Students' Cheating as a Social Interaction: Evidence from a Randomized Experiment in a National Evaluation Program

Claudio Lucifora
Marco Tonello

Students' Cheating as a Social Interaction: Evidence from a Randomized Experiment in a National Evaluation Program

Claudio Lucifora
Università Cattolica and IZA

Marco Tonello
Bank of Italy and Defap

ABSTRACT

Students' Cheating as a Social Interaction: Evidence from a Randomized Experiment in a National Evaluation Program[*]

We analyze students' cheating behavior during a national evaluation test. We model the mechanisms that trigger cheating interactions between students and show that, when monitoring is not sufficiently accurate, a social multiplier may magnify the effects on students' achievements. We exploit a randomized experiment, which envisaged the presence of an external inspector in the administration and marking of the tests, to estimate a structural (endogenous) social multiplier in students' cheating. The empirical strategy exploits the Excess-Variance approach (Graham, 2008). We find a strong amplifying role played by social interactions within classrooms: students' cheating behaviors more than double the class average test scores results. The effects are found to be larger when students are more homogeneous in terms of parental background characteristics and social ties.

JEL Classification: C31, D62, I21

Keywords: social multiplier, students' cheating, randomized experiment

Corresponding author:

Claudio Lucifora
Department of Economics
Università Cattolica
1, L.go Gemelli
20123 Milano
Italy
E-mail: claudio.lucifora@unicatt.it

"It's seen as helping your friend out. If you ask people, they'd say it's not cheating. I have your back, you have mine." senior student at Stuyvesant High School in Manhattan.

"We want to be famous and successful, we think our colleagues are cutting corners, we'll be damned if we'll lose out to them, and some day, when we've made it, we'll be role models. But until then, give us a pass." student at Harvard Graduate School of Education.

The New York Times, September 25th, 2012

1. Introduction

In many social and economic contexts individuals often face the choice to adopt different types of opportunistic or even illicit behavior to increase their welfare taking advantage of others for personal interests. Leaving aside major crimes, there is abundant evidence indicating that cheating on taxes, free riding on public goods, claiming benefits without entitlement, bribing and corrupting public officials, abusing of drug and drinking, smoking when not permitted, as well as other types of dishonest behaviors are widely diffused phenomena in most countries (Kleven et al. 2011; Fortin et al. 2007; Powell et al. 2010; Gaviria and Raphael 2001; Clark and Loheac, 2007).

In this paper, we focus attention on a specific type of such fraudulent behavior, that is students' cheating when taking an exam. Several surveys document that students' cheating has grown, over the last decades, hand in hand with the more extensive use of testing programs (Davies et al., 2009; McCabe 2005; Rimer 2003)[1], yet there is little evidence on the effects of cheating behavior for educational outcomes, as well as on the measures taken to contrast its diffusion[2]. Students' cheating behavior can have important consequences in the process of human capital accumulation and for the functioning of the labor market. For example, cheating can interfere with the evaluators ability to assess students' performance and can decrease the external validity of grades (Anderman and Murdock, 2007). 'Cheating bias' may contaminate the information used in many educational

---

[1] Large-scale cheating has been uncovered over the last year at some of the US most competitive schools, like *Stuyvesant High School* in Manhattan, the *Air Force Academy* and, most recently, *Harvard University* (The New York Times, September 7, 2012). A survey conducted as part of the Academic Integrity Assessment Project by the Center for Academic Integrity (Duke University) and covering 80,000 students and 12,000 faculties in the U.S. and Canada, between 2002 and 2005, reported that 21% of undergraduates admitted to have cheated on exams at least once a year (McCabe, 2005). Another survey run - in 2010, on a national sample of U.S. public and private high schools students - by the Josephson Institute of Ethics – *'Report on honesty and integrity'* (2011) - found that 59.3% of the U.S. students interviewed cheated at least once during a test, while more than 80% of them copied from others' homework at least once.

[2] In many countries, policy interventions make extensive use of test scores to determine the allocation of resources across schools and to evaluate teachers' work but little has been done to develop objective measures of students' cheating.

decisions, such as: promoting students from one grade to the next, or awarding a diploma without the required knowledge. In one case, cheating detracts from the signaling validity of education titles on the labor market; in the other case, it determines negative externalities on the learning processes, for example, slowing down the teaching activity[3].

