Monitoring and sanctioning cheating at school: What works?

Evidence from a national evaluation program

This version: June 2016

(Preliminary: Do not quote without permission)

Claudio Lucifora[*] Marco Tonello[♦]

Abstract

The diffusion of evaluation programs, along with the higher stake they account for, has also increased the prevalence of opportunistic behavior and cheating practices. This paper investigates the efficacy of different policy measures aimed at fighting cheating behaviors in schools. We exploit a classroom-based randomized experiment in Italian public schools, which assigned an external inspector to monitor the administration and marking of the tests, as well as different sanctioning mechanisms for schools suspect of cheating. We find that higher monitoring is effective in deterring cheating at all grades, while sanctions in general have no effect, and, under specific circumstances, may also trigger schools' strategic behaviors, such as selective pooling. The ineffectiveness of sanctioning schemes is explained by the fact that they are not embedded in a proper school accountability system.

JEL Classification: I21, I28, K42

Keywords: cheating, monitoring, incentives

[*] Università Cattolica and IZA (claudio.lucifora@unicatt.it). Corresponding author.

[♦]Bank of Italy, Economic Research Department, Law and Economics Division (Rome) and CRELI, Università Cattolica (marco.tonello@bancaditalia.it).

We are very grateful to Patrizia Falzetti and Valeria Tortora (Invalsi) for making the data available and for their guidance and insights in using them, and to Santiago Pereda Fernàndez and Sergio Longobardi for sharing with us their statistical indicators of cheating. We received helpful suggestions and comments from Lorenzo Cappellari, Paolo Sestito, Ylenia Brilli and seminar participants at the CRELI-Università Cattolica and the Invalsi Workshop 'Call for Ideas 2012-2014'. The views expressed in this papers 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, as well as to allocate resources across schools. The diffusion of national evaluation programs, along with the higher stake they account for, has also increased the prevalence of opportunistic behaviors and cheating practices in schools. Several surveys and a flourishing economic and psychological literature have documented the increase in cheating that occurred worldwide over the past years in all grades of the schooling system (Davies et al. 2009, Dee at al. 2011, 2016). Cheating practices and other types of dishonest behaviors are not confined to the education system but extend to other field such as untruthful tax declarations, free-riding on public goods, shirking on colleagues at work, as well as cheating in sports or in games (Kleven et al. 2011; Card and Giuliano 2013). However, the effect of cheating in school can be particularly disruptive due to the long-run effects that the misallocation of resources generates (Mechtenberg 2009). In particular, test scores manipulation 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).

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 countries (e.g. Germany, Italy, United Kingdom, Sweden, Mexico). These programs have generated considerable controversy over the benefits of the greater transparency and accountability that evaluation programs are expected to provide, and the costs associated to teach-to-test practices and the potential disruptive effects of probation programs and closure for under-performing schools on most disadvantaged students (Schultz et al. 2007, Fryer et al. 2012). Moreover, testing systems generate incentives for teachers, students and school administrators to manipulate the scores in order to achieve a better performance introducing distortions in the allocation of resources across schools, as well as in the information about student achievement. This is aggravated in countries where testing programs link teachers' career progress and monetary benefits to students achievements (Ahn and Vigdor 2014). Even in low-stake settings, where no explicit incentives are attached to testing outcomes, reputational concerns may induce opportunistic behaviors within schools to improve their performance indicators.

The institutional setting, the procedures adopted for the administration of the tests and a poor monitoring process can often exacerbate moral hazard problems. Typically, national evaluation programs are administered by teachers of the same school and only in some cases supervised by external inspectors.[1] Although National Authorities set precise tasks and protocols to ensure

[1] In the majority of U.S. states, the same class teachers administrate and mark the tests, but often a random sample of classes are asked to repeat the test after some weeks in order to validate the entire results. For instance, in the Chicago Public School system this happens for roughly 100 classes per year (about 2-3% of CPS classes) (Jacob and Levitt, 2003a).

adequate handling of the assessment procedure and to avoid manipulations of the scores, cases of cheating are often reported (U.S. Department of Education 2009, Eurydice 2009, UK Standard & Testing Agency 2013).

This paper contributes to the emerging literature that tries to evaluate the efficacy of different policies directed at monitoring and sanctioning dishonest in standardized testing, as well as the strategic responses through which opportunistic behavior can occur in schools. Taking advantage of a randomized experiment in the Italian national evaluation program, we compare the effectiveness of two different programs, aimed at reducing cheating in national standardized tests, either increasing the level of direct monitoring on students and teachers during the administration and correction of the test, or building a system of reputational sanctions for schools. We also document whether the two incentive schemes determine unintended side effects on students' psychological well-being.

A number of studies have documented the diffusion of cheating practices across countries, especially in universities and colleges (Carrell et al. 2008, Anderman and Murdock, 2007, Davis et al. 2009). A survey conducted in 2010 on a representative sample of U.S. public and private high schools students reported that 59.3% of the students interviewed declared to have cheated at least once during a test, while more than 80% copied form others' homework at least once in their life (Josephson Institute of Ethics 2011). Cheating practices at school or university is widespread across countries: on average 28% of the respondents of the 'Survey on Deceit' (The Wall Street Journal, 2008) 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.[2]

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 and to empirically identify and measure cheating practices in test score evaluation programs (Behrman et al. 2015, Martinelli et al. 2015, Dee et al. 2016, Diamond and Persson 2016). In this respect, the Italian schools system has proved an interesting case study, as documented by several recent works (Bertoni et al. 2013, Sestito and Paccagnella 2014, Angrist et al. 2015, Lucifora and Tonello 2015, Pereda Fernàndez 2016), and by the statistical indicators aimed at detecting opportunistic behavior during the administration of the test produced by the Italian National Education Authority (Invalsi 2010, Casellano et al. 2009).[3] However, limited evidence is available on the effectiveness of policies aimed at reducing the extent of cheating practices in schools. To the best of our knowledge, only Dee and Jacob (2012) show the effectiveness of online tutorials to reduce cheating, although their work is focused on a very specific form of academic cheating (i.e. plagiarism in take-home assignments).[4]

We expand on previous papers in the literature that investigated the effectiveness of deterrence and sanctioning measures on opportunistic behavior in schools along three main dimensions. First, we provide evidence on the effectiveness of different policy tools aimed at reducing illicit behavior in

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

[3] In a similar way, the Mexican Federal Secretary of Education calculates indicators of dishonest behavior in the national assessment program run every year in Mexican schools (Estrada 2015).

[4] Dee at al. (2016) also show that manipulation in the New York Regents Examinations disappear when passing from a local to a centralized grading system.

education. We contrast the effectiveness of either increasing deterrence (through higher monitoring) or introducing non-monetary incentives schemes in education, based on reputational mechanisms. Building our analysis on a national evaluation program, run on yearly and census basis, our results can be also easily generalized and extended not only to other education systems, but also to other contexts (e.g. dishonest behavior in public employment). Second, we document the effects of such policies on students' anxiety, both before and during the test administration. In this respect, we also contribute to the literature on the unintended effects of the introduction of evaluation programs in education on students' health and psychological well-being (Figlio and Winicki 2005, Bockhari and Schneider 2011). Finally, our work contributes to the recent literature on the effectiveness of monitoring tools in combating corruption and dishonest practices in education (Reinikka and Svensson 2011, Borcan et al. 2016).

We evaluate the effectiveness of the different programs using a statistical indicator of cheating and other indicators of illicit behaviors, such as strategic pooling. Moreover, we also match the administrative archives with a follow-up survey to obtain additional information on motivational questions concerning the test that allow us to investigate behavioral externalities such as students' stress and anxiety. Within the national evaluation program, we test the effectiveness of two alternative policies: the external monitoring program, based on the presence of an external inspector for the administration and marking of the tests; and the sanctioning program, consisting in a correction or non-return of the test scores for classes suspect of cheating. Our empirical strategy exploits a classroom-based randomized experiment run by the National Education Authority (Invalsi) in Italian public schools, which assigned an external inspector to monitor the administration and marking of the tests, as well as different sanctioning mechanisms for schools suspect of cheating.

Our main findings show that higher monitoring is effective in deterring cheating at all grades, while we find limited evidence that the 'reputational' costs associated to the sanctions may have any significant effect. We also find that higher deterrence measures and sanctions increase schools' strategic behaviors, such as selective pooling, while there is little evidence about the negative effects on students' psychological well-being. We interpret the ineffectiveness of sanctioning schemes by the fact that they are not embedded in a proper school accountability system.

The paper is organized as follows. Section 2 presents the conceptual framework. Section 3 illustrates the institutional setting. Section 4 describes the data used and provides descriptive statistics. Section 5 illustrates the empirical strategy, while section 6 presents the main results and the robustness checks. Section 7 extends the results analyzing behavioral responses in students' well-being. Section 8 concludes and derives policy implications.

2. Cheating in schools

2.1. A conceptual framework

Test-based evaluation programs often include a set of explicit, or implicit incentives for students, teachers and schools. In high-stake settings, teachers' career progress and monetary benefits are often linked to students achievements (Lavy 2009; Fryer et al. 2012); less frequently financial

incentives are offered to students for meeting some given standard of performance based on attendance, behavior, as well as standardized test scores (Levitt et al. 2016). In school-based accountability systems (such as the NCLB) schools can be rewarded when they meet the required standards, or they can be sanctioned if they fail to meet standards or make adequate progress (Gershenson 2015). The practice of using test-scores in high-stakes assessment system to measure the performance of teachers and schools has been shown to generate a number of undesirable effects, such as: inducing teachers to focus on tested subject only, inflating test scores and manipulating public information on student achievement (Neal 2013).

Even in low-stake settings, with no explicit targets and rewards attached to student achievement, test-scores evaluation programs may generate incentives for opportunistic behavior. Students competing with each other to get a good score may be induced to cheat either exchanging information or using prohibited material (Martinelli et al. 2015). Teachers, whose reputation is likely to depend - among other things - on students' performance in the assessment exercise, may be persuaded to alter the results either helping students or directly manipulating the answer sheet of the test during the proctoring phase (Jacob and Levitt 2003; Dee et al. 2016; Diamond and Persson 2016; Angrist et al. 2015). School's principals might also be concerned about the ranking achieved by the school in the evaluation exercise as a way to attract more and better students, and may encourage teachers to concentrate their effort on test taking skills or engage in strategic pooling of students (Figlio 2006). Overall, since the effort required to achieve a good performance is typically unobserved and is costly for all the actors involved, the system both in high- and low-stake environments is likely to produce incentives for opportunistic behaviors. The institutional setting and the procedures adopted for the administration of the tests can often exacerbate moral hazard problems. Typically, national testing programs are administered and proctored by teachers of the same school and only in some cases supervised by external inspectors.[5] A poor monitoring process and the absence of sanctions, by reducing the cost of opportunistic behavior, may also increase the probability of cheating occurring (Bertoni et al. 2013). Even if cases of cheating are frequently reported, sanctions for illicit behaviors are not commonly used in schools.

