paccagnella and Sestito 2014, Lucifora and Tonello 2015, Pereda Fernàndez 2016, Angrist et al. 2017, Battistin et al. 2017). Most of these studies exploit a natural experiment (i.e. the presence of an external inspector in the class, Bertoni et al. 2013), or variation in class size induced by administrative cutoff rules (Angrist et al. 2017), to assess the effect of stricter monitoring on test scores. The main findings document widespread manipulation in test scores, with evidence of higher manipulation in Southern regions originating from teachers' moral hazard and low accountability pressure.

In this paper, we evaluate the effectiveness of two main programs aimed at reducing cheating behavior in Italian schools. The first program, relies on the presence of an external inspector to monitor students' and teachers' behavior during the administration and proctoring of the testing process. The second program consists in a system of sanctions based on a 'fame and shame' policy - expected to affect the school's reputation - for schools identified as having a high likelihood of cheating and manipulation of the test scores. While the effects of the monitoring program on test score manipulation have already been investigated in the literature, to the best of our knowledge, we are the first to study the effects of the sanctions program on different indicators of cheating. In particular, we expand the set of outcomes conventionally used to assess the testing process, and investigate the timing of cheating, such as actions taking place *during* and *after* the testing process, as well as the strategic responses altering the pool of students *before* the test.

The paper contributes to this emerging literature along different dimensions. First, it provides a formal evaluation of the effectiveness of various measures taken to reduce cheating in standardized testing. Only a limited number of studies have investigated the effects of policy

interventions aimed at curbing cheating and opportunistic behavior in evaluation programs.[3] In this respect, the paper offers a novel contribution by comparing and contrasting the effectiveness of alternative programs based either on direct monitoring by external inspectors, or on sanctions imposed on schools. In practice, we investigate the effects of non-monetary sanctions to schools identified as suspect of cheating, which have been never studied before, either the Italian setting or in other countries.

Second, we extend previous work on the effects of the external inspector program (as in Bertoni et al. 2013, Angrist et al. 2017), to consider different statistical indicators of cheating (i.e. statistical anomalies in the test scores patterns built in the Cheating Propensity Indicator, CPI), as well as absence rates of students who sit the test (i.e. the number of students formally enrolled in the school compared to those who take the test). While the CPI captures the main alterations in the patterns of the tests responses (Castellano et al. 2009), indicating the presence of cheating behavior (by teachers, students or both), absence rates might be taken as an indicator of opportunistic behavior (again by teachers, students or both) with the objective of altering the composition of the pool of students that sit the test (the so-called strategic pooling).[4] This type of opportunistic behavior has been generally neglected in the literature: existing works on the Italian setting did not consider absence rates, while only Figlio (2006) provides evidence of specific forms of strategic pooling in a high-stake testing environment.

Third, our estimation framework is based on a school-level analysis, which accounts both for the existence of spill-over effects between classes – as outlined by Bertoni et al. (2013) –, as well as the possibility of endogenous allocation of inspectors to classes within schools – as discussed in Angrist et al. (2017). We exploit the random allocation of the external inspector to the schools to estimate the causal effect of external monitoring (Invalsi 2010, 2011), as well as the effect of the sanctions program, on both cheating and absence rates (Falzetti, 2013). Moreover, while all previous works have restricted the analysis to primary schools (grades 2 and 5), we focus on higher grades (junior-high and high school, grades 6 and 10), which have not been investigated so far and where strategic pooling behavior may be more relevant (Anderman and Murdock 2007).

---

[3] To the best of our knowledge, Dee and Jacob (2012) is the only paper that specifically looks at the effectiveness of a policy to reduce cheating, though it is focused on a very specific form of cheating (i.e. plagiarism in take-home assignments), while Dee at al. (2019) show that manipulation in the New York Regents Examinations disappears when passing from a local to a centralized grading system.

[4] Notice that schools must pay attention to both outcomes: CPI and absence rates. Test score results are usually returned to the schools only if the CPI (i.e. the estimated level of cheating) is under a given threshold of tolerance. Absence rates are also monitored by Invalsi, as test results are not considered representative of class average performance (and are not returned to schools), if a higher than normal share of students is absent during the test (Invalsi 2010, 2011).

Fourth, we sketch a taxonomy of cheating behavior at school which is useful to better rationalize the different forms of cheating, based on *who* takes the action (students or teachers), and on the *timing* of cheating (before, during the test or after). Our work refers to different parts of this proposed taxonomy, even if we cannot formally distinguish students' from teachers' (or principals') behavior. Namely, the analysis on the absence rates fills the gap in the literature on cheating behavior *before* the test, while the analysis on the statistical cheating indicator contributes to a better understanding of opportunist behavior *during* and *after* the test.

We find that the presence of the external inspector reduces cheating propensity (CPI) by about 20 percent, a figure that is lower with respect to previous findings in the literature on primary school grades, but it also increases absence rates by about 8 percent, altering the composition of the students who sit the test. With respect to the "fame and shame" sanctions program, we find that schools that have been sanctioned do not significantly change their cheating behavior. We argue that, in low-stake testing environments where school choice and accountability are limited, non-pecuniary sanctions are generally ineffective when it comes to discipline cheating behavior and manipulations during (or after) the testing process. However, we do find that sanctions have an effect in reducing students' absence rates, suggesting that school which received a sanction are more likely to react when it comes to strategic pooling or students' absenteeism.

While we cannot disentangle the exact mechanisms through which sanctions may or may not work, we interpret these findings as evidence that schools are not effective in taking corrective actions, besides direct monitoring, to deter the complex cheating interactions (or manipulations) which may occur during (or after) the test, but are likely to be revealed only *ex-post* (i.e. Invalsi reports the results to the schools only several months later). Conversely, changes in absence rates are readily observable and schools, which have been sanctioned in the past, may be more inclined to take *ex-ante* actions to reduce strategic pooling and students' absenteeism. In other words, we argue that sanctions are more likely to work when the school's opportunistic behavior can be better and readily observed (as with absence rates), while they do not work when the latent cheating behavior is only observed with a lag and it is measured with error (i.e. the CPI).

Also, we find heterogeneity in the effectiveness of sanctions across different contexts. In particular, in areas where trust and institutional quality are less valued and the reputational cost of sanctions is smaller, sanctions programs are unlikely to affect school behavior. Put it differently, "fame and shame" type of sanctions seem to work better where the institutional and social context can make the potential loss of reputation costlier to the school.

The rest of the paper is organized as follows. Section 2 illustrates the institutional setting. In section 3 the main features of our conceptual framework are discussed. Section 4 describes the data used and presents some descriptive statistics. Section 5 illustrates the empirical strategy,

while in section 6 we report the main results and the heterogeneous effects. In Section 7 we perform the robustness checks. Section 8 concludes and discusses the policy implications.

2. Institutional context

2.1 The Italian school system: organization and school choice

The school system in Italy is organized in five years of primary school (grades 1 to 5, corresponding to ISCED level 1) and three years of junior-high school (grades 6 to 8, ISCED level 2). At the end of the junior-high school (i.e. after completing 8 years of education) students obtain a Diploma, which entitles them to enroll in high school. The high school cycle can last two or five years (grades 9 to 13, ISCED level 3), according to the type of track chosen. Academic and technical high school tracks last for five years, and prepare student for college (academic track) or provide skills for the labor market (technical track). The vocational school track lasts for two years only and mainly endow students with vocational skills necessary to start a job. In general, children enroll in the first grade of primary school the year they turn six, start junior-high school when they turn eleven, and enroll in the first grade of high school the year they turn fourteen. The primary and junior-high schools and the first two years of high school are compulsory for all students.

In this work we focus our analysis on students in grade 6 of junior-high school and grade 10 of high school. In junior-high and high school, students have several teachers, one for each subject, and are expected to gain knowledge on a wide range of skills. Due to the larger number of teachers, the amount of time each teacher passes in the classroom in junior-high and high school is considerably lower than in primary school, as it is the formation of interpersonal relationships between students and teachers.

School choice is limited, as mainly based on residence criteria, but schools can exploit some discretion in attracting students, as resources are mainly allocated on the basis of pupils' enrollments. In primary and junior-high schools parents can choose among the set of available schools depending on their municipality of residence and proximity to school facilities.[5] Since school resources mostly depend on the number of students enrolled, bigger schools receive more money and more teachers, while small schools tend to be embedded by the closest biggest ones. Concerning high-schools, they are in principle free to compete to attract more students (and resources). Nevertheless, school formation is regulated by the Ministry of Education, that might limit the presence of high-schools of the same type in a given catchment area.

[5] For example, more choice is likely to be available in highly populated municipalities (e.g. in big cities).

2.2 The SNV Evaluation Program

Invalsi started the National Evaluation Program of Students' Achievement (henceforth, SNV Evaluation Program) in the school year 2009-10. The SNV has a yearly and census nature: every school year, between late April and May, all students in grades 2 and 5 (primary school), grade 6 (junior-high school), and grade 10 (high school) sit a language and a math test.[6] The SNV Evaluation Program is not conceived as a high-stakes test: the results of the evaluation have the purpose of assessing the school's performance from one year to another and are not part of a proper school accountability system (i.e. either granting additional funding or rewarding teachers). Since their first release in 2010, the results of the SNV assessments have been widely covered by the media and closely scrutinized by parents when selecting the school for their children. Although formal obligation for schools to disclose their SNV results became effective only starting from the school year 2014-15, schools started much before to advertise their SNV results to improve visibility, raise enrollments and attract more students.[7] In particular, the results of test scores have increased their importance over time, both for parents and other school's stake holders, creating a system of informal incentives for teachers and school principals. In this context, Invalsi has enforced a strict protocol in the administration and marking of the test (Invalsi, 2010). Students are proctored by teachers chosen from a different class and specialized in a different subject with respect to the one tested. The answer sheet of each student is marked contemporaneously by several school teachers in order to cross-check each other. The answers of the entire class are reported into one answer sheet and then sent to Invalsi for the computation of the scores. Test scores are then returned to each school and class in September. Several bodies within and outside the school (i.e. teachers, school principal, students' and parents' representatives) have direct access to the results, and may decide to make them public, to promote the school reputation and visibility.