Moreover, students' cheating raises a number of concerns not just for the unfairness with respect to students who do not cheat, but more generally for the externalities that are created on others (McCabe, Treviño and Butterfield, 1999; Carrel et al. 2008; Dee and Jacob, 2012). In particular, when a student breaks an ethical code of behavior exchanging information, cooperating with other students or using any prohibited materials during an exam (Cizek, 2003), many others – who might otherwise have behaved honestly - end up being influenced thus reacting to such behavior. Many students may feel that they cannot afford to be disadvantaged by those who cheat without being reported or punished by school authorities[4]. In this context, even an isolate cheating behavior may propagate and become larger through social interactions. Hence, as widely discussed in the social interaction literature, the aggregate outcome is likely to depend on a direct effect (a reaction *via* private incentives to cheat) and an indirect effect on behavior (a reaction to the cheating behavior of others): the ratio between the equilibrium aggregate response and the sum of the reactions of individuals to cheating is the so-called social multiplier (Glaeser et al., 2003). The cheating outcome is amplified by the multiplier generating large differences in variance across different groups (i.e. school, classroom, etc.) with otherwise similar characteristics. While unobserved heterogeneity and sorting of individuals across groups may account for part of the differences in cheating behavior, social interactions within group of students linked by different types of contextual ties are often necessary to explain the excess variation that is observed in the data (Manski, 1993; Brock and Durlauf, 2001).

Note that in many circumstances the driving force for dishonest or illicit student's behavior during an exam may be found in some personal traits, such as: greed, envy, competitive pressure, etc.; however, social norms, low trust, a widespread acceptance of illicit behavior and other background characteristics may also increase the likelihood of dishonesty within students.

In other words, cheating behavior can be seen a genuine free-riding problem, where students, for any given level of effort, try to maximize their performance (i.e. pass-rate probability, exam grades,

---

[3] The consequences of cheating can be even more severe in educational settings in which the school system is based on a strict tracking system (e.g. Germany).

[4] Note that reporting the offenders, as contemplated in many schools' ethical codes, is required to halt the diffusion of cheating behaviors, nevertheless it should be noted that small transgressions and dishonest behavior are very often overlooked or tolerated within many schools, either because students do not like to be directly involved in the accusation, or because schools themselves do not want to be associated to the judiciary procedures required to support the allegations of student's dishonesty.

test scores, etc.) and exploit the possibility of opportunistic behavior – i.e. exchanging information or cooperating - anytime the monitoring system tolerates it or is not efficient in reporting the offenders[5]. The interdependencies between students' decisions to cheat are at the basis of the positive covariance in individual behavior that triggers the (social) multiplier effect. In terms of the framework introduced by Manski (1993), and extensively discussed in the literature on social interactions, the above cheating behavior represents the endogenous part of social effects (Bramoullé et al. 2009; De Giorgi et al. 2010)[6].

The literature on social interactions in education has largely focused on peer effects in students achievements in classrooms and schools, or on social outcomes within fraternity (or sorority) membership (Sacerdote 2001; Zimmerman 2003; Stinebrickner and Stinebrickner 2006; Foster 2006; Graham 2008; Hanushek, et al. 2003; Lyle, 2007; Lefgren 2004; Carrel et al. 2009; Lavy et al. 2012). Conversely, the effect of students' cheating interactions has not received much attention and even less is known about the potential mechanisms that may drive cheating behavior.

An extensive literature in educational psychology has documented cheating behavior in schools[7], while only few papers have addressed the issue of social interactions in cheating behavior using a credible identification strategy. Most papers in the literature use statistical techniques that cannot reliably separate the endogenous and exogenous effects – i.e. the effect of the group upon an individual from the effect of an individual upon the group due to the well-known reflection problem (Carrel et al. 2008). Starting from an early study by Stanard and Bowers (1970), where it was shown that cheating tended to be higher among members of a fraternity or sorority, the psychological literature has focused attention on how social norms, peer pressure, environmental pressure and self-perception of cheating behavior affect individual cheating decision. McCabe and Trevino (1997), for example, found peer-related contextual factors to be the strongest predictors of

---

[5] Monitoring activities are introduced to validate testing procedures in national evaluation programs. However, 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 -, national assessments programs are conducted on a census basis and the same school teachers supervise students while tacking the exam (U.S. Department of Education, 2009; Eurydice, 2009).

[6] Manski (1993) identifies three main factors that are likely to influence social interactions: exogenous (or contextual) effects (i.e. when the propensity of an individual to behave in some way varies with the exogenous characteristics of the group), correlated effects (i.e. common shared group-level factors) and endogenous social interactions (i.e. when the propensity of an individual to behave in some way varies with the behavior of the group). Only the latter effect can determine the social multiplier.