The above setup can be framed as a standard agency problem. The Principal (the State) cares about the outcome of the educational production function and implements a test-score technology relating measured education achievements to different inputs. While some of these inputs can be directly controlled by the Principal (number of teachers, class-size, infrastructures, etc.), in our analysis we place particular emphasis on the discretionary actions of the Agents (i.e. teachers and students) in terms of cheating behaviors (or lack of effort), that are key to the educational outcome. Notice that, since illicit behaviors are unobservable and can take different forms, contingent contracts cannot be generally signed. In this context, stricter monitoring – by increasing the probability of cheaters being caught – is an effective way to reduce opportunistic behaviors, at least during the administration of the test. However, it should be noted that extensive monitoring is not generally feasible – or cost-effective – in national evaluation programs that are run on a census basis.

[5] Contrary to international programs of students' assessments (e.g. PISA, TIMSS, PIRLS) which are usually conducted on a survey basis and sampled students sit the test under the supervision of inspectors who are responsible to validate the overall testing procedure, national evaluation programs are conducted on a census basis and proctored by teachers of the same school.

Alternatively, incentives can be designed as threat to punish for violators, rather than rewards for compliers, with sanctions or penalties introduced to 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).[6] One problem with the sanctioning measures is that cheating can only be inferred through statistical indicators that 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.

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

[Table 1 about here]

Students' cheating materializes prevalently during the test and can take two main forms: collaborative effort during the test in exchanging information or copying from the peers, or using prohibited materials and technologies. In both cases, students' cheating can be interpreted as a form of 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. 2015).

Conversely teachers' contribution to cheating behavior can take several forms. First, before the test, cheating may occur as 'strategic pooling' when educators attempt to raise a 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 focusing 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; Cohodes 2016). Third, during the administration 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 student 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 proctoring 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. 2016; Diamond and Persson 2016; Angrist et al. 2015).

2.2. Measuring cheating

Cheating in school is typically unobserved and thus difficult to measure and sanction. While ethical codes, which define and discipline illicit behaviors, have been introduced to restraint the diffusion

[6] 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.

of cheating, still most opportunistic and dishonest behavior are very often overlooked and tolerated within many schools. However, even if 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.[7] Building on that literature, with the purpose of measuring and monitoring the propensity to cheat in Italian schools, the National Education Authority (Invalsi) has developed, since 2009, a Cheating Propensity Indicator (CPI) that is class and subject specific. The CPI can be interpreted as the probability that cheating occurred in each classroom during the test: typically, a classroom is suspect of cheating the more homogeneous is the pattern of responses and non-responses to each single item in each of the part the test, and the higher is the average and the lower is the variability of the scores (Castellano et al. 2009). For the purpose of the present study, we obtained from Invalsi an *ad hoc* calculation of the CPI for the universe of Italian schools in the various grades, which 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 a class-level indicator of strategic pooling (SPI), based on the share of students absent on the day of the test.

Some caveats are in order. First, the CPI approach is likely to underestimate the incidence of cheating, since only major and systematic manipulation are likely to be captured, while more subtle or moderate cheating behavior are likely to go undetected. Second, we cannot exclude the presence of 'false positive' in the CPI indicator. In other words, some classes, due to some exceptional circumstances, such as high ability and very homogeneous pool of students, may be coded as suspect of cheating while they are not. These cases, however, should be quite negligible in our data, and we also test the robustness of our results using an alternative indicator of cheating, less subject to this issue. Third, from the CPI indicator it will prove almost impossible to disentangle the contribution of students and teachers to the observed cheating behavior, that said for student to be able to cheat there must always be a negligent (or benevolent) attitude of the teacher. We return the details of the CPI and SPI indicators and some descriptive statistics to a later section.

3. Institutional context

3.1 The Italian school system and the SNV Evaluation Program

[7] 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 patterns of suspect 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. (2015) 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 (2015) exploits indicators based on Error Similarity Analysis and calculated by the Mexican Federal Secretary of Education.

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). The primary and junior-high schools are compulsory for all students. 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 track chosen: academic high schools last for five years and mainly prepare for college; technical high schools also last for five years, while the vocational ones last for two years and mainly endow students with the technical 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 school system is organized in institutions and schools: each institution is managed by a School principal and by an administrative body and it may counts one or more schools that can be also settled in different municipalities. Primary school is quite different from junior-high and high schools in terms of organization and types of teaching activities. In primary schools, pupils have two reference teachers, teaching different subjects, who usually follow them from the first to the fifth grade, establishing a strong personal link. On the contrary, 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 length of time each teacher passes in the classroom in junior-high and high school is considerably less than in primary school, as it is the formation of interpersonal relationships between students and teachers.[8]

The National Education Authority (Invalsi, from the Italian acronym) 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 all students in grades 2 and 5 (primary school), grade 6 (junior-high school), and grade 10 (high school) sit in two consecutive days in the month of May a language and a math test. In one of the two days, students (with the exception of grade 2) are also asked to fill in a Student Questionnaire containing, among others, questions on their feelings and stress before and while sitting the test.

In order to minimize illegal and opportunistic behaviors, Invalsi enforces a strict protocol for the administration and making of the test (Invalsi, 2010). Students are proctored by teachers from the same school, although from a different class and specialized in a different subject with respect to the one tested. The answer sheet of each student is marked by school teachers, but all teachers must do the marking contemporaneously so to cross-check each other. Then, each teacher transcripts the answers of all students into one single answer sheet for the entire class and send it to Invalsi for the collection of the data and the calculation of the scores. The test scores records, by class, are sent back to each school in September.

The SNV Evaluation Program has not a high-stake nature: the test scores are sent by Invalsi to the school principal, to the teachers and to the stake holders (e.g. the bodies of the representatives of the students or parents) with the purpose of documenting the situation of each school and class, and

[8] The passage from primary to junior-high school also coincides with the transition from the childhood to the pre-adolescent and adolescent periods, which has implications on the teacher-student relationship, since the teacher is no longer perceived as a sort of parental care-giver, but more as an assertive and detached instructor.

also to document the changes in performance from one year to another. The results of the SNV Evaluation Program are not embedded in a proper school accountability system to give additional funding to schools or teachers. Rather, they are intended for school principals and teachers to assess school and class performance. Schools are allowed to use their SNV score results to increase enrollments and to attract more and better students.[9]

3.2 Policy measures to fight cheating behavior in schools

Invalsi has implemented several deterrence measures to fight cheating since the first wave of the SNV Evaluation Program. These involve increased monitoring, to prevent cheating *ex-ante* (implemented from school year 2009-10), and sanctions for schools suspect of cheating, to punish cheating *ex-post* (implemented from school year 2012-13).

The external monitoring program. In every SNV wave, 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. We define as a 'treated class' a class for which the test is proctored and marked by an external inspector, and a 'treated school' a school in which there is at least one treated class. The external inspector represents a random and unexpected shock altering the monitoring technology: it increases monitoring by establishing a 'non-cheating' environment, where the possibility of cheating on the part of both students and teachers, both during and after the test, is remarkably reduced.[10]

The external inspectors are sampled each year from a pool of retired teachers. The school principal is made aware of the presence of the external inspector in his school a few days before the SNV tests take place, so that some opportunistic behaviors are still possible (Angrist et al. 2015). For instance, the school principal could manipulate the choice of the class that the external inspector is intended to proctor to favor a better class (we will come back to this point while discussing the identification strategy), as well as he or the teachers might be able to select the pool of students that sit the test.

The 'correction/non-return' sanctioning program. After the SNV 2011-12 was completed, but before the scores were returned to the schools, Invalsi implemented a new sanctioning policy to reduce cheating based on a 'fame and shame' mechanism. The policy was totally unexpected by the schools and implemented in September 2012 (Falzetti, 2013). It consisted of two different measures: (i) a *correction* (deflation) of the class test scores, or (ii) a *non-return* of the class test scores when there was a high suspect that cheating occurred during the administration of the SNV 2011-12, in May 2011.

[9] See for example the Appendix Figure B.1 which shows a school web page advertising the recent SNV results. However, this is not compulsory: there are not official 'School League Tables', as it is done, for instance, in the UK. Nevertheless, starting from the school year 2014-15 all schools are recorded on the official web pages of the Ministry of Education. In these web pages the schools are obliged to disclose several information about their facilities, programs and teachers (*http://cercalatuascuola.istruzione.it/cercalatuascuola/*). Among other things, the schools are also obliged to publish their average performance in the previous SNV.

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

In details, a strict algorithm was implemented when returning test scores of the 2011-12 SNV to the schools in September 2012. The algorithm combines the statistical indicator of cheating, 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*).[11] The implementation and the strictness of the sanctioning policy was then set according to the CPI_{cgj} with respect to the level of the $TCSA_{gj}$. Hence, the test score results of each class could be returned to the school either without any correction, or returned with a 'correction' using a cheating deflator, alternatively in cases of high levels of cheating the results could be not returned at all. The above measures thus introduced *de facto* three different regimes:

a) if $CPI_{cgj} \leq TCSA_{gj}$, the scores were returned without any correction;
b) if $TCSA_{gj} < CPI_{cgj} \leq 0.5$ Invalsi applied a statistical correction and returned to each class the 'scores corrected for cheating':[12] the school is made aware of the fact that the correction procedure has been applied because of high suspect that cheating occurred;
c) if $CPI_{cgj} > 0.5$ the scores of the class were not returned (and they were excluded from the calculation of the school average score).[13]

Hence, starting from September 2012 (i.e. at the start of the school year 2012-13) and as a result of the implementation of the new sanctioning measures, any given school could be receiving a different sanction for any of its classes according to the following cases: *a)* real test scores returned (no-cheating), *b)* test scores returned with a correction (suspect of cheating), *c)* no test scores returned at all (high suspect of cheating). 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 suspect cheating or non-return at all of the test scores. In the empirical analysis, we shall use the number of classes receiving a sanction within any given school to define our treatment variable, while the control group is made up of schools which did not receive any sanction in all classes.

A comparison of the two programs. To get a rough idea of how the different policy measures used by Invalsi are expected to fight cheating behavior, it is useful to compare their different implementation mechanisms. The external monitoring program mainly works through a change in the monitoring technology that strictly follows the Invalsi protocol for the administration and marking of the tests. The program, by increasing the probability of detecting illicit or opportunistic behavior, is expected to remove opportunities for cheating behavior of both students and teachers, both during, as well as after the test (see Table 1).