2.3 Invalsi programs to fight cheating in schools

[6] We do not have data for 8th grade students who are tested under a specific assessment program which is not formally part of the SNV and cannot be fully compared to the other grades (e.g. it follows a different timing and it contributes to the formal examination taken by the students at the end of the 8th grade to gain the junior-high school Diploma).

[7] In the Appendix, Figure B.1, we show the online webpage of a school where the most recent SNV results are advertised. Since the school year 2011-2012, the results of each school's SNV evaluation have been made available to the school head, to the teachers and to the representative bodies of parents and students. Invalsi also established the duty for schools to disclose the SNV results whenever parents asked for it. Starting from the school year 2014-15 all schools have to disclose several information about their facilities, programs and teachers, as well as their average performance in previous SNV (*http://cercalatuascuola.istruzione.it/cercalatuascuola/*).

In order to minimize illegal and opportunistic behaviors in the SNV Evaluation Program, Invalsi has further introduced two main deterrence measures: an "external monitoring program", for the administration and proctoring of the tests, and a "sanctions program" for schools whose performance in the test scores displays a high likelihood of cheating.

The external monitoring program. Every schooling year, for the SNV evaluation, Invalsi sends external inspectors to a random and representative sample of classes. The external inspectors have the duty to administer the tests and are responsible for the marking process. Each external inspector receives a compensation of about 200€ for carrying out his duties. We define as 'monitored class', a class in which the test is proctored and marked by an external inspector, and as 'monitored school', a school in which there is at least one monitored class. The external inspector represents a random event which changes the monitoring technology. Stricter monitoring implies a 'non-cheating' environment, where the possibility of cheating both by students and teachers, during and after the test, is remarkably reduced.[8] Once a monitored school has been selected, the selection of the class within the school is totally random (Invalsi, 2011).

The external inspectors are sampled each year mainly from a pool of retired teachers. The school principal is informed of the presence of an external inspector in the school only few days before the SNV tests take place. The short notice is intended to reduce *ex-ante* opportunistic behavior, but we cannot exclude that school principals put in place some strategic reactions. For example, the school principal or the school teachers may try to select the pool of students that sit the test by inducing lower ability student to be absent on the day of the test, thus altering the composition of the class. Alternatively, the school principal might try to manipulate the assignment of the external inspector to the class he is randomly assigned to proctor in favor of a class with better quality students (we return the implications of this behavior when we discuss the identification strategy).

The sanctions program. After the SNV 2011-12 was completed, but before the scores were returned to the schools, Invalsi implemented a new sanctions program to reduce cheating. The policy was unanticipated by the schools and implemented in September 2012 (Falzetti, 2013). It consisted of two different measures: (i) *correction* (deflation) of the class test scores, or (ii) *non-return* of the test scores to the class, depending on the cheating score detected during the administration of the SNV 2011-12, in May 2012.

In details, an algorithm was implemented when returning test scores of the 2011-12 SNV to the schools in September 2012. The algorithm combines a statistical indicator of cheating

[8] Additional details on this policy and on the randomization scheme can be found in Invalsi (2011), Bertoni et al. (2013), Lucifora and Tonello (2015), Angrist et al. (2017). These works show that the presence of the external inspector reduces test scores, social interactions and teachers' shirking.

Cheating Propensity Indicator, CPI_{cgj}, within class (c), grade (g) and subject (j), with a 'threshold of cheating statistical acceptability' ($TCSA_{gj}$) defined at the national level but specific to each grade (g) and subject (j).[9]

The implementation and the strictness of the sanctions was then set comparing the CPI_{cgj} with the reference level of the $TCSA_{gj}$. In other words, the test scores of each class were returned directly to the school either without any correction, or after 'correcting' them by means of a cheating deflator. Alternatively, in cases of high levels of cheating the results were not returned at all. Notice that in all the above cases the school is made aware of the fact that the correction/non-return procedure was applied because cheating was detected during the test. The above measures thus introduced *de facto* three different regimes:

a) $CPI_{cgj} \leq TCSA_{gj}$ test scores are returned without correction;

b) $TCSA_{gj} < CPI_{cgj} \leq 0.5$ test scores are returned to each class after being 'corrected' by an implicit cheating deflator;[10]

c) $CPI_{cgj} > 0.5$ test scores are not returned to the class at all (and are excluded from the calculation of the school average score).[11]

Hence, starting from September 2012 (i.e. at the beginning of the school year 2012-13) and as a result of the implementation of the new sanctions program, any school could be receiving a sanction for any of its classes according to the above three scenarios. Clearly, any combination of the above cases is possible within any school, from no classes receiving any correction measures, to one or more classes receiving a correction for high likelihood of cheating detected or non-return at all of the test scores.

3. Cheating in schools

To highlight the margins upon which the Invalsi monitoring and the sanctions programs are expected to work, we discuss the main features of cheating behavior along two main dimensions: the agents' behavior and the timing with respect of the administration of the test. Next, we

[9] Invalsi calculates the median of the CPI in the 5 macro-areas of the country (North-West, North-East, Centre, South, Islands) and sets this threshold at the median CPI of the lowest cheating macro-area.

[10] The correction for cheating in the Invalsi SNV Protocol is such that the test scores are multiplied by ($1 - CPI_{cgj}$), that is, the test scores are 'deflated' by a factor that is proportional to the likelihood of cheating.

[11] The threshold of 0.5 was chosen arbitrarily (Falzetti 2013), and it broadly corresponds to the 95th percentile of the CPI overall distribution. If in a given school s, more than 50 percent of the classes satisfy condition (c), then the test scores results were not returned to the entire school. We do not consider this additional treatment in our analysis because it was so rare in the SNV wave used (between 2 and 3 percent of the schools, depending on the subject tested) that a formal evaluation exercise is not feasible. The results do not change if we exclude these schools from the analysis.

compare and contrast how the implementation mechanisms of the monitoring and the sanctions programs work.

3.1. A taxonomy of cheating behavior in schools

Cheating in school may arise from different opportunistic behaviors. Teachers and school principals, for example, in order to achieve higher ratings or attract better prospective students may focus attention on tested subject only or manipulate test scores (Finn 2015). Students themselves may cheat to achieve a better performance and gain admission to selective schools (Martinelli et al. 2018, Diamond and Persson 2016). Overall, since the effort required to achieve a good performance is costly for all the agents involved, test scores systems are likely to produce incentives for opportunistic behaviors. Also, since illicit behaviors are unobservable and can take different forms, contingent contracts cannot be generally signed. The institutional setting, the procedures adopted for the administration of the tests and a poor monitoring process can often exacerbate moral hazard problems and the incidence of cheating. Also, even if cases of cheating are frequently reported, sanctions for illicit behaviors are not commonly used in schools.

Stricter monitoring, which increases the probability of cheaters being caught, has been found to reduce opportunistic behaviors and the incidence of cheating, however it is generally not feasible, or cost-effective, in national evaluation programs that are run on a census basis (Bertoni et al. 2013). Alternative systems of incentives and sanctions are often designed as threat to punish violators and avoid undesirable outcomes. Incentives of this type are generally introduced to hold negligent agents liable for breaking the rule of law (De Geest and Dari-Mattiacci 2014).[12] One problem with the sanctioning measures is that, in most cases, latent cheating can only be inferred through statistical indicators that are imperfectly measured and cannot typically distinguish among different types of cheating (unless specific experimental designs are in place), which makes particularly difficult the design and the implementation of the penalties.

[Table 1]

To fix ideas, in Table 1 we highlight the main mechanisms that may be expected to drive cheating practices in testing. We distinguish between agents' contribution to cheating behavior (i.e. teachers and students) and by the timing of cheating over the administration of the test (i.e. before the test is administered, during the test itself, or after in the marking phase).

Students' driven cheating behavior materializes prevalently during the test and can take two main forms: collaborative effort, during the test in exchanging information or copying from the

[12] In this set-up, the decision to cheat weights the expected payoff from cheating with the disutility from being caught and sanctioned. While the benefit of cheating are in terms of lower effort and (expected) higher performance, the costs crucially depend on the (ex-ante) probability of being caught and on the (ex-post) severity of the sanction.

peers; or use of prohibited materials and technologies. In both cases, students' cheating can be interpreted as a form of complex social interaction, in terms of collaborative behavior – between students exchanging information –, or peer pressure – originating from other students' cheating behavior (Carrel et al. 2008; Lucifora and Tonello 2015; Martinelli et al. 2018).

Conversely teachers' contribution to cheating behavior can take several forms. First, before the test, cheating may occur as strategic pooling when teachers and school principals attempt to raise the school's overall performance profile by reshaping the pool of students who sit the test (e.g. retaining low-scoring students in grade, or classifying more students in 'special needs' to exclude their scores from school averages) (Figlio 2006). Second, teachers can concentrate on 'teaching to the test' strategies by focusing on tested subject only, or devoting extra effort on students at the margin of passing the test or failing it, while lavishing their attention to students that are not likely to pass the test or that are almost sure of passing it (Lazear 2006; Neal and Schanzenbach 2010). Third, during the proctoring of the test, teachers can adopt a benevolent attitude by lowering monitoring standards to let students use prohibited materials and collaborate. Alternatively, teachers can suggest answers or hints directly to students during the exams. The educational psychology literature suggests that altruistic behaviors tend to increase with the length of time the teacher has been with the students (Anderman and Murdock, 2007). Finally, after the test is administered, and during the marking process (when carried out by teachers of the same school), teachers can directly manipulate the students' answer sheets (Jacob and Levitt 2003; Dee et al. 2019; Diamond and Persson 2016; Angrist et al. 2017). While cheating behavior is generally unobservable, the different configurations of cheating vary in their observability and measurement affecting the effectiveness of the sanctioning process.

Even if we cannot formally distinguish students' from teachers' (or principals') behavior, our work is related to all parts of this taxonomy. The analysis on absence rates fills the gap in the literature on cheating behavior *before* the test, while the analysis on the statistical cheating indicator (CPI) contributes to the evaluation of how direct monitoring and sanctions might reduce opportunist behavior *during* and *after* the test, both on the part of students and teachers.