[7] Stephens and Gehlbach (2007) count more than an hundred empirical studies on this issue over the last decade. Research in this area documents that cheating occurs among students from all grades, from elementary schools to colleges, and even in graduate schools. From a developmental perspective, Miller et al. (2007) find that cheating tend to occur less in younger children than in adolescents. These developmental differences are due to changes both in students' cognitive abilities and in the social structure of the educational contexts in which children and adolescents interact (Murdock et al., 2001). From a motivational perspective, Anderman and Murdock (2006) document different reasons for engaging in academic cheating: some students cheat because they are highly focused on extrinsic outcomes such as grades; others cheat because they are concerned with maintaining a certain image to themselves or to their peers or because they lack the requisite self-efficacy to engage in complex tasks.

cheating in their multi-campus investigation of individual and contextual influences related to academic dishonesty. Students who perceived that their peers disapproved academic dishonesty were less likely to cheat, while those who perceived higher levels of cheating among their peers were more likely to report cheating. Grimes and Rezek (2005) estimate a probit regression model to determine the factors that contribute to the probability of cheating. Their results indicate that the most important determinants are personal beliefs about the ethics and social acceptability of cheating and various attributes of the classroom environment[8]. Carrel et al. (2008) are the first to analyze cheating behavior as a social interaction using separate estimation procedures to identify an exogenous (contextual or pre-treatment) peer effect and an endogenous (during treatment) peer effect. Their model assumes that peer effects are completely driven either through experiences of cheating behavior at high school or completely through peers' behavior while at college. Their results for the endogenous peer effects indicate that one additional college cheater 'creates' approximately 0.61–0.75 additional college cheaters.

There is also a parallel literature that has focused on other forms of cheating, for example cheating on taxes is one of the most interesting cases. Kleven et al. (2011) analyze a tax enforcement field experiment in Denmark confronting different types of tax reporting methods (i.e. third-party reporting vs. self-reported income), as well as different auditing methods faced by tax filers. The authors show that tax cheating is close to zero for income subject to third-party reporting, but substantial for self-reported income and that prior audits and threat-of-audit letters have significant effects in reducing cheating on self-reported income. Galbiati and Zanella (2012) estimate a social multiplier effect in tax cheating generated by the congestion of the auditing resources. They use a rich dataset from Italian Local Tax Authorities and find that an exogenous shock altering concealed income independently across individuals produces an equilibrium variation that is up to three times the initial response.

We develop a simple theoretical model to highlight the mechanisms that may drive social interactions in cheating behavior and to derive testable predictions. We show that students may optimally decide whether to engage in cooperative effort exchanging information (e.g. conform to other student cheating behavior) and do so taking into account other students' best response. The equilibrium solution takes the form of a linear-in-means model with (endogenous) social interactions *à-la* Manski, so that we can attach a structural interpretation to our estimate of the multiplier (Cooley 2010a,b). In particular, our model posits a specific social spillover: by observing

---

[8] Kerkvliet and Sigmund (1999) explore the determinants of source-specific cheating behavior including student characteristics and deterrent measures. They conclude that large alcohol consumption and low grade point average increase the probability of cheating. Jordan (2001) finds a significant correlation between college students' perceived social norms and their self-reported cheating.

or expecting that student achievements depend also on cheating interactions, students adjust their behavior in response to the cheating behavior in the classroom.

We use a unique data set drawn from the 'National Survey of Students' Attainments' (henceforth SNV) (in both Mathematics and Language), which is compulsory for all schools and students attending different grades of primary and junior-high school in Italy, and exploit a randomized experiment which envisaged the presence of an external inspector in the administration and marking of the tests.

In particular, we contrast the behavior of students in classrooms where the test is administered only by the school teachers, with the behavior of students in classroom where an external inspector invigilates over students' behavior during the exam, to identify students' social externalities in cheating behavior. In the non-monitored classrooms (i.e. our control group), we may expect monitoring to be more 'benevolent' *vis-à-vis* student interactions during the exam, while no interactions are expected to occur in the monitored classrooms (i.e. our treated group).

In this context, we interpret the presence of a positive covariance in students' behavior, when exchanging information or engaging in any sort of collaborative behavior during the test in the non-monitored classroom, as a form a behavioral externality which may produce a social multiplier[9]. Students' cheating behavior during the test has few relevant implications. First, it generates excess variance in individual behavior with respect to individual and group characteristics in the monitored classrooms. Second, it introduces a difference among the between-group and the within-group variance of individual behavior. These two features are the foundations of the empirical strategy proposed by Graham (2008), which exploits the Excess-Variance (henceforth, E-V) approach to separate the part of variability due to individual and group level heterogeneity from the excess variability genuinely originating from social interactions.