[11] Invalsi calculates the median of the CPI in the 5 macro-areas of the country (North-West, North-East, Centre, South, Islands) and determines the threshold as the median CPI of the most 'virtuous' macro-area.
[12] 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 suspect of cheating. The threshold of 0.5 was chosen arbitrarily as the mean value of the CPI_{cgj}.
[13] If in a given school *s*, more than 50% of the classes satisfy condition (*c*) (i.e. $CPI_{cgj} > 0.5$), 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 5 and 1 percent of the schools, depending on the grade) that a formal evaluation exercise using our econometric strategy is not feasible. The results do not change if we exclude these schools from the analysis.

The 'correction/non-return' sanctioning program instead is supposed to work by changing the incentives to cheat through a 'fame and shame' mechanism, without requiring any additional direct cost or extra resources from the State (Falzetti, 2013). Schools that present anomalies in the patterns of answers within classes, and thus are suspect of cheating behavior, are sanctioned with a correction of the results or by withholding the class test score performance from the evaluation program. While the lack of a proper accountability system, with explicit pecuniary sanctions or loss of resources linked to cheating behavior, is likely to reduce the effectiveness of the 'correction/non-return' sanctioning program, reputational concerns for teachers and schools of being stigmatized as cheaters may be expected to be relevant. Moreover, when parents have free choice between public and private schools - or within public school themselves -, even a moderate level of school competition can generate losses by reducing the attractiveness of the school for prospective students. Clearly the size and significance of the expected losses and the real bite of the deterrence measures is likely to depend on both the institutional environment and the level of school competition.

4. Data and descriptive statistics

4.1 The SNV micro-data

We exploit 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 'correction/non-return' sanctioning program. While the external monitoring program started in 2009-10, in this paper we focus on the 2011-12 wave, which corresponds to the last school year when only the monitoring program was implemented. However, similar results are found when the analysis is extended to previous schools years. Conversely, the sanctioning 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, thus it did not have any influence on the 2011-12 SNV wave that we exploit for identification.[14]

The SNV archives contain individual level records on students' test scores, basic demographic characteristics, while additional information can be drawn from the Student Questionnaire administered after the test. We focus attention primarily on the CPI indicator for language proficiency in primary schools (grade 2 and 5), junior-high (grade 6) and high schools (grade 10), though the results for math are not qualitatively different. The CPI is computed by Invalsi from test scores statistics using 'fuzzy clustering' techniques, it is continuous and bounded between 0 (no

[14] From the universe of classes and schools in the SNV 2012-13 we exclude those subject to the external monitoring, 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 classes, we can safely exclude them from the analysis. Starting from the SNV 2012-13, the test sheets were given to the students in five different versions, obtained changing the order of the items (see Report SNV 2012-13 pag. 11). This was intended to decrease cheating on the part of the students. Given that we do not exploit the panel dimension of the SNV data, and that the change was the same for all grades and subjects, our results are not affected by this additional institutional change.

cheating) and 1 (maximum suspect of cheating), and it can be interpreted as the probability that cheating occurred in each classroom during the test (Invalsi, 2010).[15] Notice that since the technique for the computation of the CPI does not take into account any class-level observable characteristics, monitored and non-monitored classrooms are treated exactly the same way. This aspect is fundamental to implement the identification strategy illustrated in the next section.

We also compute from the SNV data an alternative measure of opportunistic behavior, endorsed by the teachers or by the school principal, based on selective absenteeism on the day of the test, which we use to proxy strategic pooling behavior. The SPI indicator is computed on the share of students who are formally enrolled in the class in September but do not take the Invalsi test in May.[16] While it would have been useful to investigate whether lower-performing students (e.g. non-natives or grade-retained) are over represented in the group of students absent in the day of the test, since these students do not sit the test their socio-demographic characteristics are not available in the SNV data (in some grades, missing data account for more than 70% of the observations).

Students' well-being is an important feature of the assessment exercise that has been largely overlooked in the evaluation of the (unintended) consequences of testing systems (Figlio and Winicki, 2005). In this regard, we also evaluate the effects of the two policies with respect to some behavioral responses of the students, such as their anxiety, stress and tension, both before and during the test. To this purpose, we draw additional information from the Student Questionnaire of the SNV data and compute the share of students who declared to be worried before taking the test, or alternatively reported to be nervous or afraid of doing bad while sitting the test.[17]

4.2 Descriptive statistics

Descriptive statistics concerning the external monitoring and the 'correction/non-return' sanctioning program are reported in Table 2. Overall, about 7.5 percent of the classes were subject to the external monitor (treated classes), while about 15 percent of the schools had at least one class with external monitoring (treated schools). Concerning the 'correction/non-return' sanctioning program, on average, about 31 percent of the classes were subject to either the correction or the non-return policy: the correction treatment is far more common (26 percent) than the non-return treatment (5 percent), while there are not sizeable differences across grades.

[Table 2 about here]

Table 3 shows the descriptive statistics of the dependent and control variables used in the empirical analysis, distinguishing between the treated and the control units. The CPI is two percentage points lower in the classes that received the external inspector (panel A), while no differences can be observed in the SPI. Concerning the measures of students' psychological well-being, on average, a

[15] For further details on the 'fuzzy clustering' techniques used to compute the CPI indicator see Castellano et al. (2009). In a robustness check, we will also use alternative indicators of cheating based on Pereda Fernández (2016).

[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, should be quite negligible in our data.

[17] Each student has to state whether he completely agree, agree, not agree or completely disagree with the statements 'I was already worried before taking the test', 'I was nervous while sitting the test', 'I had the impression of doing bad while sitting the test'. We focus on the share of those who completely agree with each statement. Experimentations with alternative definitions do not qualitatively change the results.

non-negligible share of students (between 16 and 20 percent) declared to be worried before sitting the test or afraid of doing bad during the test, thus showing anxiety both before and during the test. A lower share of students (about 5 percent) declared to be so nervous during the test to find it difficult to figure out the correct answers. In the regression analysis we include as control variables some observable characteristics at the class and school level, such as the share of females, grade-retained students, and non-natives, the class size (and its square) and the school size (and its square), which do not show remarkable differences depending on the treatment status.

[Table 3 about here]

Finally, it is worth recalling that the effect of the two programs on cheating works *via* two different policy parameters. The external monitoring program is based on a randomization: for this reason, the share of treated classes or schools does not vary significantly across the country. Conversely, the 'correction/non-return' sanctioning program makes leverage on the reputational costs of being stigmatized as cheaters and the potential lower attractiveness for prospective students. In this context the policy is implemented as a punishment for high levels of cheating detected, so that we find that its implementation was highly heterogeneous across the country. The average number of classes treated by the 'correction/non-return' policy is higher by almost 15 percentage points in southern regions as compared with the northern ones: we will come back on this aspect in a later section.

5. Empirical strategy

We exploit two different identification strategies to assess the effects (and the effectiveness) of the deterrence and sanctioning programs on different types of illicit and opportunistic behaviors in the SNV Evaluation Program. The identification strategies are similar in spirit and both based on an instrumental variable approach to deal with the potential sources of endogeneity that need to be taken into account in the policy evaluation exercise.

5.1 The external monitoring program

We specify and estimate the following equation:

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

where the outcome variable (y_{csg}) could be the CPI, the SPI or the measures of students' psychological well-being calculated from the SNV 2011-12 for every class *c* in school *s* and grade *g*; EI_{csg} indicates whether the class is treated (i.e. monitored by an external inspector); X'_{csg} is a vector of class level characteristics (share of females, share of grade retained students, share of immigrant students, class size and its square, school size and its square), and φ_g represents various

fixed effects (FE), such as: grade FE, high school type FE and province FE.[18] Both outcomes and control variables are defined at the class-level, the level at which the policy is implemented.

If the presence of the external inspector in a given class were random (see Section 3), the causal effect of the external monitoring policy on the outcomes described above (α_1) can be estimated by OLS. Since, as discussed in the previous sections, it cannot be excluded that the school principal retains some degree of discretion in allocating the external inspector to a different class (plausibly, a better class) from the one selected by the Invalsi randomization process, some additional care should be used in the estimation.[19] In this context, if the assignment is not independent from the quality of the students in the class (i.e. there is positive selection into treatment), we may expect OLS to underestimate the true effect of the external monitoring program. Nevertheless, while the school principal may be able to manipulate the choice of the treated classes within a school, it is not possible to manipulate the allocation of the external inspector in the selected school the day of the test. We thus adopt the estimation strategy proposed by Angrist et al. (2015), and estimate equation 1 in a 2SLS framework, in which a valid instrumental variable for EI_{csg} is given by a dummy variable equal to 1 if a school is treated, and zero otherwise ($\overline{EI}_{sg}$).

The first-stage regression takes the following form:

$$EI_{csg} = \alpha_0 + \alpha_1 \overline{EI}_{sg} + \alpha_2 X'_{csg} + \varphi_g + \varepsilon_{csg} \qquad (2)$$

In the following tables we show that the results for the first stage regressions confirm the statistical relevance of the instrumental variable, while the validity of the instrument rests on the reasonable assumption that the school principal cannot manipulate the choice of the school, but only the choice of the class.[20]

5.2 The sanctioning program

In order to evaluate the effectiveness of the 'correction/non-return' sanctioning program, we start from a baseline specification:

$$y_{csg} = \beta_0 + \beta_1 T_{sg} + \beta_2 X'_{csg} + \varphi_g + \varepsilon_{csg} \qquad (3)$$

[18] Italian provinces (about 110, NUTS level 3) broadly correspond to the School Districts. 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 (Angrist et al. 2015). Experimentations with alternative FE do not alter the results.

[19] Angrist et al. (2015) show evidence in line with the possibility that the school principal can manipulate the random choice of the class made by Invalsi, thus retaining some degree of freedom in sending the inspector to a preferred class (within a selected school).

[20] Notice that Angrist et al. (2015) apply this strategy at the institution level, rather than at the school level. While our results do not change substantially moving from an instrument calculated at the school to one calculated at the institution level (results are available upon request to the authors), we keep the school level because it ensures a higher identification power in the first stage, especially concerning the evaluation of the sanction program. In principle, the Invalsi randomization process specifies a class, and the address of school and the institution where the class is. While it could be possible to manipulate the choice of the class, it would be hard to change the address, and this is equally true for the school or the institution level.

where both the outcomes, y_{csg}, and control variables, X'_{csg}, are defined as in equation 1 above, but due to the different timing and implementation of the policy, they are calculated from the SNV wave 2012-13 (i.e. the first wave after the implementation of the sanction policy); T_{sg} indicates the share of classes in each school *s* that received any treatment as defined in the previous section (the two types of treatments can be the correction or the non-return of the test scores as defined in the previous section); while φ_g represents the same set of fixed effects previously specified. The estimation of equation (3) by OLS entails a main threat to the identification of the causal effect since the definition of the treatment variables (T^p_{sg}) is itself a function of the level of cheating observed in the previous wave of the SNV (SNV 2011-12). In other words, serial correlation may bias our results since those schools that are more likely to feature as 'high suspect of cheating' are also more likely to be subject to the sanctioning process (which is based on the level of cheating observed in the previous period). Hence, a positive effect between the sanctioning treatment and the cheating indicator (measured in the following school year) may partly depend from the spurious serial correlation.