3.2. A comparison of the Invalsi programs

To get a rough idea of how the different programs used by Invalsi are expected to address cheating behavior, it is useful to contrast their different implementation mechanisms. The external monitoring program mainly works through a change in the monitoring technology that strictly follows the Invalsi protocol during the administration and marking of the tests. The program, by increasing the probability of detecting illicit or opportunistic behavior, is likely to remove opportunities for cheating behavior both for students and teachers, during and after the

test (see Table 1). However, since inspectors are present in the school during or immediately after the testing process, any cheating that occurs before the testing takes place is likely to go undetected. In particular, since the assignment of an inspector to a class within a school is announced few days in advance, school can alter strategically, *ex-ante*, the pool of student who sit the test without being affected by the monitoring program.

The sanctions program instead is expected to work by changing the incentives to cheat through a 'fame and shame' mechanism, leveraging on the reputational concerns for schools being stigmatized as cheaters, without requiring any additional direct cost or extra resources (Falzetti, 2013). Schools that exhibit anomalies in the patterns of answers within classes, and thus are identified with a high likelihood of cheating, are sanctioned either with the correction of the test score results or by withholding the class test score performance from the evaluation program. Notice that, while the external monitoring program works directly during the administration of the test in a given school year, the correction/non-return sanctions refer to a latent cheating behavior and are *de-facto* deferred to the following school year (i.e. effectively takes place between one school year and the next). These features of the sanctions program, jointly with the lack of a proper accountability system with explicit pecuniary sanctions or reallocation of resources, is generally considered a low-power scheme with a weak deterrence effect. Moreover, the complex interactions underlying cheating behavior and the difficulties in measuring it with precision make the reputational cost more uncertain adding further difficulties to the effectiveness of the sanctions program. In other context, however, the 'fame and shame' sanction program may work, even without a proper accountability system, provided that school's opportunistic behavior is observable and easier to measure. In such context, when schools compete to attract prospective students and increase their resources, the potential loss of reputation can generate sizeable losses by reducing the attractiveness of the school. Clearly the size and significance of the expected losses and the real bite of the deterrence measures are likely to depend on the transparency and reliability of the information and the level of school competition.

Finally, notice that absence rates, constructed similarly to ours, are routinely monitored by Invalsi since the first waves of the SNV Evaluation Program. Indeed, the test results are not considered representative of a class average performance, nor returned to schools, if a high proportion of students absent in the day of the test is detected (Invalsi 2010, 2011). This generates a trade-off in the strategic pooling behavior by the school (or students): on the one hand, selecting the pool of students who sit the test may improve average performance, but, on the other hand, high absence rates could backfire with sanctions enforced by the monitoring institution.

3.3. Measuring cheating

Cheating in school is typically regulated by ethical codes, which define and discipline illicit behaviors. Still most opportunistic and dishonest behavior are very often overlooked and tolerated within many schools. Even if mostly unobservable, cheating behavior usually determines unexpected and unusual patterns in test scores answers within a classroom that can be analyzed and measured. Cheating may result in block of identical answers (either correct or wrong), strange patterns of correlations across students' answers, as well as anomalous association of average and dispersion of the scores.

Building on the education measurement literature (Wollack et al. 2001) and the seminal work of Jacob and Levitt (2003), a growing literature has tried to develop algorithms to detect cheating behavior.[13] Since 2009, Invalsi has developed a Cheating Propensity Indicator (CPI) with the purpose of measuring and monitoring the propensity to cheat in Italian schools. The CPI, which is class and subject specific, can be interpreted as the probability that cheating occurred in each classroom during the test. Typically, a classroom displays a high likelihood that some cheating occurred the more homogeneous is the pattern of (right or wrong) responses and non-responses to each single item, as well as the higher is the average and the lower the variability of the scores (Castellano et al. 2009, Invalsi 2010). For the purpose of the present study, we obtained from Invalsi a calculation of the CPI for the universe of Italian schools in the various grades, which is what we use in the empirical analysis. To gain insights on other cheating practices not directly captured by test scores, as discussed in the previous section, we also construct an indicator of strategic pooling, based on the share of students absent on the day of the test (*Absence rate*).

Some caveats are in order. First, the CPI approach is likely to underestimate the incidence of cheating, since only major and systematic manipulations are likely to be captured, while subtler or moderate cheating behavior might go undetected. Second, the CPI indicator does not allow to disentangle the contribution of students and teachers to the observed cheating behavior: still, for student to be able to cheat during the test there must always be a negligent (or a benevolent neglect) attitude of the teacher. Finally, while the *Absence rate* is intended to capture anomalies in the number of students' absent on the day of the test, the lack of information on their

[13] The detection of cheating in test scores is generally based on statistical or sequential indicators. For instance, Jacob and Levitt (2003) exploit the panel dimension of their data and additional information on teachers and class codes, to identify plausible patterns of cheating based on sequential indicators and unexpected jumps in test scores performance. On the contrary, Dee and Jacob (2012) use statistical software to detect plagiarism in take-home assignments. Martinelli et al. (2018) make use of several statistical indicators of cheating in exams developed in the education measurement literature and based on exact matches between each possible couple of students sitting the exam in the same classroom. Estrada (2019) exploits indicators based on Error Similarity Analysis calculated by the Mexican Federal Secretary of Education.

performance (i.e. since they are absent) does not allow to compute a more precise measure based on the share of low-scoring students who are absent. We return further details on these indicators and some descriptive statistics in the following section.

4. Data and descriptive statistics

We use the SNV archives for the school year 2011-12 to evaluate the external monitoring program, and the SNV archives for the school year 2012-13 to evaluate the sanctions program. While the external monitoring program started in 2009-10, in this paper we focus on the 2011-12 wave, for which Invalsi provided us with cheating indicators that are comparable with the following wave. Conversely, the sanctions program was first implemented when the results of the SNV 2011-12 wave were returned to the schools, thus the effects on cheating behavior can only be expected from the following SNV assessment (i.e. SNV 2012-13). Using the first year of implementation of this policy is crucial for our empirical strategy: as the policy was not announced, it did not have any influence on the SNV 2011-12 wave that we exploit for identification.[14]

The SNV archives contain individual level records on students' test scores and basic demographic characteristics. We consider two grades covered by the SNV program: junior-high (grade 6) and high schools (grade 10). The CPI is computed by Invalsi from test scores statistics using 'fuzzy clustering' techniques; it is continuous and bounded between 0 (no cheating) and 1 (maximum likelihood that cheating occurred), and it can be interpreted as the probability that cheating occurred in each classroom during or after the test (Invalsi, 2010), and thus capturing both cheating on the part of students and teachers (see Table 1).[15] For our analysis, we average at the school level the CPI calculated by Invalsi for each class. We look separately at the CPI in the language and math test, though the two do not differ a lot. Notice that since the technique for the computation of the CPI does not take into account any class-level observable characteristics,

[14] From the universe of the schools in the SNV 2012-13 we exclude those subject to the external monitoring (about 15 percent), not to confound the effects of the two policies in the evaluation exercise. Given that the external monitoring program is administered in a representative and random sample of schools, we can safely exclude them from the analysis. Also notice that starting from the SNV 2012-13, the tests were produced in five different versions changing the order of the items (see Report SNV 2012-13 pag. 11). Since the change was the same for all grades and subjects, and given that we do not exploit the panel dimension of the SNV data, this feature is unlikely to affect our results.

[15] The CPI is computed by Invalsi using a fuzzy clustering technique on the principal component analysis of four main statistical indicators of 'suspected behaviors' (share of correct answers, share of missing answers, variability and homogeneity in the response patterns) suggested by the statistical literature on the detection of cheating in tests scores (Castellano et al. 2009). The fuzzy clustering technique allows to establish a probability of belonging to a certain group, rather than a simple dichotomous indicator, thus making possible to evaluate the intensity of the cheating phenomena, which is more intense and more likely as the CPI approaches 1. For further details see Castellano et al. (2009) and Invalsi (2010, 2011). See also section 3.

monitored and non-monitored classrooms are treated exactly the same way. This aspect is fundamental in order to implement the identification strategy illustrated in the next section.

We also compute an alternative measure of cheating behavior based on absenteeism on the day of the test, which we use as a proxy of teachers' and school principals' strategic pooling behavior. The *Absence rate* is computed as the share of students who are formally enrolled in the school in September but do not sit the test in May. As a matter of fact, this indicator might include both cases in which the pool of students who sit the test is selected by schools' staff, as well as cases in which families themselves (or students) decide not to send their children to school on the day of the test.[16] The indicator, of course, also includes students who are sick on the day of the test or those who change school during the school year, these circumstances, however, are expected to be negligible and likely to be randomly distributed across schools.

[Table 2]

Descriptive statistics concerning the external monitoring and the sanctions program are reported in Table 2. Overall, about 24 percent of the schools were subject to the external monitoring program (monitored schools). About a half of the schools were sanctioned, while in each school, on average, about 30 per cent of the classes received a sanction (either a correction or a non-return sanction). A non-negligible share of students, between 14 to 16 percent, were reported to be absent on the day of the test in both programs.

5. Empirical strategy

We exploit two different identification strategies to assess the effects (and the effectiveness) of the monitoring and sanctions programs on different types of opportunistic behaviors in the SNV Evaluation Program. The identification strategies are similar in spirit and both based on the random presence in the school of the external inspector.

5.1 The external monitoring program

If the presence of the external inspector in a given class were truly random (see section 2), the causal effect of the external monitoring policy on the outcomes could be estimated by simple OLS regressions at the class-level (Brunello et al. 2013). However, two main estimation problems are to be mentioned. First, as discussed in the previous sections, we cannot exclude that the school principal uses some discretion in the assignment of the external inspector to

[16] Unfortunately, it is observationally impossible to distinguish between the two. While it would have been useful to investigate whether low-performing students (e.g. non-natives or grade-retained) are over represented in the group of students absent on the day of the test, their socio-demographic characteristics are not available in the SNV data.

classes within a school – i.e. for example, selecting higher than average scoring classes as compared to those selected by Invalsi (Angrist et al. 2017). Second, the presence of the inspector in a given class can generate spillover effects on the behavior of students and teachers in the non-monitored classes within the same school – i.e. the so-called, monitored school (Bertoni et al. 2013). In this context, the assignment of the inspector is not independent from the quality of the students in the class - i.e. there is positive selection into treatment -, and the class-level model is likely to underestimate the true effect of the external monitoring program; alternatively, the presence of spillover effects may underestimate the monitoring effect at the classroom level (e.g. in a school fixed-effect model).

To take into consideration both types of potential bias, we run the analysis at the school level rather than at the class level. Aggregation at the school level allows us to account for both types of bias as school principal cannot manipulate the assignment of the external inspectors to schools, and both the direct and the spill-over effects of the external inspector are embedded in the school-level analysis. We thus specify the following equation:

$$y_{st} = \alpha_0 + \alpha_1 EI_{st} + X'_{st}\alpha_2 + \varphi_g + \varepsilon_{st} \quad (1)$$

where the outcome variable (y_{st}) is either the CPI or the *Absence rate* computed from the SNV 2011-12 for every school *s*. The variable EI_{st} indicates whether the school is treated (i.e. monitored by an external inspector), X'_{st} is a vector of school level characteristics (share of females, share of grade retained students, share of immigrant students, average class size and its square, school size and its square), and φ_g represents a set of fixed effects (FE) including: grade, high school type and province FE.[17] We also include the interaction between region FE and school size (in terms of number of students enrolled) to control for the strata of the random sampling scheme (Invalsi 2011).

5.2 The sanctions program

To evaluate the effectiveness of the sanctions program, we start from specifying the following equation:

$$y_{st} = \beta_0 + \beta_1 T_{s(t-1)} + X'_{st}\beta_2 + \varphi_g + \varepsilon_{st} \quad (2)$$

where dependent, y_{st}, control variables, X'_{st}, and fixed effects, φ_g, maintain the same definition as in equation (1) above, but, due to the different timing and implementation of the policy, they are computed on the SNV wave 2012-13 (i.e. the first wave after the implementation of the sanctions program). The variable $T_{s(t-1)}$ indicates whether a school was sanctioned or,

[17] Italian provinces (about 110, NUTS level 3) broadly correspond to the School Districts. Experimentations with alternative FE do not alter the results.

alternatively, the share of classes in each school *s* that received any sanction, correction or the non-return, capturing, respectively, the extensive and intensive margin of the treatment. The estimation of equation (2) by simple OLS would not deliver the causal effect of the policy since the definition of the treatment variables ($T_{s(t-1)}$) is itself a function of the level of cheating observed in the previous wave of the SNV (SNV 2011-12). In such context, serial correlation may bias our results since the same schools that are more likely to display a higher 'likelihood of cheating' are also more likely to be the target of the sanctions. Hence, a positive relationship between receiving the sanction (observed in SNV 2011-12) on cheating behavior (measured in SNV 2012-13) may partly depend from a spurious serial correlation.

To address this issue in our estimation, we use an instrumental variable approach based on the presence of the external inspector in the school in the SNV 2011-12 ($EI_{s(t-1)}$, is a dummy variable as specified in the previous section). Notice that the sanctions program was administered to the entire population of schools, independently on whether the school had received the monitoring treatment too (Falzetti, 2013). Hence, the presence of an external inspector in the SNV 2011-12 provides a random variation in cheating behavior on which the correction and non-return sanctions were determined (see section 6.1). In practice, schools that received the monitoring program, due to the (random) presence of the external inspector, also experienced a lower share of class scores corrected or non-returned.

One potential threat to our identification strategy, to be discussed, is related to the possibility that the presence of the external inspector in the previous wave of assessment has a direct effect on cheating behavior in the current wave, as this would violate the exclusion restriction. To address this concern, in the Robustness section, we use SNV waves prior to the introduction of the sanctions program, and show that there is no correlation between the presence of the inspector in one year and the school cheating behavior (both in terms of CPI and *Absence rate*) in the following year. Also a number of administrative procedures support this finding. First, most teachers, whose classes are tested one year, are unlikely to be under assessment in the following year. Second, teachers and school principals - by their experience in previous SNV waves (i.e. since the school year 2009-10) - were informed and accustomed to the external inspector monitoring program. Hence we argue that the presence of an inspector in the school is unlikely to have a significant and persistent effect on students' performance. The deterrence effect of monitoring, as discussed in the literature, comes directly from the contemporaneous presence of the inspector in the class at the time of the test, while there is no effect from the

presence of inspectors in previous SNV editions.[18] Moreover, conditional on school size, which we add as control variable in all the specifications, the probability of receiving an external inspector in any school year is independent from the presence of the inspector in the previous school year. This is because the random procedure that selects the monitored classes and schools is run every year by Invalsi without taking into account whether the inspector was present in the school in previous SNV waves (Invalsi 2010, 2011).

Finally, it is worth noticing that in the sanctions program considered here, the correction or non-return sanctions were originally targeted towards the class (at *t-1*), but effectively intended to warn the school as a whole about the cheating detected (Falzetti, 2013). Moreover, given the census and cross-sectional nature of the Invalsi SNV Evaluation Program (which every school year assesses the performance of the same school grades), the same class cannot be observed the following year, thus ruling out the possibility to implement a sharp regression discontinuity design exploiting the thresholds discussed in section 2.

6. Results

In this section, we investigate the impact of the different deterrence and sanctions programs on cheating behavior. We first estimate the effect of the external monitoring program on both cheating behavior and absence rates. Next, we focus on the effects of the sanctions program. We also analyze the heterogeneous impact of the above programs and discuss potential channels that might drive our results. The robustness checks are presented in the following section.

6.1 The external monitoring program

The baseline results of the effects of the external monitoring program are reported in Table 3. The presence of the inspector in a school is associated with a reduction in the propensity to cheat, as measured by the CPI, and an increase in strategic behavior, as measured by the *Absence rate*. In detail, the CPI decreases by about 1 percentage point (p.p.), corresponding to a reduction in cheating of about 0.1 standard deviation. The effect is similar across subjects and specifications (with or without the fixed effects). Conversely, the *Absence rate* increases by 1.2 p.p. (about 0.1 standard deviations). While results from the existing literature have documented the deterrence effect of the presence of an external inspector in monitored classes on test scores and cheating (Bertoni et al. 2013, Angrist et al. 2017, Pereda Fernàndez 2016), evidence on the

[18] As a matter of fact, also the other studies in the literature argue that it is not necessary to control for the presence of inspectors in previous years (Bertoni et al. 2013; Angrist et al. 2017).

existence of strategic pooling has been largely neglected. Only Figlio (2006) found evidence of selective pooling in the high stake context of the application of the No Child Left Behind Act in the U.S. public schools.

[Table 3]

Using the cheating taxonomy previously discussed (see Table 1), we interpret some of the results in light of the possible mechanisms driving cheating behavior in schools. For example, while we found that the presence of the external inspector is effective in reducing cheating behavior *during* and *after* the test administration, we also reported evidence of strategic responses, in monitored schools, which shifted cheating behavior *before* the test administration altering the pool of students who takes the test. While in principle such strategic behavior should be minimized by the protocol enforced by Invalsi that notifies only some days in advance the presence of the external inspector to the school, this span of time apparently does not eliminate the risk of manipulation of the composition of the students taking the test, for example by retaining low-performing students or students' own self-selection.

[Table 4]

In Table 4, we further analyze the role of the external inspector by documenting heterogeneous effects with respect to some relevant institutional characteristics which have been shown to be relevant determinants in the differences of cheating behavior in the literature, such as regional differences and trust. Available evidence has consistently found sizeable differences, both in achievement and cheating, between Northern and Southern regions in Italy (Paccagnella and Sestito 2014, Angrist et al. 2017, Battistin et al. 2017, Lucifora and Tonello 2015). This is in line with the long standing literature that suggests the existence of a deeply rooted divide in socio-economic, as well as cultural features across regions (Guiso et al. 2004). A North-South divide is apparent in both formal and informal institutions, as regions in the South are characterized by lower economic development, lower levels of trust and civicness, higher corruption and diffused organized crime (Pinotti, 2015).

In Panels A and B of Table 4, we report evidence on the different effectiveness of monitoring across regional clusters and splitting geographical areas according to the level of trust in institutions and in the collectivity (i.e. higher or lower than the country median).[19] We find that the deterrence effect of the external monitoring program on cheating is higher in the South of the country and in areas with low trust endowments (Guiso et al. 2004). Also the *Absence rate* is found to be larger in areas with lower levels of trust.

[19] The *High* and *Low Trust* subsamples are defined according to the school being located in a province (NUTS 3 level) above or below the median value of the variable *trust* as defined in Guiso et al. (2004), which is an index of the level of trust based on the *World Value Survey* for Italy run among 2,000 individuals in years 1990 and 1999.

6.2 The sanctions program

The effects of the sanctions program are reported in Tables 5 and 6. The treatment variable in Table 5 is defined as a dummy equal to 1 if the school received any sanction, and it is thus aimed at capturing the extensive margin of the treatment; the treatment variable in Table 6 is defined as the share of classes in the school that received any sanction, thus capturing the intensive margin of the treatment.

[Tables 5 and 6]

The results from the OLS regressions, shown in Panels A of Tables 5 and 6, display the expected positive correlation between cheating outcomes (especially in terms of CPI) and the sanctions, highlighting the fact that schools which have a higher propensity to cheat are also more likely to receive the sanction. The first stage regressions reported in the Appendix Table B.1 show that the presence of the inspector in the previous SNV wave (i.e. in the SNV 2011-12) decreases by about 3 p.p. the probability that a school is sanctioned, and by 4 p.p. the share of sanctioned classes. The F-statistics always reject the null that the instrument is weakly correlated to the treatment (Stock and Yogo 2005), while using the intensive margin definition of the treatment allows for a stronger first stage (as documented by the higher values of the F-statistics in Table 6 as compared to Table 5). The 2SLS results presented in Panels B of Tables 5 and 6 show no statistically significant effect of the sanctions on the CPI, and a decrease in the absence rates, when considering the intensive margin of the treatment (Table 6).