To overcome this problem we exploit an instrumental variable approach based on the presence of the external inspector in the school in the SNV 2011-12 ($\overline{EI}_{sg}$, as specified for equation 2). In details, the sanctioning program was applied to the entire population of the schools, thus neglecting the fact that some of them had received, randomly, the external inspector treatment (Falzetti, 2013). The presence of the external inspector in the school in the SNV 2011-12 randomly lowered the CPI on which the correction and non-return sanctions were determined. Thus, some schools received a lower share of class scores corrected or non-returned because of the presence of the external inspector. Given that the presence of the external inspector in the school was based on the randomized process described above for the external monitoring program, the variable $\overline{EI}_{sg}$ can be used as an instrument as it introduces a source of exogenous variability in the assignment of the treatments. The statistical relevance of the instrument is confirmed by the first stage regressions that we will show in the following tables, while in the robustness section we will also show indirect evidence in support of its validity. In this respect, our identification strategy is likely to identify a local average treatment effect of the sanctioning program (LATE), the causal effect being mainly estimated exploiting the part of the distribution of the schools that effectively changed their cheating behavior because of the presence of the external inspector (and, as a consequence, they were assigned to a different treatment). This is likely to have occurred for schools suspected of high cheating levels.

6. Results

To examine the impact of different deterrence and sanctioning programs on cheating behavior, in all grades of primary and secondary school, we begin by estimating the causal effect of the external monitoring treatment on both our cheating (CPI) and strategic pooling (SPI) indicators (Table 4, panel A). Next, we replicate the same analysis, on both the indicators listed above, first pooling the different sanctioning treatments (Table 4, panel B) and then separately splitting the correction and non-return treatments (Table 5). In the reminder of the section, we also investigate the heterogeneous impact of the above treatments according different dimensions of the schooling

system: first, by school levels (compulsory and high school levels) (Table 6); second, by social capital endowment and geographical areas (Table 7); and third, by school density (number of schools at the local level) (Table 8). A number of robustness checks are also presented and discussed in the following section.

6.1 Baseline results

The baseline results are reported in Table 4. In panel A, we show the impact of the external monitoring program, while in panel B we report the effect of the correction/non-return sanctioning program. The effect of the external monitoring program estimated by OLS shows that the random presence of the inspector in a class decreases the CPI and the SPI by, respectively, 3 and 1 percentage points (p.p.). As previously discussed, to address the possibility of endogenous selection of the treated class on the part of the school principal we implement a 2SLS strategy using as instrument the random assignment of the inspector at the school level. The first stage regression shows that any class in a school where an inspector was present on the day of the test faces a probability of being supervised of about one third (0.34). In the case of the sanctioning program, the presence of the inspector in the previous SNV wave (i.e. in the SNV 2011-12) decreases by 5 p.p. the share of classes of the school that received either a correction or a non-return treatment in September 2011. In both cases, the F-statistics always reject the null that the instrument is weakly correlated to the treatment (Stock and Yogo 2005).

[Table 4 about here]

The 2SLS results for the external monitoring program confirm previous OLS results, though the size of the coefficient (in absolute terms) is larger. The external inspector decreases the propensity to cheat by 5 p.p. – approximately one quarter (0.26) of a standard deviation in the CPI indicator [OR: which corresponds to a reduction of about 80% with respect to the mean CPI/ or of about 70% with respect to the mean CPI of the untreated group] –, showing the effectiveness of the strict monitoring technology associated to the presence of the external inspector in reducing the propensity to cheat in treated classes. Notice that the smaller effect we detect in OLS estimates, compared to 2SLS estimates, confirm the presence of attenuation bias due to some manipulation in the allocation of classes. These results are somehow in line with Borcan et al. (2016) who find that a stricter monitoring technology based on Closed Circuit TV (CCTV) cameras in national evaluations in Romania reduced the probability of passing the test by 8.3 p.p. (about 12% with respect to the mean passing rate).

We also find evidence that classes treated with the external inspector exhibit an increase in strategic pooling. In practice we find that the share of students absent in the day of the test (SPI) increases by 2 p.p. in classes subject to the external inspector, corresponding to an increase of about 16% with respect to the mean SPI. Namely, knowing that an inspector will be present the day of the test, we find evidence that schools (i.e. school principal and teachers) are able to strategically manipulate the pool of students taking the test. While in principle this kind of strategic behavior should be minimized by the protocol enforced by Invalsi that notifies only one week in advance to the school the presence of the external inspector, this short span of time seems enough to allow different types of opportunistic behaviors to occur (Angrist et al. 2015).

Before considering the effect of the overall correction/non-return sanctioning program, it useful to recall that, due to the different implementation mechanisms and the set of incentives, the latter is expected to influence the propensity to cheat in a class or prevent strategic pooling behavior only indirectly through the reputational concerns of the school principal or the teachers. In other words, we may expect a milder effect of this policy when contrasted with the external monitoring program. We report the main estimates in Table 4, panel B. OLS results show a positive, statistically significant, correlation between the treatment and the indicator of cheating. This is likely to reflect a spurious serial correlation: namely, classes in schools treated for cheating in the previous wave, quite independently from the treatment received, are more likely to show up as suspect of cheating. When we re-estimate the model by 2SLS, the effect of the sanctioning program on the CPI indicator vanishes becoming non-statistically significant. Similar results are found with respect to the SPI indicator, as the effect of the sanctioning program becomes non-statistically significant when also fixed effects are introduced.

[Table 5 about here]

This evidence suggests that the sanctioning program did not have any effect in reducing cheating or in influencing strategic pooling on the part of the school. However, it might be that only harsher sanctions, such as those involving the higher reputational cost of non-return of the test scores, can have an effect on cheating and strategic pooling, while milder sanctions (i.e. those based the mere correction of test score) in general do not. In Table 5 , we test the different effectiveness of the alternative sanctioning policy, by separately estimating the impact of test score correction and non-return policies. In practice, we estimate the effect of the share of classes in the school that received the correction treatment as opposed to the share of classes in the school that were treated with a non-return policy. The results in Table 5 do not differ significantly from the previous baseline estimates (Table 4 panel B), showing that the sanctioning policy, quite independently of the design of the incentives, did not produce any response on the part of the school.

6.2 Heterogeneous effects

School grades. To account for differences in the school organization and the length of interpersonal relationship between teachers and students illustrated in the previous sections, we analyze separately primary schools, on the one hand, and high school grades (both junior-high and high) on the other. Results are reported in Table 6.

[Table 6 about here]

We find a larger impact of the external monitoring treatment on cheating in primary, as compared to high schools grades (panel A.1). Conversely, opportunistic behaviors in terms of strategic pooling are detected only in the higher grades (panel A.2). These results can be explained considering some of the institutional differences mentioned. First, primary school teachers are more likely to have developed strong personal ties with the students, which may induce them to adopt a more benevolent attitude when supervising the student during the test or in the marking phase thus increasing manipulation in the absence of the external inspector (Lucifora and Tonello, 2015). Second, strategic pooling behavior appears easier to be perpetrated when students are older, since in this case they need not adults to take care of them when they are away from school. When considering the 'correction/non-return' sanctioning policy, we find no evidence of any

heterogeneous effects according to the school grades (panels B.1 and B.2). This points again to an overall ineffectiveness of the policy in the entire schooling system.

Regional differences and social capital. 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. 2015, Battistin et al. 2015, 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 these 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.

[Table 7 about here]

To assess the existence of a regional divide in the effectiveness of the deterrence and sanctioning policies, in Table 7 we perform the analysis separately by regional clusters and according to the stock of social capital available in each local area. In this respect, notice that while the monitoring program is applied in a uniform way across all regions, the correction/non-return sanctioning policy is likely to work indirectly through a set of reputational concerns which are grounded in the cultural attitudes of the local area and are shared by students and teachers. Also we should expect a harsher punishment where cheating is more pervasive. As a result, 68 percent of the schools located in Southern regions received either the correction or the non-return treatment (for at least one class), as compared to 35 percent of the schools in the North of the country (17 percent if we restrict the sample to the non-return treatment, as compared to 5 percent in the North).

Consistent with the above discussion, we find that the monitoring program reduces cheating more in the South of the country and in areas with lower endowments of social capital, where the latter indicates a low value of trust (i.e. below the median) (Guiso et al. 2004).[21] Also strategic pooling is found to be larger in areas with low social capital. On the contrary, no differential effects are detected in the effectiveness of the correction/non-return sanctioning program in reducing cheating across either Northern or Southern regions.

School density interaction effects. The mechanism through which the sanctioning program is expected to impact on cheating behavior is rooted in the reputational concerns (the so-called 'fame and shame' mechanism) faced by schools that are found suspect of cheating behavior, and thus receive either the correction or the non-return treatment. The sanctioning scheme could also serve the purpose of an 'early warning' to the schools to induce them to enforce better internal monitoring and proctoring of the testing process. However, the results we obtained so far point to an overall ineffectiveness of the sanctioning policy, for both the correction and the non-return measures. A number of reasons may be advanced to explain why this is the case. The first is the absence of real sanctions were attached to schools' suspect of cheating or engaging in strategic pooling. The second is the lack of transparency, that is no obligation for the schools to make public their scores. Third is the low-stake of the testing and the absence of any established school accountability system, as no resources are attached to schools' performance in testing. All these aspects are likely to have

[21] Experimentations with alternative measures of social capital do not qualitatively change the results.

seriously hindered the reputational costs of the sanctions and consequently the effectiveness of the whole scheme. There could be however contexts, in which even a low-stake and low-enforcement sanctioning scheme, such as the one described above, could involve a substantial penalty for the school. For example, the reputational costs of being stigmatized as suspect cheaters may be magnified by the presence of competition among schools.

In the Italian context, schools generally are not given many incentives to compete. However, where the density of schools in a local area is high, parents enjoy more freedom in the choice of the school for their children and schools themselves often compete in attracting prospective students. For example, schools advertise their extra-curricular activities, their facilities, the school-time scheduling that they can offer to the families, and their results in the Invalsi tests in the web, but also in 'Open-days' and participating to local events and exhibitions (see Figure B.1).