In terms of our taxonomy of cheating (see Table 1), these results can be interpreted as an indication that sanctions generally are not effective in settings characterized by complex cheating interactions (or manipulations), when the outcome is measured with error (the CPI) and it is revealed *ex-post*, several months later. Conversely, sanctions could work when the outcome is timely observed and measured (as with absence rates). In such context, schools which had been sanctioned in the past, may take *ex-ante* actions to reduce strategic pooling and students' absenteeism to avoid loss of reputation and a lower school's attractiveness.[20]

[Tables 7 and 8]

As before, in Tables 7 and 8, we explore the potential heterogeneous effects of the sanctions program across macro-regions and areas characterized by different levels of trust. Results for the extensive margin of the definition of the treatment variable (Table 8) show that the reduction in

[20] A much higher absence rate on the day of the test would negatively affect the school's ranking in the SNV and its overall reputation. While schools could also enforce other leverages, such as a stricter control on students' and/or teachers' behavior during the resting process, these are less likely to determine substantial changes in behaviors and show up in the statistical indicators of cheating.

the absence rates is mainly driven by the Northern regions of the country and by areas endowed with a higher level of trust. The sanctions program seems to work better where trust and institutional quality are higher, such that the potential loss of reputation associated with sanctions is costlier. Conversely, in areas characterized by low levels of trust and institutional quality sanctions appear unable to significantly change school behavior (Paccagnella and Sestito 2014).

6.3 Low-stake assessments and external validity: a discussion

One relevant question, given the low-stake nature of the assessment program considered here, concerns the external validity of our findings and their relevance for other institutional contexts. While the existing literature has generally detected the prevalence of cheating behavior in high-stake settings (e.g. Jacob and Levitt 2003, Figlio 2006, Dee et al. 2019), relatively less is known about the effectiveness of measures to deter cheating, as well as their success in different contexts (Dee and Jacob 2012, Dee et al. 2019). Moreover, even though the focus has generally been placed on direct monitoring programs (i.e. which reduce cheating during and after the test), our work also contributes to enlarge the set of policies which have been formally evaluated, also documenting the unintended consequences associated to such policies, such as the increase in strategic responses (i.e. shifting the timing of cheating behavior to moments before the test administration). Given that the literature has shown how opportunistic behavior typically increases along with the stakes held by the agents (De Geest and Dari Mattiacci 2014, Neal 2013), we believe that our findings on the effects of sanctions in a low-stake context can be extended to other low-stake environments, as well as to more traditional high-stake systems characterized by a more formal accountability system. Notice, however, that we can only speculate about such external validity, as we are only able to document the effectiveness of the sanction program on *ex-ante* opportunistic behavior such as strategic pooling and students' absenteeism.

7. Robustness

As discussed in the empirical strategy, a main threat to the identification of the effects of the sanctions on cheating is given by the validity of the instrumental variable. Indeed, the exclusion restriction would be violated if the presence of the external inspector in one year has *per se* a direct effect on the cheating of the same school the following year. While we have already discussed a number of institutional features that should exclude such correlation, here we provide a more formal test on the statistical relevance of the presence of the external inspector in a school, on the levels of cheating in the same school in the following year.

In the Appendix Table B.2 we report the main results of this exercise, using two consecutive SNV waves before the introduction of the sanctions program. In practice, we regress a dummy variable for the presence of the external inspector in the school (in the SNV wave 2010-11), on the cheating indicators of the same school in the following SNV wave (in the SNV wave 2011-12).[21] We do not find any statistically significant correlation suggesting that the presence of the inspector generally does not have direct effects on cheating behavior in the same school in the following year.

We also perform a complier characterization, in the spirit of Angrist (2004) and Angrist and Pischke (2009), to establish whether the population of the complier of our IV estimates differ substantially from the average school, in terms of observable characteristics. First, we discretize the school level covariates in dummies equal to 1 if the school attributes are above the median. The results show that set of compliers does not differ substantially from the average school in the sample (Appendix Table B.3). "Compliers" are in line with the average school in terms of female share and socio-economic status (ESCS), while non-native and grade-retained students are slightly more represented (13 to 14 p.p. more likely to be above the median). Compliers differ from the average only in terms of class and school size, but this is likely to be due to the randomization process that selects larger schools, and for which we control for in all specifications (Invalsi, 2011).

Finally, in Appendix Table B.4 we show that monitored and non-monitored schools are quite similar in terms of observable characteristics (share of females, non-natives and grade-retained students, average class size), thus confirming the randomness in the assignment process of the external inspector at the school level followed by Invalsi (see section 2). Monitored schools are larger in terms average number of students enrolled, for the reason we discussed above, and differ with respect to non-monitored ones in terms of average ESCS. This can be due to differences in the accuracy of reporting the additional information needed to compute the socio-economic indicator when the external inspector is present.

8. Conclusions and policy implications

A growing recent literature has shown that cheating in school standardized testing can be particularly disruptive as it contaminates the information provided by the educational system about student achievement, altering students' careers, and even their wages (Dee et al. 2019, Diamond and Persson 2016). The design of testing systems and incentives to reduce

[21] We exclude monitored schools in the 2011-12 wave.

opportunistic behavior appears to be key in preventing their long-term distortionary effects (Mechtenberg 2009). However, to date only few studies have investigated the effects of policy interventions aimed at curbing cheating and opportunistic behavior in evaluation programs.

In this study we evaluated the effectiveness of different policies introduced in the Italian SNV evaluation program to monitor and sanction schools which were identified as having a high likelihood of cheating and manipulation over the testing process. In particular, we compared and contrasted the effectiveness of two alternative policies: an external monitoring program, based on the presence of an external inspector for the administration and proctoring of the tests; and a "fame and shame" sanctions program, consisting in a correction or non-return of the test scores for classes with cheating patterns that exceed the statistical threshold of tolerance. In the empirical analysis we exploited a randomized experiment to estimate the causal effect of the monitoring and sanctions programs on cheating behavior and other forms of opportunistic behavior in school standardized testing.

The main findings show that higher monitoring is effective in deterring cheating: the presence of the external inspector in monitored schools reduces cheating propensity by 1 p.p., – i.e. about 20 percent when computed at the mean CPI – a figure that is lower with respect to previous findings in the literature on primary school grades. We also present new evidence suggesting that the external monitoring program can trigger a strategic response in schools shifting cheating *before* the testing process, such as altering the composition of the students who sit the test (i.e. strategic pooling and students' self-selection). The estimated share of students who are absent on the day of the test in monitored schools is about 8 percent higher, compared to non-monitored schools. Conversely, we report mixed results with respect to the "fame and shame" sanctions program. In particular, we found that schools, which have been sanctioned in the past year, do not significantly change their cheating propensity *during* (or *after*) the testing process. Sanctions, however, do show an effect on students' absence rates on the day of the test, suggesting that school which received a sanction are more likely to react when it comes to strategic pooling or students' absenteeism.

We also report evidence on the heterogeneous effects of the different programs across different contexts. First, the impact of monitoring on cheating is shown to be higher in Southern regions where cheating is more diffused. Second, the estimated effect of sanctions is lower or null in areas with low trust and poor institutional quality, suggesting that where the reputational cost of sanctions is smaller, sanctions programs are unlikely to significantly change opportunistic behavior.

These results provide a number useful insights concerning the effectiveness of policies directed at deterring cheating in standardized testing. Comparing alternative programs that target

cheating with direct monitoring or those that leverage on incentives and sanctions, we contribute to the debate on the design of proper accountability systems also discussing the cost-effectiveness of the programs. In this respect the Italian case is interesting as it shows that cheating behavior can be pervasive even in low-stakes environments, and not just in high-stakes settings as commonly believed.

Monitoring programs with external inspectors are found to be highly effective, since by increasing the probability of detecting cheaters in monitored classes and raising the cost of opportunistic behavior, they remove all opportunities for cheating during and after the test administration. It should be noted, however, that the amount of resources necessary to implement them are quite considerable. In the Italian SNV Evaluation Program, the total budget devoted every year to the external monitoring program (in a sub-sample of schools) is in the order of 1,500,000€, which corresponds to approximately 20 percent of the total budget devoted to the Evaluation Program.[22] Hence, while monitoring remains a fundamental pillar of any deterrence policy, it cannot be the solution for national programs that are run on a census basis (Borcan et al. 2017).

We highlight the challenges posed by schools' strategic responses to deterrence policies, when for example opportunistic behavior is shifted from one outcome to another. In this respect, to reduce the margins for schools, or students, to be able to game the systems – such as in strategic pooling or students' absenteeism –, the implementation of the external monitoring should not only be random but also unexpected on the part of the schools (i.e. without advance notice).

The effectiveness of the deterrence and sanctions programs crucially depends on the incentive structure implied by the alternative programs. The "fame and shame" sanctions program, by leveraging on the potential loss of reputation for schools' identified as cheaters, mainly works when sizeable losses are associated to the school's (lower) ranking or attractiveness. In this respect, we show that sanctions are generally ineffective when the latent cheating behavior originates from complex agents' interactions (or manipulations) which occur during (or after) the test but are measured with error (i.e. the CPI) and revealed only with a significant lag several months later, thus making the reputational cost more uncertain. Conversely, sanctions work better when the school's opportunistic behavior can be better and readily observed, such as with

[22] The figures reported are drawn from a " back of the envelope" exercise obtained multiplying the 200€ fee, each external inspector receives to complete the supervision and proctoring tasks for each single class, by the number of monitored classes in the SNV 2011-12. The shares are calculated with respect to the Invalsi Budget for the year 2012 (total revenues of about 3,700,000) (available at: http://www.invalsi.it/operazionetrasparenza/documenti/invalsi_bilancio_previsione_2012.pdf, accessed on February 28, 2019).

absence rates and strategic pooling. The lack of any obligation for the schools to make public their scores or sanctions is likely to reduce further the deterrence potential of sanctions programs.