To test for this hypothesis, we construct a measure of school density given by the normalized ratio of the number of schools (by school track and high school type) in a Local Labor Market (LLM) with respect to its size (in terms of squared km).[22] The school density measure takes values from 0 (lowest school density) to 1 (highest school density), showing higher density in highly urbanized areas and in the LLMs of the biggest cities.

[Table 8 about here]

The results reported in Table 8 show the effects of the correction or non-return treatment on both cheating and strategic pooling behaviors in an interacted model with school density. To ease the interpretation and to improve the precision of the estimates, we construct a dummy equal to 1 for the LLM with a school density greater than the median (in any given school track and high school type), and 0 otherwise. The effect of the interacted term on the CPI cheating indicator bears a negative sign, though it is never statistically significant at conventional levels. This result reinforces the lack of effectiveness of the sanctioning policies on cheating, even where reputational concerns are supposed to be higher. Conversely, we do find a positive and statistically significant coefficient on the interaction term in the strategic pooling equation, particularly in Southern regions, in areas with a low social capital endowment, and in high school grades.[23] This evidence is somewhat surprising, as it seem to suggest that the sanctioning policies triggered more strategic pooling in context where school density is higher and especially in areas where the correction/non-return sanctions were more extensively implemented. Speculating on this finding, one could argue that the sanctioning policies by correcting or withholding the information about the schools test scores contributed to raise the stake of the testing program in context where competition for prospective students is higher, such as in high school grades. Moreover, in context where schools are numerous and social norms encompass higher incidence of cheating or lower trust, the implementation of sanctions instead of discouraging cheating generated additional opportunistic behavior raising schools' strategic pooling.

6.3 Robustness

[22] Similar indicators have been also used in the literature to assess the degree of competition among schools in a given area (Hanushek and Rivkin 2003).
[23] Effects for primary schools (not reported in the table) are not statistically significant.

The external monitoring program has been exploited in other works that show that the randomization is effective, and that standard balancing tests between the treatment and the control group are verified (Brunello et al. 2013, Angrist at al. 2015, Lucifora and Tonello 2015, Pereda Fernàndez 2015). As discussed in previous sections, in our empirical analysis we follow the strategy outlined in Angrist et al. 2015, which makes use of the most updated versions of the SNV archives. We use as instrumental variable the presence of the inspector in the school, while Angrist et al. 2015 use the presence of the inspector at the comprehensive school level (more schools belongs to a same institution). The results of our analysis are confirmed (both for the CPI and the SPI) also when we adopt this latter specification.[24]

The analysis of the sanctioning program is a novelty in the body of the literature focusing on opportunistic behaviors in education. The identification strategy hinges on the assumption that the instrument (i.e. the presence of the external inspector in the school in the SNV 2011-12) does not influence directly cheating and strategic pooling the SNV 2012-13. In other words, the validity of the instrument rests on the absence of any correlation between the presence of the external inspector in the previous school-year and cheating behavior in the following year. Supportive evidence showing the lack of any statistically significant correlation is reported in Table 9, where we show the results of OLS regressions between the instrumental variable ($\overline{EI}_{sg}$, the presence of the inspector in the school in the SNV 2011-12) and the different outcomes (y_{csg}) measured in the SNV 2012-13.

[Tables 9 and10 about here]

Since our treatment variable is defined at the school level, while we conduct the analysis at the class level to maintain comparability between the estimated effects of the two policies and allow more flexibility in the heterogeneity analysis, we also replicate the analysis using a school-level aggregation. Our main results are shown to be robust to the different levels of aggregation (class vs. school) (Table 10, columns 1 and 2).[25]

In the main analysis we use the cheating indicator exploited by the Invalsi to detect and sanction cheating, which is based on Castellano et al. (2009). A possible concern with statistical indicators of cheating relates to the false positives, that is identifying as high cheaters classes that are exceptionally good, and thus show very high test scores and low within class variability. The new statistical indicator of cheating developed by Pereda Fernàndez (2015) and applied to Invalsi data, improves on this aspect showing a lower risk of incurring in false positives.[26] Our results do not change replicating the empirical analysis using this new indicator of cheating as outcome variable (Table 10, columns 3 and 4).

[24] Results are not reported here but are available upon request with the authors.

[25] We also try alternative specifications in which the treatment variables are defined as dummies according to whether or not the school experienced the correction, the non-return, or both policies. Results (not reported) do not show statistically significant effects, though less precisely estimated.

[26] We gratefully acknowledge Santiago Pereda Fernàndez for having shared with us his indicator. Note that, by construction, the indicator developed in Pereda Fernàndez (2015) is only available for the classes without the external inspector, thus it cannot be used in the comparative analysis of the two policies.

7. Behavioral externalities on students' stress and psychological well-being

The introduction of high-stake assessments programs can have important side effects on students' well-being. Anecdotal evidence suggest that, particularly in high stake testing programs, both parents and teachers often exert significant pressure on students before the test, thus increasing their level of stress. Cases of more subtle methods to increase students' performance have also been reported. For example, Figlio and Winicki (2005) report evidence showing that during the testing periods, in order to boost students performances, some schools in the US increased the caloric intake provided by the cafeterias. In another paper, Bockhari and Schneider (2011) show that diagnosis of attention deficit (AD) is more frequent in the US states that adopted stronger accountability laws, suspecting that some schools try to boost students' performances in the tests using psycho-stimulant drugs. While these are, hopefully, extreme cases, it seems important to evaluate whether policies aimed at reducing opportunistic behaviors have had any effect on students psychological well-being. In the case of the external monitoring program, we may anticipate that the presence of an external inspector could cause stress or anxiety on students while sitting the test or before the test if teachers put pressure on the students, knowing that the inspector will be there. In the case of the sanctioning program, additional pressure on students could come from the teachers or the school principals, whom after receiving a sanction in the previous SNV (correction or non-return treatment), now want to improve the performance of the class, or of the whole school. This may induce anxiety both before and during the test.

[Table 11 about here]

Specifically, our dependent variables are the share of students who declared to have been already *worried* before taking the test, *nervous* while sitting the test, and *afraid* of doing bad while sitting the test. In this exercise, we apply the same estimation strategies as in the previous analysis.[27] Results are reported in Table 11. We find evidence of increased anxiety and stress for the external monitoring program, but exclusively in high schools. The share of students declaring to be worried before the test increases by 2 p.p., those afraid of doing bad increase by 3 p.p., while those nervous during the test by 1 p.p. in classes subject to the external monitoring treatment. These effects are non-negligible and correspond to increase between 30 and 50% of the share of students showing any form of psychological stress in the high school (with respect to the mean). Students in lower grades (primary and junior-high schools) do not seem to be affected.[28] In the case of the sanctioning program the effects are mixed and do not show any evidence in favor of the hypothesis that teachers put pressure on students and this may hurt students' psychological well-being.

8. Conclusions and policy implications

Recent works have shown that cheating in test scores, both on the part of students and teachers, can influence students' education careers, raising their grades in future classes, high school graduation

[27] For this part of the analysis, the Primary schools only include grade 5, as the Student Questionnaire is not taken by students in grade 2.

[28] This results is in line to what found by Brunello et al. (2013) and Lucifora and Tonello (2015) on primary school students in the first wave of implementation of the external monitoring program in the SNV.

rates and even their wages (Dee et al. 2016, Diamond and Persson 2016). The design of effective testing systems that reduce the possibility of opportunistic behavior appears to be fundamental to reduce such behaviors and prevent their long run costs (Mechtenberg 2009, Schwager 2012).

In this study we evaluate the effectiveness of different measures introduced in Italy to monitor and sanction schools suspect of manipulating students test scores. Opportunistic and illicit behavior in the SNV testing program is measured for all Italian schools – from grade 2 of primary school to grade 10 of secondary school – using a statistical indicator of class-based cheating and by an indicator of strategic pooling. Within the national evaluation program, we test the effectiveness of two alternative policies: the external monitoring program, based on the presence of an external inspector for the administration and marking of the tests; and the sanctioning program, consisting in a correction or non-return of the test scores for classes suspect of cheating. In the empirical analysis, we used a randomized experiment to estimate the causal effect of the monitoring and sanctioning policies on school cheating behaviors. We also examined whether such policies trigger schools' strategic pooling behavior, or affect students psychological well-being increasing their stress and anxiety before or during the test.

We find that higher monitoring is effective in deterring cheating at all grades, the presence of the external examiner in treated classes reduces the incidence of cheating by 5 p.p. as compared to classes in the control group, a reduction of about 80% with respect to the mean CPI. The estimated impact of monitoring on cheating is shown to be heterogeneous: increasing with the length of the relationship between teachers and students, and being more pervasive where social norms value trust less and where social capital is low. We also find evidence that stricter monitoring increases schools attempts to manipulate strategically the pool of students taking the test: the share of students absent in the day of the test is about 16% higher in treated classes. Conversely, we find generally no effect for the sanctioning policy, whether correction of the class test scores or even the harsher non-return measure, on either cheating or strategic pooling. We find that sanctions, in specific circumstances, that is when the reputational costs are potentially magnified by the presence of competition among schools in attracting prospective students and in contexts displaying lower civicness, might also trigger strategic pooling behavior.

We interpret the different results we obtain, contrasting the incentive structure implied by the two policies. We argue that the external monitoring program, by raising the cost of opportunistic behavior - through an increase in the probability of detecting cheaters in treated classes -, is effective in removing all opportunities for cheating. Conversely, the sanctioning program only works indirectly, by increasing the reputational costs of being stigmatized as cheaters, *via* a 'fame and shame' mechanism. In this case, we explain the overall ineffectiveness of the sanctioning mechanisms considering the low-stake of the testing and the lack of a proper school accountability system, linking school performance and cheating behavior to a system of rewards and sanctions.

These results provide a number useful insights concerning the effectiveness of policies directed to fight cheating behaviors in schools. For instance, we contribute to the discussion on whether policy interventions based on monitoring are more or less effective as compared to policy tools that make leverage on incentives and sanctions, and highlight key challenges for the design of proper accountability systems.

First, as in the case of Borcan et al. (2016), monitoring technology is found to be highly effective in removing cheating behavior. However, the amount of resources necessary to implement it even in a sub-sample of schools is substantial: back of the envelope calculations suggest that the total budget devoted every year to the external monitoring program in the Italian SNV Evaluation Program is in the order of 1,500,000€. This amount corresponds to approximately 20 percent of the total budget devoted to the Evaluation Program and to 40 percent of the resources Invalsi receives every year from the Central Government.[29] 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.

Second, the effectiveness of the deterrence and sanctioning policies crucially depend on whether they are embedded in a proper school accountability system. In the Italian context, the lack of any obligation for the schools suspect of cheating to make public their scores or sanctions is found to reduce the deterrence potential of such policies. Moreover, while the high-stake of the testing is expected to increase the incentives of students and teachers to engage in opportunistic behavior, the Italian experience show that even in a low-stake environment cheating can be pervasive. We also find that cheating is higher where trust and social capital are low, suggesting a role of cultural factors and social norms in opportunistic behaviors. The design of accountability systems should then trade the high- low-stake of the testing system with the prevalent cultural and institutional setting.

Third, even a moderate level of school competition could be effective in raising the effectiveness of sanctioning policies through the increase in the costs of being stigmatized as suspect cheaters and the associated loss of reputation in attracting prospective students. This, however, is likely to work better where schools density is high and parents have more freedom in the choice of the school. In this vein, we believe that the evidence resulting from this study shows that when sanctioning policies are not embedded in a proper school accountability system or when school competition is lacking, they are unlikely to provide an appropriate set of incentives and a suitable environment to foster students achievements and curb opportunistic behaviors. Moreover, lacking these conditions, we do find evidence that in areas where trust and civicness endowments are poor, and corruption and criminal behavior socially accepted, higher reputational costs may also trigger schools' strategic behaviors, such as selective pooling.

[29] This figure is obtained multiplying the 200€ fee each external inspector receives to complete the supervision and proctoring tasks for each single class, by the number of treated 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).

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, J., E. Battistin, and D. Vuri (2015). In a Small Moment: Class Size and Moral Hazard in the Mezzogiorno. IZA Discussion Paper No. 8959.

Battistin, E., M. De Nadai, and D. Vuri (2014). Counting Rotten Apples: Student Achievement and Score Manipulation in Italian Elementary Schools. IZA Discussion Paper No. 8405.

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 (2016). Fighting Corruption in Education: What Works and Who Benefits? American Economic Journal: Economic Policy, forthcoming.

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, S. (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.

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

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 (2016). The Causes and Consequences of Test Score Manipulation: Evidence from the New York Regents Examinations. NBER WP No. 22165.

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. (2015). Rules rather than discretion: teacher hiring and rent extraction. European University Institute, Max Weber Program WP 2015/14.

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

Figlio, D. N. and J. Winicki (2005). Food for thought: the effects of school accountability plans on school nutrition. Journal of Public Economics 89(2-3), 381-394.

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

Fryer Jr, R. G., S. D. Levitt, J. A. List, and S. Sadoff (2012). Enhancing the Efficacy of Teacher Incentives through Loss Aversion: A Field Experiment. NBER Working Paper No. 18237.

Gershenson S. (2015). Performance Standards and Employee Effort: Evidence from Teacher Absences. IZA Discussion Paper No. 9203.

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

Guiso, L., P. Sapienza, and L. Zingales (2013). Long-term persistence. EIEF Working Paper 23/13.

Hanushek, E. and M. Raymond (2005). Does School Accountability Lead to Improved Student Performance? NBER Working Paper No. 10591.

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).

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.

Jordan, A. E. (2001), 'College student cheating: The role of motivation, perceived norms, attitudes, and knowledge of institutional policy', Ethics & Behavior, 11(3), 233-247.

Kerkvliet, J.. and C. L. Sigmund (1999), 'Can we control cheating in the classroom?', Journal of Economic Education 30 (Fall): 33 -143.

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.

Koch, A., J. Nafziger, and H. S.Nielsen (2014). Behavioral economics of education. Journal of Economic Behavior and Organization 115, 3-17.

Lavy, V. (2009). Performance Pay and Teachers' Effort, Productivity, and Grading Ethics. American Economic Review 99(5), 1979-2011.

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

Levitt, S. D. (2004). The Economics of Education. NBER Reporter OnLine: Winter 2004.

Levitt, S. D., J. A. List, and S. Sadoff0 (2016). The Effect Of Performance-Based Incentives On Educational Achievement: Evidence From A Randomized Experiment. NBER Working Paper 22107.

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 (2015). Cheating and Incentives: Learning from a Policy Experiment. Interdisciplinary Center for Economic Science, george Mason University, Working Paper, October 2015.

McCabe, D. L., and L. K. Trevino (1997). Individual and contextual influences on academic dishonesty: a multi campus investigation. Research in Higher Education 38(3), 379-376.

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

McCabe, D. L., and L. K. Trevino (1993). Academic Dishonesty: Honour Codes and Other Contextual Influences. Journal of Higher Education 64(5), 522-38.

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).

Pereda-Fernández, S. (2016). A new method for the correction of test scores manipulation. Working Paper No. 1047, Bank of Italy, Economic Research and International Relations Area.

Reinikka, R. and J. Svensson (2011). The power of information in public services: Evidence from education in Uganda. Journal of Public Economics 95(7-8),956-966.

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.

Schultz, G., M. R. West and L. Wößmann (2007). School Accountability, Autonomy, Choice, and the Equity of Student Achievement: International Evidence from PISA 2003. OECD Education Working Paper No. 14.

UK Standard & Testing Agency (2013). 2012 Maladministration report. National Curriculum assessments. https://www.gov.uk/government/publications/2012-maladministration-report.

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.

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 a peer (Carrel et al. 2008, Martinelli et al. 2015, 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 2015,Bertoni et al. 2013, Angrist et al. 2015, Lucifora and Tonello 2015)	(i) Manipulating students test scores (Jacob and Levitt 2003, Dee et al. 2016, Diamond and Persson 2016) (ii) Shirking in correction procedures (Angrist et al. 2015)

Table 2. Descriptive statistics: share of treated classes in the monitoring program and in the sanctioning program, by grade.

	Mean	Sd	N
Panel A: share of treated classes in the external monitoring program			
Primary school	6.24	24.19	54359
Junior-high school	7.97	27.08	24089
High school	10.30	30.4	21728
Total	7.54	26.4	100176
Panel B: share of treated classes in the sanctioning program			
Primary school			
Correction or Non-return	32.34	38.01	47906
Correction only	26.96	34.75	47906
Junior-high school			
Correction or Non-return	30.09	29.6	20357
Correction only	24.25	25.83	20357
High school			
Correction or Non-return	29.71	30.46	14530
Correction only	24.88	26.75	14530
Total			
Correction or Non-return	31.33	34.86	82793
Correction only	25.93	31.46	82793

Notes. A treated class is a class where the external inspector is present (Panel A); the share of treated classes in the sanctioning program indicates the share of classes (within each school) which received each treatment (Panel B). The *Correction or Non-return* and the *Correction only* treatments are defined as the share of classes in each school that received, respectively, both the correction or the non-return treatment or only the correction treatment in September 2012 (i.e. based on the results of the SNV 2011-12). The share of classes that received the *Non-return only* treatment can be obtained as the difference between the other two treatments. **Source:** Invalsi SNV 2011-12, 2012-13.

Table 3. Descriptive statistics: dependent and control variables, by treatment status.

	All			*Treated*			*Controls*		
	Mean	Sd	N	Mean	Sd	N	Mean	Sd	N
Panel A: external monitoring program									
Dependent variables									
Indicators of cheating:									
Cheating Propensity (CPI)	0.06	0.19	100176	0.05	0.16	22944	0.07	0.2	77232
Strategic Pooling (SPI)	0.12	0.11	100176	0.13	0.13	22944	0.12	0.11	77232
Students stress:									
Worried before the test	0.21	0.16	72720	0.05	0.06	19045	0.06	0.07	53675
Nervous while sitting the test	0.05	0.07	72720	0.18	0.15	19045	0.23	0.16	53675
Afraid of doing bad while sitting the test	0.18	0.13	72720	0.15	0.12	19045	0.19	0.13	53675
Control variables									
Inspector in the class	0.08	0.26	100176	0.33	0.47	22944	0	0	77232
Inspector in the school (Z)	0.23	0.42	100176	1	0	22944	0	0	77232
Female (share)	0.49	0.18	100176	0.49	0.2	22944	0.49	0.17	77232
Non-native (share)	0.11	0.13	100176	0.11	0.13	22944	0.11	0.13	77232
Grade retained (share)	0.08	0.13	100176	0.1	0.15	22944	0.07	0.12	77232
ESCS (adjusted)	0.05	0.47	100176	0.04	0.49	22944	0.05	0.47	77232
Class size (no. of students)	20.62	4.73	100176	22.09	4.08	22944	20.19	4.82	77232
School size (no. of classes)	6	2.88	100176	7.35	3.15	22944	5.6	2.67	77232
Panel B: sanctioning program									
Dependent variables									
Indicators of cheating:									
Cheating Propensity (CPI)	0.04	0.11	82793	0.04	0.1	15325	0.04	0.11	67468
Strategic Pooling (SPI)	0.12	0.12	82793	0.13	0.13	15325	0.12	0.11	67468
Students stress:									
Worried before the test	0.2	0.15	58849	0.17	0.15	12239	0.21	0.16	46610
Nervous while sitting the test	0.05	0.06	58849	0.04	0.06	12239	0.05	0.06	46610
Afraid of doing bad while sitting the test	0.16	0.12	58849	0.14	0.11	12239	0.16	0.12	46610
Control variables									
Correction or Non-return	0.31	0.35	82793	0.26	0.29	15325	0.32	0.36	67468
Correction only	0.26	0.31	82793	0.23	0.26	15325	0.27	0.33	67468
Non-return only	0.05	0.17	82793	0.04	0.11	15325	0.06	0.18	67468
Inspector in the school in the previous wave (Z)	0.19	0.39	82793	1	0	15325	0	0	67468
Female (share)	0.49	0.17	82793	0.49	0.19	15325	0.49	0.16	67468
Non-native (share)	0.1	0.13	82793	0.1	0.13	15325	0.1	0.13	67468
Grade retained (share)	0.07	0.11	82793	0.08	0.13	15325	0.06	0.11	67468
ESCS (adjusted)	-0.02	0.46	82793	-0.01	0.48	15325	-0.02	0.46	67468
Class size (no. of students)	20.68	4.18	82793	21.69	3.85	15325	20.45	4.22	67468
School size (no. of classes)	5.79	2.68	82793	6.86	2.88	15325	5.55	2.57	67468

Notes. The SES is an indicator of the Socio-Economic Status of the students families: it is standardized with 0 mean in the entire sample. We adjust the indicator placing a zero for all students in grade 2 for whom the indicator is not calculated. The variables on Students' stress are obtained as the share of those who completely agree with each statement: 'I was already worried before taking the test', 'I was nervous while sitting the test', 'I had the impression of doing bad while sitting the test'. **Source:** Invalsi SNV 2011-12, 2012-13.

Table 4. Monitoring vs. incentives: baseline results.