Contextual factors such as culture and social norms also matter in the diffusion of opportunistic behaviors and in the efficacy of cheating deterrence. In particular, in contexts and areas where trust and institutional quality are poorer, the design of accountability systems should consider the reputational costs associated to the cultural and institutional setting. In this vein, we believe that the evidence resulting from this study shows that sanctions programs not embedded in a proper school accountability system, with poorly measured or delayed reporting of outcomes, as well as involving low expected reputational costs in terms of schools' ranking or attractiveness, are unlikely to provide an appropriate set of incentives and a suitable environment to reduce opportunistic behaviors.

While we cannot disentangle the exact mechanisms through which sanctions may or may not work, we interpret the findings reported in this study as evidence that schools are not effective in taking corrective actions, besides direct monitoring, when the accountability system provides poorly designed incentives and in context and areas where low levels of trust and poor institutional quality weaken the reputational costs of cheating behavior.

References

Anderman, E. M. and T. B. Murdock (2007). The psychology of academic cheating: who does it and why? Academic Press.

Ahn, T. and J. Vigdor (2014). The impact of No Child Left Behind's accountability sanctions on school performance: regression discontinuity evidence from North Carolina. NBER WP 20511.

Angrist, Joshua D. (2004). Treatment Effect Heterogeneity in Theory and Practice. The Economic Journal 114 (494): C52-83.

Angrist, Joshua D. and J.S. Pischke (2009). Mostly Harmless Econometrics: An Empiricist's Companion. Princeton University Press.

Angrist, J., E. Battistin, and D. Vuri (2017). In a Small Moment: Class Size and Moral Hazard in the Mezzogiorno. American Economic Journal: Applied Economics 9(4), 216-249.

Battistin, E., M. De Nadai, and D. Vuri (2017). Counting Rotten Apples: Student Achievement and Score Manipulation in Italian Elementary Schools. Journal of Econometrics 200 (2), 344-362.

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

Behrman, J. R., S. W. Parker, P. E. Todd, and K. I. Wolpin (2015). Aligning Learning Incentives of Students and Teachers: Results from a Social Experiment in Mexican High Schools. Journal of Political Economy 123(2), 325-364.

Borcan, O., M. Lindahl, and A. Mitrut (2017). Fighting Corruption in Education: What Works and Who Benefits? American Economic Journal: Economic Policy, 9(1), 180-209.

Brunello, G. and D. Checchi (2007). Does school tracking affect equality of opportunity? New international evidence. Economic Policy 22, 781-861.

Card, D., and L. Giuliano (2013). Peer effects and multiple equilibria in the risky behavior of friends. Review of Economics and Statistics 95(4), 1130-1149.

Carrell, E. S., F. V. Malstrom, and J. E. West (2008). Peer effects in academic cheating. Journal of Human Resources 63(1):173-206.

Castellano, R., S. Longobardi, and C. Quintano (2009). A fuzzy clustering approach to improve the accuracy of Italian student data. Statistica & Applicazioni 7(2):149-171.

Cohodes, Sarah (2016). Teaching to the Student: Charter School Effectiveness in Spite of Perverse Incentives. Education Finance and Policy 11(1)

Davis, F. S., P. F. Drinan, and T. Bertram Gallant (2009). Cheating in school. U.K.: Wiley-Blackwell.

Dee, S. T. and B. A. Jacob (2012). Rational Ignorance in Education: A Field Experiment in Student Plagiarism. Journal of Human Resources 47(2): 397-434.

Dee, T.S., W. Dobbie, B.A. Jacob, and J. Rockoff (2019). The Causes and Consequences of Test

Score Manipulation: Evidence from the New York Regents Examinations. American Economic Journal: Applied Economics, 11(3), 382-423.

De Geest, G. and G. Dari Mattiacci (2014). Carrots Versus Sticks, in Francesco Parisi ed. Oxford Handbook of Law and Economics, Oxford University Press.

Diamond, R. and P. Persson (2016). The long-term consequences of teacher discretion in grading of high-stakes tests. NBER WP No. 22207.

Estrada, R. (2019). Rules rather than discretion: teacher hiring in Mexico. Journal of Labour Economics, 37(2), 545-579.

Eurydice (2009). National testing of pupils in Europe: objectives, organization and use of the results. http://eacea.ec.europa.eu/education/eurydice/documents/.

Falzetti, P. (2013). L'esperienza di restituzione dei dati al netto del cheating. Workshop Metodi di Trattazione, Analisi e Correzione del Cheating. Roma, 8 febbraio 2013.

Figlio, D. N. (2006). Testing, crime and punishment. Journal of Public Economics 90(4-5), 837-851.

Finn, B. (2015) Measuring Motivation in Low-Stakes Assessments, Educational Testing Service, Research Report 15-19, Princeton, NJ.

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

Hanushek, E. and J. Rivkin (2003). Does Public School Competition affect Teacher Quality? In: The Economics of School Choice, C. M. Hoxby (Ed.), University of Chicago Press.

Invalsi (2010). Il Servizio Nazionale di Valutazione. Aspetti operativi e prime valutazioni sugli apprendimenti degli studenti (Rapporto completo) 2009-10. Invalsi (Roma).

Invalsi (2011). Il Servizio Nazionale di Valutazione 2010-11. Le rilevazioni degli apprendimenti A.S. 2010-11. Invalsi (Roma).

Jacob, B.A., and S. D. Levitt (2003). Rotten Apples: An Investigation Of The Prevalence And Predictors Of Teacher Cheating. The Quarterly Journal of Economics 118(3), 843-877.

Josephson Institute of Ethics (2011). Report on honesty and integrity. Los Angeles, CA.

Kleven, H. J., Knudsen, M. B., Kreiner, T., Pedersen, S., Saez, E., 2011. Unwilling or unable to cheat? Evidence from a tax audit experiment in Denmark. Econometrica 79(3): 651–692.

Lazear, P. E. (2006). Speeding, terrorism and teaching to the test. The Quarterly Journal of Economics 121(3), 1029-1061.

Lucifora, C. and M. Tonello (2015). Cheating and social interactions. Evidence from a randomized experiment in a national evaluation program. Journal of Economic Behavior and Organization 115(C), 45-66.

Martinelli, C., S. W. Parker, A. C. Pérez-Gea, and R. Rodrigo (2018). Cheating and Incentives:

Learning from a Policy Experiment. American Economic Journal: Economic Policy, 10 (1): 298-325.

McCabe, L. D. (2005). Cheating among college and university students: A NorthAmerican perspective. International Journal Educational Integrity 1(1).

Mechtenberg, L. (2009). Cheap talk in the classroom: how biased grading at school explains gender differences in achievements, career choices and wages. Review of Economic Studies 76, 1431-1459.

Neal, D. (2013). The Consequences of Using One Assessment System To Pursue Two Objectives. The Journal of Economic Education 44(4).

Neal, D. and D. W. Schanzenbach (2010). Left Behind by Design: Proficiency Counts and Test-Based Accountability. The Review of Economics and Statistics, 92(3),263-283.

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

Pereda Fernández, S. (2016). A new method for the correction of test scores manipulation. Working Paper No. 1047, Bank of Italy.

Pinotti, P. (2015). The economic costs of organized crime. Evidence from Southern Italy. The Economic Journal 125, F203-F232.

Schwager, R. (2012). Grade inflation, social background, and labour market matching. Journal of Economic Behavior & Organization 82, 56-66.

Stock, J. H. and M. Yogo (2005). Testing for weak instruments in linear IV regression. In D. Andrews (Ed.), Identification and Inference for Econometric Models, pp. 80-108. New York: Cambridge University Press.

UK Standard & Testing Agency (2013). 2012 Maladministration report. National Curriculum assessments.

US Department of Education (2009). Standards and Assessments Peer Review Guidance: Information and Examples for Meeting Requirements of the No Child Left Behind Act of 2001. http://www2.ed.gov/policy/elsec/guide.

Wall Street Journal (2008), The Survey on deceit. https://www.wsj.com/articles/SB121448862810107085.

Wollack J. A., A. S. Cohen, and R. C. Serlin (2001). Defining Error Rates and Power for Detecting Answer Copying. Applied Psychological Measurement 25.

Tables

Table 1. Cheating behavior in school

		Timing		
		Before the test	*During the test*	*After the test*
Agents	*Students*		(i) Copying or collaborating with peers (Carrel et al. 2008, Martinelli et al. 2018, Bertoni et al. 2013, Lucifora and Tonello 2015) (ii) Using prohibited materials or ICT tools (Dee and Jacob 2012)	
	Teachers	(i) Teaching to the test (Lazear 2006, Neal and Schanzenbach 2010, Cohodes 2016) (ii) Strategic pooling (Figlio 2006)	(i) Give suggestions to students and loose monitoring (Estrada 2019, Bertoni et al. 2013, Angrist et al. 2017, Lucifora and Tonello 2015)	(i) Manipulating students' test scores (Jacob and Levitt 2003, Dee et al. 2019, Diamond and Persson 2016) (ii) Shirking in correction procedures (Angrist et al. 2017)

Table 2. Descriptive statistics: dependent and control variables by type of program

	Mean	Standard deviation	N. obs.
Panel A: external monitoring			
Dependent variables			
Cheating Propensity (CPI) - Language test	0.065	0.138	12484
Cheating Propensity (CPI) - Math test	0.041	0.106	12484
Absence rate	0.144	0.114	12484
Other variables			
Monitored school	0.236	0.425	12484
Female share	0.482	0.190	12484
Non-native share	0.113	0.109	12484
Grade retained share	0.142	0.155	12484
Average ESCS	0.045	0.521	12484
Average class size	19.368	4.620	12484
School size	108.149	69.427	12484
Panel B: sanctions			
Dependent variables			
Cheating Propensity (CPI) - Language test	0.036	0.071	10290
Cheating Propensity (CPI) - Math test	0.049	0.083	10290
Absence rate	0.156	0.133	10290
Other variables			
Sanctioned school - Language test	0.53	0.50	10290
Share of sanctioned classes in the school - Language test	0.29	0.35	10290
Sanctioned school - Math test	0.55	0.50	10290
Share of sanctioned classes in the school - Math test	0.33	0.37	10290
Inspector in the school in the previous wave (instrument)	0.19	0.40	10290
Female share	0.483	0.184	10290
Non-native share	0.100	0.112	10290
Grade retained share	0.128	0.145	10290
Average ESCS	-0.026	0.520	10290
Average class size	21.071	4.203	10290
School size	75.288	61.959	10290

Notes. Sanctioned schools include those for which the test scores results of at least one class were returned with c*orrection* or *non-returned* in September 2012 (i.e. based on the results of the SNV 2011-12). Absence rate refers to the share of students who do not sit the test with respect to those formally enrolled. The SES is an indicator of the Socio-Economic Status of students' household: it is provided by Invalsi and standardized with 0 mean in the entire sample; the class and school size are expressed in terms of number of students enrolled.
Source: Invalsi SNV 2011-12, 2012-13.

Table 3. Monitoring: baseline results

	(1)	(2)	(3)	(4)	(5)	(6)
	CPI Language		*CPI Math*		*Absence rate*	
Monitored school	-0.020***	-0.013***	-0.008***	-0.009***	0.016***	0.012***
	(0.003)	(0.003)	(0.003)	(0.002)	(0.003)	(0.003)
N.Observations	12484	12484	12484	12484	12484	12484
Controls	yes	yes	yes	yes	yes	yes
Fixed effects		yes		yes		yes

Notes. CPI indicates the Cheating Propensity Indicator. School level regressions weighted by the number of students in the school (school size). The set of control variables includes school characteristics (share of females, grade-retained and non-native students, SES indicator, class size and its square, school size and its square as defined in Table 2). The set of fixed effects includes: fixed effects for the Italian provinces (110 provinces), grade fixed effects (grades 6 and 10), type of high school fixed effects (academic, technical, vocational), sampling strata controls (20 fixed effects for the Italian regions and their interaction with school size). Robust standard errors in parenthesis, clustered at the province level. Asterisks denote statistical significance at the * p<0.1, ** p<0.05, *** p<0.01 levels.
Source: Invalsi SNV 2011-12, 2012-13.

Table 4. Monitoring: heterogeneous effects

	(1)	(2)	(3)	(4)	(5)	(6)
	CPI Language		*CPI Math*		*Absence rate*	
Panel A: geographical areas	*North*	*South*	*North*	*South*	*North*	*South*
Monitored school	-0.008**	-0.024***	-0.003***	-0.019***	0.012***	0.011***
	(0.003)	(0.004)	(0.001)	(0.005)	(0.004)	(0.003)
N.Observations	7424	5060	7424	5060	7424	5060
Panel B: trust	*High*	*Low*	*High*	*Low*	*High*	*Low*
Monitored school	-0.007*	-0.021***	-0.003***	-0.015***	0.007**	0.016***
	(0.004)	(0.004)	(0.001)	(0.004)	(0.003)	(0.004)
N.Observations	6342	6142	6342	6142	6342	6142
Controls and fixed effects	yes	yes	yes	yes	yes	yes

Notes. CPI indicates the Cheating Propensity Indicator. School level regressions weighted by the number of students in the school. For the set of control variables and fixed effects see Table 3. Robust standard errors in parenthesis, clustered at the province level. Asterisks denote statistical significance at the * p<0.1, ** p<0.05, *** p<0.01 levels.
Source: Invalsi SNV 2011-12, 2012-13.

Table 5. Sanctions Program: baseline results (extensive margin)

	(1)	(2)	(3)	(4)	(5)	(6)
	CPI				*Absence rate*	
	Language test		*Math test*			
Panel A: OLS						
Sanctioned school	0.017***	0.013***	0.021***	0.008***	0.006	0.004
	(0.002)	(0.002)	(0.003)	(0.002)	(0.005)	(0.003)
Panel B: 2SLS						
Sanctioned school	0.056	0.060	-0.078	-0.030	-0.906	-0.210
	(0.049)	(0.072)	(0.131)	(0.060)	(0.638)	(0.140)
First stage F-statistic	10.32	4.51	2.38	8.18	2.38	8.18
N.Observations	10290	10290	10290	10290	10290	10290
Controls	yes	yes	yes	yes	yes	yes
Fixed effects		yes		yes		yes

Notes. CPI indicates the Cheating Propensity Indicator. School level regressions weighted by the number of students in the school. For the set of control variables and fixed effects see Table 3. The First stage F-statistics refers to the Kleibergen-Paap rk Wald F-statistics. Robust standard errors in parenthesis, clustered at the province level. Asterisks denote statistical significance at the * p<0.1, ** p<0.05, *** p<0.01 levels.
Source: Invalsi SNV 2011-12, 2012-13.

Table 6. Sanctions Program: baseline results (intensive margin)

	(1)	(2)	(3)	(4)	(5)	(6)
	CPI				*Absence rate*	
	Language test		*Math test*			
Panel A: OLS						
Share of sanctioned classes	0.049***	0.039***	0.049***	0.026***	0.017*	0.017**
	(0.003)	(0.004)	(0.005)	(0.005)	(0.010)	(0.008)
Panel B: 2SLS						
Share of sanctioned classes	0.037	0.039	-0.039	-0.020	-0.456**	-0.143*
	(0.030)	(0.041)	(0.057)	(0.040)	(0.183)	(0.078)
First stage F-statistic	42.47	20.95	12.14	46.57	12.14	46.57
N.Observations	10290	10290	10290	10290	10290	10290
Controls	yes	yes	yes	yes	yes	yes
Fixed effects		yes		yes		yes

Notes. CPI indicates the Cheating Propensity Indicator. School level regressions weighted by the number of students in the school. For the set of control variables and fixed effects see Table 3. The First stage F-statistics refers to the Kleibergen-Paap rk Wald F-statistics. Robust standard errors in parenthesis, clustered at the province level. Asterisks denote statistical significance at the * p<0.1, ** p<0.05, *** p<0.01 levels.
Source: Invalsi SNV 2011-12, 2012-13.

Table 7. Sanctions Program: heterogeneous effects (extensive margin)

	(1)	(2)	(3)	(4)	(5)	(6)
	CPI Language		*CPI Math*		*Absence rate*	
Panel A: geographical areas	*North*	*South*	*North*	*South*	*North*	*South*
Sanctioned school	0.060	0.088	0.018	-0.027	-0.516	0.238
	(0.131)	(0.056)	(0.059)	(0.101)	(0.357)	(0.161)
First stage F-statistic	1.39	6.28	3.12	7.21	3.12	7.21
N.Observations	6243	4047	6243	4047	6243	4047
Panel B: trust	*High*	*Low*	*High*	*Low*	*High*	*Low*
Sanctioned school	0.034	0.107	0.050	-0.076	-0.320	-0.005
	(0.166)	(0.073)	(0.051)	(0.135)	(0.216)	(0.257)
First stage F-statistic	0.60	6.88	3.90	4.21	3.90	4.21
N.Observations	5208	5082	5208	5082	5208	5082
Controls and fixed effects	yes	yes	yes	yes	yes	yes

Notes. CPI indicates the Cheating Propensity Indicator. School level regressions weighted by the number of students in the school. For the set of control variables and fixed effects see Table 3. The First stage F-statistics refers to the Kleibergen-Paap rk Wald F-statistics. Robust standard errors in parenthesis, clustered at the province level. Asterisks denote statistical significance at the * p<0.1, ** p<0.05, *** p<0.01 levels.
Source: Invalsi SNV 2011-12, 2012-13.

Table 8. Sanctions Program: heterogeneous effects (intensive margin)

	(1)	(2)	(3)	(4)	(5)	(6)
	CPI Language		*CPI Math*		*Absence rate*	
Panel A: geographical areas	*North*	*South*	*North*	*South*	*North*	*South*
Share of sanctioned classes	0.052	0.054*	0.014	-0.016	-0.401***	0.140**
	(0.099)	(0.030)	(0.045)	(0.057)	(0.138)	(0.068)
First stage F-statistic	5.72	33.24	13.95	38.22	13.95	38.22
N.Observations	6243	4047	6243	4047	6243	4047
Panel B: trust	*High*	*Low*	*High*	*Low*	*High*	*Low*
Share of sanctioned classes	0.033	0.062*	0.066	-0.030	-0.424**	-0.002
	(0.148)	(0.038)	(0.056)	(0.049)	(0.212)	(0.103)
First stage F-statistic	2.28	34.95	7.69	57.88	7.69	57.88
N.Observations	5208	5082	5208	5082	5208	5082
Controls and fixed effects	yes	yes	yes	yes	yes	yes

Notes. CPI indicates the Cheating Propensity Indicator. School level regressions weighted by the number of students in the school. For the set of control variables and fixed effects see Table 3. The First stage F-statistics refers to the Kleibergen-Paap rk Wald F-statistics. Robust standard errors in parenthesis, clustered at the province level. Asterisks denote statistical significance at the * p<0.1, ** p<0.05, *** p<0.01 levels.
Source: Invalsi SNV 2011-12, 2012-13.