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
	OLS				*First stage*	*2SLS*			
	CPI		SPI		Treatment	CPI		SPI	
Panel A: external monitoring program									
Inspector in the class	-0.04***	-0.03***	-0.01***	-0.01***		-0.05***	-0.05***	0.02***	0.02***
	(0.00)	(0.00)	(0.00)	(0.00)		(0.01)	(0.01)	(0.01)	(0.00)
Z (Inspector in the school)					0.34***				
					(0.00)				
First stage F-stat.						1018.45	8801.75	1018.45	8801.75
N	100176	100176	100176	100176	100176	100176	100176	100176	100176
Panel B: sanctioning program									
Correction or non-return treatment	0.052***	0.040***	0.002	0.002		0.031	0.016	-0.175***	-0.041
	(0.002)	(0.002)	(0.001)	(0.001)		(0.026)	(0.025)	(0.036)	(0.029)
Z (Inspector in the school in the previous wave)					-0.051***				
					(0.005)				
First stage F-stat.						93.76	97.81	93.76	97.81
N	82793	82793	82793	82793	82793	82793	82793	82793	82793
Control variables	yes	yes	yes	yes	yes	yes	yes	yes	yes
Fixed effects		yes		yes	yes		yes		yes

Notes. CPI indicates the Cheating Propensity Indicator, SPI the Strategic Pooling Indicator. Class level regressions weighted by the number of students in the class. The set of control variables includes class and 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 3). The set of fixed effects includes: fixed effects for the Italian provinces (110 provinces), grade fixed effects (grades 2, 5, 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). In the First stage regression the dependent variable (Treatment) is the dummy variable indicating the presence of the inspector in the class. The First stage F-statistics refers to the Kleibergen-Paap rk Wald F-statistics. Robust standard errors in parenthesis, clustered at the school 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. The sanctioning program: heterogeneous effects by incentive type.

	(1)	(2)	(3)	(4)	(5)
	OLS		*First stage*	*2SLS*	
	CPI	SPI	Treatment	CPI	SPI
Panel B.1					
Correction treatment	0.020***	0.006***		-0.005	-0.059
	(0.002)	(0.002)		(0.030)	(0.041)
Z (Inspector in the school in the previous wave)			-0.040***		
			(0.005)		
First stage F-stat.				63.31	63.31
N	71205	71205	71205	71205	71205
Panel B.2					
Non-return treatment	0.126***	0.000		0.039	-0.109
	(0.007)	(0.003)		(0.061)	(0.076)
Z (Inspector in the school in the previous wave)			-0.027***		
			(0.003)		
First stage F-stat.				71.49	71.49
N	39001	39001	39001	39001	39001

Notes. CPI indicates the Cheating Propensity Indicator, SPI the Strategic Pooling Indicator. Class level regressions weighted by the number of students in the class. All the regressions include class and school characteristics and fixed effects as specified in Table 4. In the First stage regression the dependent variable (Treatment) is the dummy variable indicating the presence of the inspector in the class. The First stage F-statistics refers to the Kleibergen-Paap rk Wald F-statistics. Robust standard errors in parenthesis, clustered at the school 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. Heterogeneous effects by school grade.

	(1)	(2)
	Primary school	Junior-high and high school
Panel A: external monitoring program		
Panel A.1: CPI		
Inspector in the class	-0.06***	-0.04***
	(0.01)	(0.01)
First stage F-stat.	5344.55	6727.42
N	54359	45817
Panel A.2: SPI		
Inspector in the class	-0.01***	0.04***
	(0.00)	(0.01)
First stage F-stat.	5214.88	6727.42
N	54359	45817
Panel B: sanctioning program		
Panel B.1: CPI		
Correction or non-return treatment	0.032	0.007
	(0.028)	(0.048)
First stage F-stat.	102.46	23.98
N	47906	34887
Panel B.2: SPI		
Correction or non-return treatment	0.013	-0.095
	(0.018)	(0.070)
First stage F-stat.	102.46	23.98
N	47906	34887

Notes. Class level 2SLS regressions weighted by the number of students in the class. CPI indicates the Cheating Propensity Indicator, SPI the Strategic Pooling Indicator. All the regressions include class and school characteristics and fixed effects as specified in Table 4. The First stage F-statistics refers to the Kleibergen-Paap rk Wald F-statistics. Robust standard errors in parenthesis, clustered at the school 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. Heterogeneous effects by North vs. South divide and social capital.

	(1)	(2)	(3)	(4)
	North	South	High Social Capital	Low Social Capital
Panel A: external monitoring program				
Panel A.1: CPI				
Inspector in the class	-0.02***	-0.09***	-0.02***	-0.07***
	(0.01)	(0.01)	(0.01)	(0.01)
First stage F-stat.	4503.93	6252.84	7536.16	3423.44
N	60720	39456	50560	49616
Panel A.2: SPI				
Inspector in the class	0.02***	0.02***	0.01**	0.03***
	(0.00)	(0.01)	(0.00)	(0.01)
First stage F-stat.	4432.62	6260.87	7543.48	3458.28
N	60720	39456	50561	49615
Panel B: sanctioning program				
Panel B.1: CPI				
Correction or non-return treatment	0.031	0.014	-0.001	0.031
	(0.030)	(0.037)	(0.036)	(0.033)
First stage F-stat.	47.58	58.99	33.89	67.55
N	51258	31535	42334	40459
Panel B.2: SPI				
Correction or non-return treatment	-0.054	-0.005	-0.052	-0.021
	(0.044)	(0.035)	(0.046)	(0.036)
First stage F-stat.	47.58	58.99	33.89	67.55
N	51258	31535	42334	40459

Notes. Class level 2SLS regressions weighted by the number of students in the class. CPI indicates the Cheating Propensity Indicator, SPI the Strategic Pooling Indicator. All the regressions include class and school characteristics and fixed effects as specified in Table 4. The High and Low Social capital subsamples are defined according to the school being located in a province above or below the median value of the variable 'trust' (Guiso et al. 2004). The First stage F-statistics refers to the Kleibergen-Paap rk Wald F-statistics. Robust standard errors in parenthesis, clustered at the school 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. Heterogeneous effects by school density.

	(1)	(2)	(3)	(4)	(5)	(6)
	All sample		North		South	
	CPI	SPI	CPI	SPI	CPI	SPI
Panel A						
Correction or non-return treatment	0.046	-0.106***	0.005	-0.090	0.049	-0.090**
	(0.032)	(0.039)	(0.045)	(0.064)	(0.043)	(0.043)
Correction or non-return treatment * High density	-0.057	0.119**	0.048	0.071	-0.082	0.175***
	(0.046)	(0.054)	(0.064)	(0.094)	(0.059)	(0.058)
High density	0.018	-0.021	-0.012	-0.006	0.036	-0.053**
	(0.014)	(0.017)	(0.016)	(0.023)	(0.024)	(0.024)
First stage F-stat.	44.87	44.87	13.17	13.17	28.99	28.99
N	82793	82793	51258	51258	31535	31535
	Junior-high and high school		High Social Capital		Low Social Capital	
	CPI	SPI	CPI	SPI	CPI	SPI
Panel B						
Correction or non-return treatment	0.022	-0.282**	0.005	-0.125	0.055	-0.077*
	(0.057)	(0.120)	(0.055)	(0.081)	(0.038)	(0.040)
Correction or non-return treatment * High density	-0.030	0.384**	-0.013	0.136	-0.049	0.105*
	(0.093)	(0.162)	(0.084)	(0.113)	(0.050)	(0.056)
High density	0.012	-0.093*	0.004	-0.022	0.019	-0.023
	(0.028)	(0.048)	(0.019)	(0.026)	(0.020)	(0.022)
First stage F-stat.	9.65	9.65	7.09	7.09	29.73	29.73
N	34887	34887	42334	42334	40459	40459

Notes. Class level 2SLS regressions weighted by the number of students in the class. CPI indicates the Cheating Propensity Indicator, SPI the Strategic Pooling Indicator. All the regressions include class and school characteristics and fixed effects as specified in Table 4. High and low social capital subsamples are defined as in Table 7. The dummy High density takes value one for the schools located in a LLM where the school density is higher than the median for each track, and zero otherwise. The school density is constructed as the normalized ratio between the number of schools (by school track) in a given LLM and the LLM dimension (in squared km). The First stage F-statistics refers to the Kleibergen-Paap rk Wald F-statistics. Robust standard errors in parenthesis, clustered at the school 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 9. Robustness checks: a test for instrument validity.

	(1)	(2)	(3)	(4)	(5)
	All sample	Primary	Junior-high and high school	North	South
Panel A: CPI					
Z (Inspector in the school in the previous wave)	-0.001	-0.002	-0.000	-0.001	-0.001
	(0.001)	(0.002)	(0.002)	(0.001)	(0.003)
N	82793	47906	34887	51258	31535
Panel B: SPI					
Z (Inspector in the school in the previous wave)	0.002	-0.001	0.003	0.002	0.000
	(0.001)	(0.001)	(0.002)	(0.002)	(0.002)
N	82793	47906	34887	51258	31535

Notes. Class level OLS regressions weighted by the number of students in the class. CPI indicates the Cheating Propensity Indicator, SPI the Strategic Pooling Indicator. All the regressions include class and school characteristics and fixed effects as specified in Table 4. Robust standard errors in parenthesis, clustered at the institution level in columns 1-5 and 8-9, at the school level in columns 6-7. Asterisks denote statistical significance at the * p<0.1, ** p<0.05, *** p<0.01 levels. **Source:** Invalsi SNV 2011-12, 2012-13.

Table 10. Robustness checks: alternative specifications and alternative cheating indicators.

	(1)	(2)	(3)	(4)
	SPI		CPI	
Panel A				
Correction or non-return treatment	-0.012	0.036	0.007	0.007
	(0.032)	(0.025)	(0.010)	(0.009)
First stage F-stat.	132.87	134.68	97.73	134.78
N.Observations	34601	35113	82763	35110
Panel B				
Correction treatment	-0.024	0.006	0.013	0.011
	(0.046)	(0.030)	(0.015)	(0.013)
First stage F-stat.	75.07	77.26	63.37	77.35
N.Observations	31235	31718	71181	31716
Panel C				
Non-return treatment	-0.025	0.037	0.034	0.032
	(0.094)	(0.068)	(0.029)	(0.025)
First stage F-stat.	84.78	78.91	71.31	78.90
N.Observations	19912	20260	38987	20257
Specifications:				
School level	yes	yes		yes
Class level			yes	
Invalsi CPI		yes		
Alternative CPI (Pereda Fernàndez 2015)			yes	yes

Notes. Class and school level 2SLS regressions weighted by the number of students in the class. CPI indicates the Cheating Propensity Indicator, SPI the Strategic Pooling Indicator, the Alternative CPI is obtained from Pereda Fernàndez (2015). Regressions in columns 1-2 and 4 are at the school level; regression in column 3 is at the class level. All the regressions include class and school characteristics and fixed effects as specified in Table 4. The First stage F-statistics refers to the Kleibergen-Paap rk Wald F-statistics. Robust standard errors in parenthesis, clustered at the institution level in columns 1-2 and 4, at the school level in column 3. Asterisks denote statistical significance at the * p<0.1, ** p<0.05, *** p<0.01 levels. **Source:** Invalsi SNV 2011-12, 2012-13 and Pereda Fernàndez (2015).