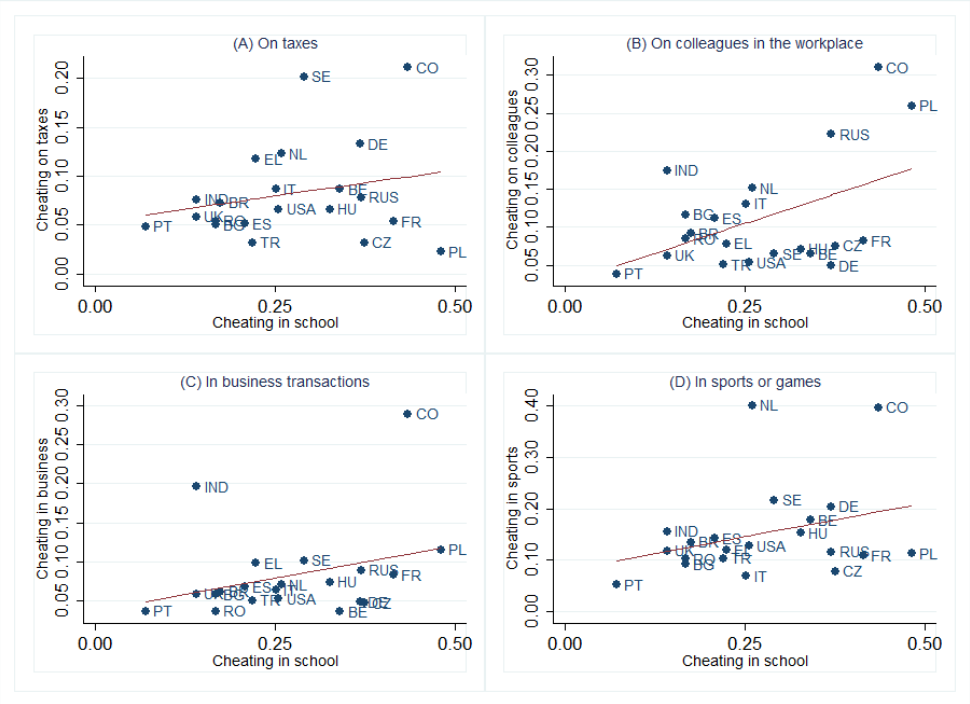
Appendix

(intended for Online publication only)

Appendix A. The prevalence of cheating: stylized facts

A number of surveys document the increase in opportunistic behavior and cheating practices that occurred over the last decades in test-based evaluation programs (Davies et al. 2009). McCabe (2005) surveyed 80,000 students and 12,000 faculties in the U.S. and Canada between 2002 and 2005, and reported evidence that 21% of undergraduates admit to have cheated on exams at least once a year. A survey conducted in 2010, on a representative sample of U.S. public and private high schools students, found that 59.3% of the students interviewed affirm to have cheated at least once during a test, while more than 80% say they have copied form others' homework at least once (Josephson Institute of Ethics, 2011).

Figure A.1. Cheating at school and in other fields.



Notes. The scatter plots show the correlation between the share of cheaters at school or university defined as the share of individuals answering 'Yes' to the question 'Have you personally ever cheated at school or university?', and the share of cheaters in other fields. The line depicts the linear fit. **Source:** based on Survey on Deceit, The Wall Street Journal (2008).

A similar survey conducted by The Wall Street Journal (2008) on national representative sample of individuals across different countries provides additional details on individuals' perceptions about the diffusion of cheating practices.[23] Cheating practices at school or university

[23] The 'Survey on Deceit' was conducted by the market-research private enterprise *GfK Custom Research Worldwide* in April and March 2008 on behalf of The Wall Street Journal, which published the results and the data

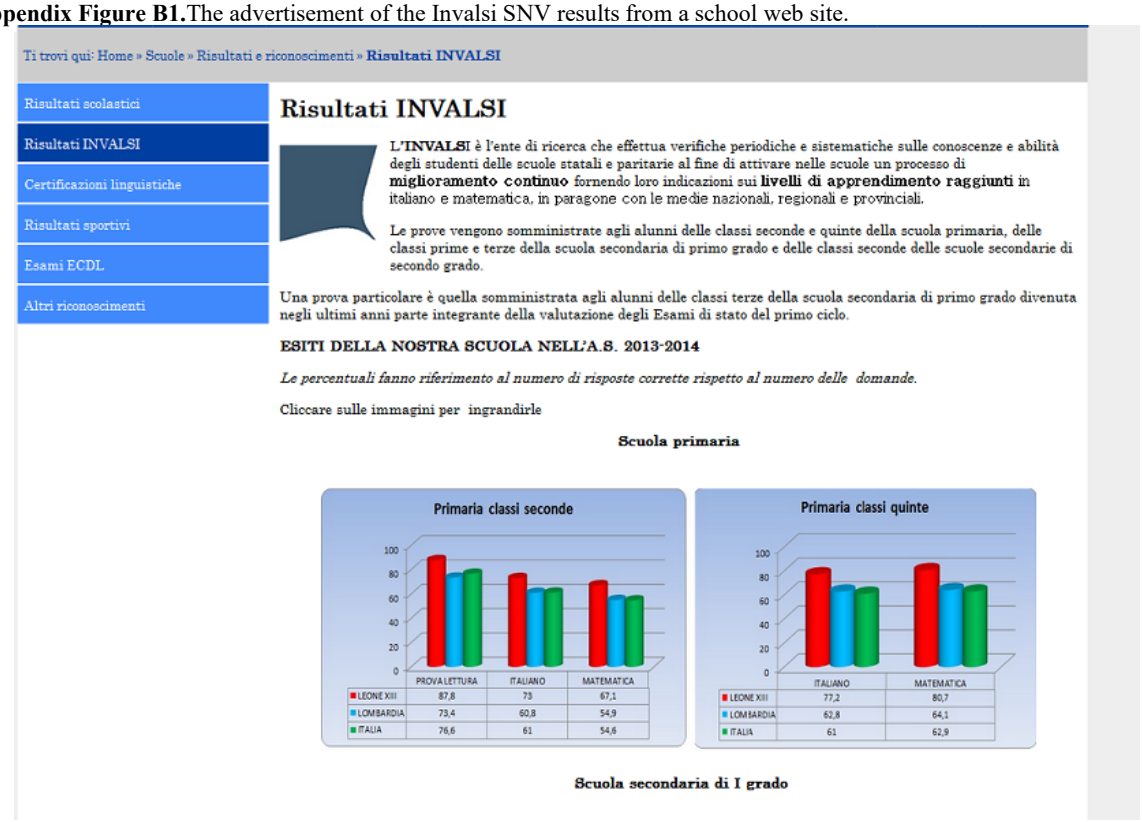
– as declared by the respondents with respect to their own experience – is widespread across countries: on average 28% of the respondents admitted to have ever cheated at school, ranging from a low 15% in the UK, to a figure of 37% in Germany and Russia, up to a 41% in France. Moreover, when restricting the focus to younger individuals (aged between 14 and 29), the above figures increase substantially in all countries (44% on average, and 59 and 66% in Italy and France, respectively).

Cheating practices are not confined to the schooling system, as they often reflect societal values and norms and extend to other domains. Figure A.1 shows that the prevalence of cheating practices at school are positively correlated to cheating on taxes, on business.

in June 2008. The survey covered about 20,000 individuals (older than 13) in 20 countries (16 European countries, plus Russia, Turkey, India and the US), focusing on a wide range of issues such as: taxes, business, academics, sports and romantic relationships. Here we prevalently focus on academic cheating (i.e. cheating in school or university). The survey was conducted face-to-face or by telephone interviews.

Appendix B. Additional Figures and Tables

Appendix Figure B1. The advertisement of the Invalsi SNV results from a school web site.



Notes. The figure shows the page of a school website showing the results in the SNV 2013-14.

Appendix Table B.1. First stage regressions: the effect of the external monitor in the school year 2011-12 on sanctions to schools

	(1)	(2)	(3)	(4)
	Sanctioned school		*Share of sanctioned classes*	
	Language test	*Math test*	*Language test*	*Math test*
School monitored in the previous wave	-0.026**	-0.027***	-0.040***	-0.040***
	(0.012)	(0.010)	(0.009)	(0.006)
N.Observations	10290	10290	10290	10290
Controls and fixed effects	yes	yes	yes	yes

Notes. See Table 3 for the list of controls and fixed effects. Robust standard errors in parenthesis, clustered at the province level. Asterisks denote statistical significance at the * p<0.1, ** p<0.05, *** p<0.01 levels.
Source: Invalsi SNV 2011-12, 2012-13.

Table B.2. Robustness checks: indirect effects of monitoring across years

	(1)	(2)	(3)	(4)	(5)	(6)
Dependent variables:	*CPI in 2011-12 wave*				*Absence rate*	
	Language test		*Math test*		*in 2011-12 wave*	
Panel A: all grades						
School monitored in 2010-11 wave	-0.001	0.001	0.003	0.001	0.004	0.003
	(0.003)	(0.003)	(0.004)	(0.003)	(0.003)	(0.003)
N.Observations	9156	9156	9156	9156	9156	9156
Panel B: junior-high school						
School monitored in 2010-11 wave	-0.015	-0.006	0.006	-0.006	-0.000	0.007
	(0.010)	(0.010)	(0.013)	(0.010)	(0.007)	(0.006)
	5556	5556	5556	5556	5556	5556
Panel C: high school						
School monitored in 2010-11 wave	0.001	0.003	0.002	0.003	0.005	0.001
	(0.003)	(0.003)	(0.004)	(0.003)	(0.003)	(0.004)
N.Observations	3600	3600	3600	3600	3600	3600
Controls	yes	yes	yes	yes	yes	yes
Fixed effects		yes		yes		yes

Notes. CPI indicates the Cheating Propensity Indicator. School level regressions weighted by the number of students in the school. For the set of control variables and fixed effects see Table 3. Robust standard errors in parenthesis, clustered at the province level. Asterisks denote statistical significance at the * p<0.1, ** p<0.05, *** p<0.01 levels.
Source: Invalsi SNV 2011-12, 2012-13.

Appendix Table B.3. Complier characteristics ratios

Variable	*Complier characteristics ratios*
Female share	1.086
Non-native share	1.135
Grade retained share	1.149
Average ESCS	1.046
Average class size	1.266
School size	1.492

Notes. The table reports the relative likelihood that compliers have the characteristics indicated on the left above the median value in the each sample.
Source: Invalsi SNV 2012-13.

Appendix Table B.4. Mean differences in outcome variables and observable characteristics between monitored and non-monitored schools

	All schools	*Monitored schools*	*Non-monitored schools*
Dependent variables			
Cheating Propensity (CPI) - Language test	0.065	0.046	0.071
Cheating Propensity (CPI) - Math test	0.041	0.032	0.044
Absence rate	0.144	0.159	0.139
Other variables			
Monitored school	0.236	1.000	0.000
Female share	0.482	0.481	0.482
Non-native share	0.113	0.116	0.113
Grade retained share	0.142	0.154	0.138
Average ESCS	0.045	0.003	0.059
Average class size	19.368	20.833	19.080
School size	108.149	142.313	97.579
N	12484	2950	9534

Source: Invalsi SNV 2011-12.