Table 11. Behavioral effects: the effects of monitoring and incentives on students stress.

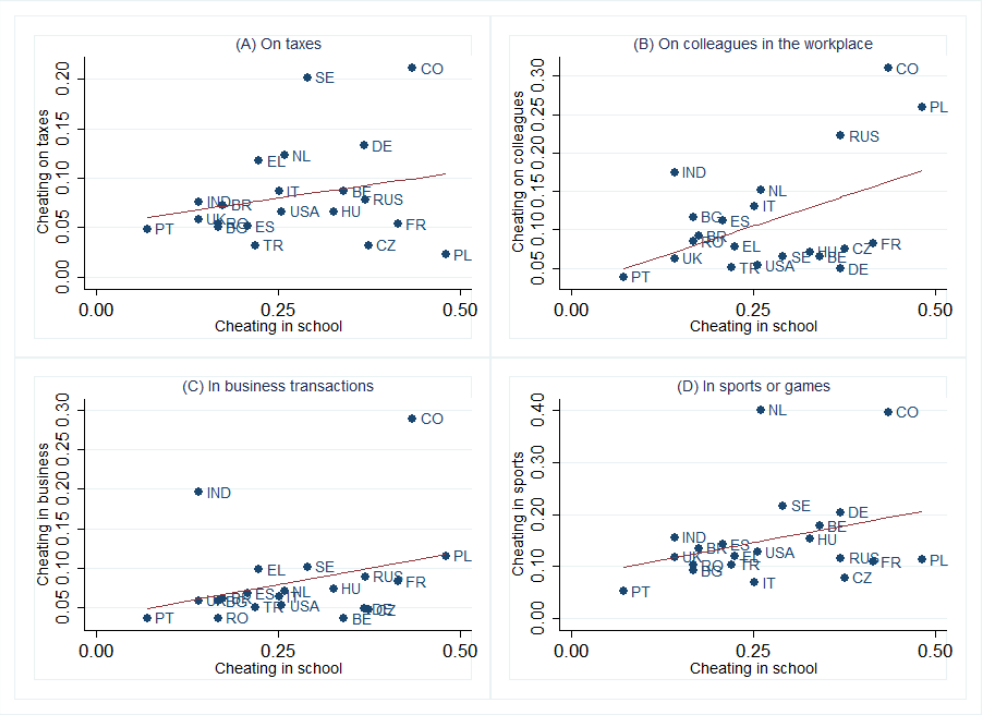
	(1)	(2)	(3)	(4)
	All grades	Primary school	Junior-high school	High school
			Worried before the test	
Panel A.1: External monitoring program				
Inspector in the class	0.00	-0.00	0.00	0.02***
	(0.00)	(0.01)	(0.01)	(0.01)
First stage F-stat.	8479.75	4717.38	2646.76	5541.02
N	72720	27061	23944	21715
Panel B.1: Sanctioning program				
Correction or non-return treatment	0.029	0.049	-0.047	0.040
	(0.035)	(0.050)	(0.076)	(0.049)
First stage F-stat.	50.41	36.92	12.17	14.46
N	58849	24006	20317	14526
			Afraid of doing bad during the test	
Panel A.2: External monitoring program				
Inspector in the class	0.00	0.01	-0.01**	0.03***
	(0.00)	(0.01)	(0.01)	(0.01)
First stage F-stat.	8479.75	4717.38	2646.76	5541.02
N	72720	27061	23944	21715
Panel B.2: Sanctioning program				
Correction or non-return treatment	0.020	-0.015	-0.009	0.082*
	(0.028)	(0.039)	(0.059)	(0.047)
First stage F-stat.	50.41	36.92	12.17	14.46
N	58849	24006	20317	14526
			Nervous during the test	
Panel A.3: External monitoring program				
Inspector in the class	-0.00	-0.00	-0.00	0.01*
	(0.00)	(0.00)	(0.00)	(0.00)
First stage F-stat.	8479.75	4717.38	2646.76	5541.02
N	72720	27061	23944	21715
Panel B.3: Sanctioning program				
Correction or non-return treatment	-0.030*	-0.055**	-0.019	-0.004
	(0.016)	(0.024)	(0.035)	(0.022)
First stage F-stat.	50.41	36.92	12.17	14.46
N	58849	24006	20317	14526

Notes. Class level 2SLS regressions weighted by the number of students in the class. In the Student Questionnaire each student has to state whether he completely agree, agree, not agree or completely disagree with the statements 'I was already worried before taking the test', 'I was nervous while sitting the test', 'I had the impression of doing bad while sitting the test'. The dependent variables are obtained as the share of those who completely agree with each statement. Primary school excludes students in grade 2 who do not take the Questionnaire. **Source:** Invalsi SNV 2011-12, 2012-13.

Appendix A. The prevalence of cheating: Some 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).

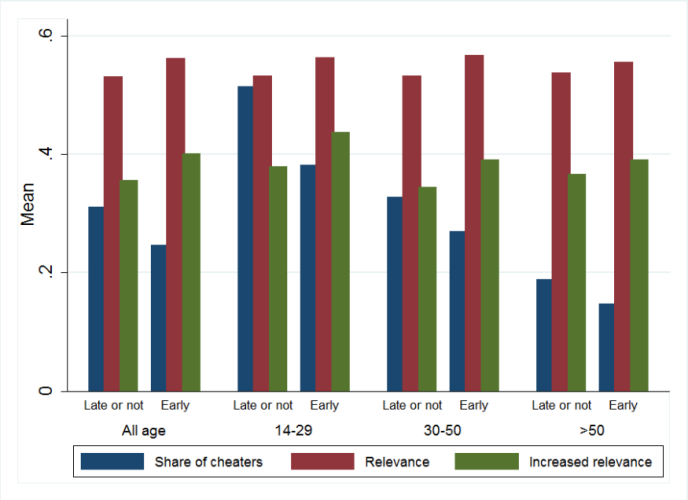
A similar survey conducted by The Wall Street Journal (2008) on national representative sample of individuals across different countries provides additional details on individuals' perceptions about the diffusion of cheating practices.[30] Cheating practices at school or university – as declared by the 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

[30] 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 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.

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 transactions, in work practices, as well as in sports and games. Moreover, in Figure A.2, we show that countries, in which national assessment programs were adopted later or not adopted at all, have a higher share of students who declare to have cheated in school (ISCED levels 1 or 2) and a lower share reporting that cheating at school is a problem (Eurydice, 2009).

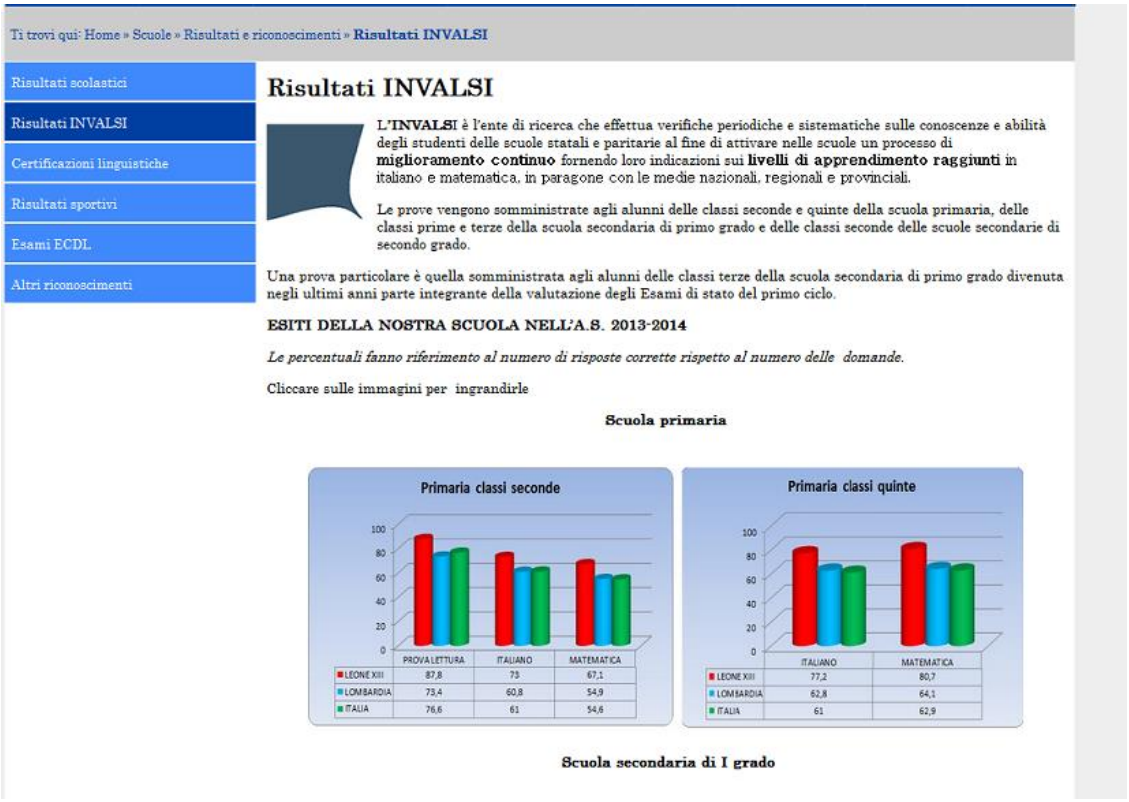
Figure A.2. The social relevance of cheating at school or university, by age group and countries which adopted national assessments programs: European countries.



Notes. We define as 'Early adopters' (as opposed to 'Late or not adopters') those countries that adopted forms of national assessments at the ISCED levels 1 or 2 before the year 2000, as reported in Eurydice (2009). The 'Share of cheaters' is defined as in Figure A.1; 'Relevance' is defined as the share of individuals who answered 'Yes' to the question 'In my country, is cheating at school or university a major problem?'; 'Increase relevance' is defined as the share of individuals who answered 'Yes' to the question 'Is cheating at school or university, generally speaking, more or less common now than it was 10 years ago?'. **Source:** based on Survey on Deceit, The Wall Street Journal (2008) and Eurydice (2009).

Appendix B. Additional Figures and Tables

Appendix Figure B.1. 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.