We contribute to different strands of literature. First, to the literature on the identification of grade inflation due to various types of cheating behaviors (Dee and Jacob, 2012, for plagiarism; Carrel et al., 2008, for students' cheating; Jacob and Levitt, 2003, Jacob 2005 for teachers' cheating)[10]. Focusing on students' cheating behavior, our approach departs from Carrel *et al.* (2008) since we do not identify the effects of a given 'share of cheaters' on individual test score, rather we provide a measure of endogenous interactions due to students' cheating behavior. In this sense, we contribute

---

[9] Note that students' cheating during an exam is an interesting case study of social interactions in the classroom, since it is likely to capture the same network of friendships and cooperative behaviors that take place during the school year. Students are more likely to collaborate with closer friends, with classmates they share out-of-the school activities (like sport practice), as well as with classmates sitting closer.

[10] Evidence of cheating behavior mostly refers to academia (Mc Cabe and Trevino, 1999; Mc Cabe, 2005; Carrel et al. 2008), less from other type of schools. In the Italian context, Ferrer-Esteban (2012) and Bertoni et al. (2012) use SNV dataset to study the effects of supervision on students' performance.

to the part of the literature on social interactions which tries to overcome the 'reflection problem' and directly estimate the effects of the endogenous social multiplier (among others: De Giorgi et al. 2010; Calvò-Armengol et al., 2009; Bramoullé et al. 2009)[11]. Second, we use data on test scores and other individual characteristics drawn from the whole student population at different grades in a national evaluation test, which is a significant improvement from studies which rely on representative samples. We also match our data with other administrative archives, at the school level, and with a follow-up survey to get additional information on parental background characteristics as well as motivational questions concerning the test. Third, we implement a rather innovative estimation method based on the Excess-Variance approach to estimate (endogenous) social interactions by exploiting an exclusion restriction provided by a randomized experiment and illustrate the presence of heterogeneity in the estimated social effects[12].

We find a strong amplifying role played by cheating social interaction within students in the classroom: in the baseline estimates we identify a social multiplier ranging between 2.26 and 2.43 for Math, and between 2.05 and 2.18 for Language. This implies that students' cheating behavior more than double the class average test scores results and the effects are found to be larger when students are more homogeneous in terms of parental background characteristics and 'social ties'.

Our findings show that tolerating cheating behavior, as it is often done, can amplify the negative effects on students' performance, alter the signaling role of education in the labor market, and raise collective indulgence with respect to various forms of dishonest practices. Also, given that increasing competition in school achievement and in the job market are likely to exerts considerable pressure on students to perform well in exams, more resources should be devoted to monitoring activities in order to avoid cheating interactions to become widespread.

The paper is organized as follows. In Section 2 we build a theoretical framework to define the social multiplier parameter. Section 3 describes the institutional setting, the data and the randomized experiment, and provides some descriptive statistics. Section 4 discusses the identification strategy

---

[11] Grounding on Manski's seminal works (1993), empirical literature on peer effects has focused on the estimation of reduced form equations which collapse the endogenous and the exogenous effects into one parameter of interest, that is identifiable and defined to as 'social effect' parameter (Ammermüller and Pischke, 2009; Lavy et al. 2012). Recent works in the field of social interactions in education addressed the reflection problem in the estimation of the classical linear-in-means model à la Manski (1993) using data where social groups are endogenously defined (i.e. networks, Calvo-Armengol et al. 2009; Bramoullé et al. 2009), introducing appropriate exclusion restrictions (e.g. partially overlapping groups in De Giorgi et al. 2010; group-size variations in Davezies et al. 2009), or even just plugging into the reduced form equation a lagged value of peers' achievement as proxy of the contemporaneous one (Hanushek et al. 2003, 2009). In these cases, random assignment is usually ensured in the specific characteristics of the data used (e.g. random assignment to classes and courses at the first year of college), or controlled for using multiple levels fixed-effects.

[12] Galbiati and Zanella (2012) also implement the Excess-Variance approach to tax cheating behavior using a more standard exclusion restriction given by group-size variations.

while Section 5 and 6 present the main results and some robustness checks. Section 7 concludes and provides some policy implications.

2. Theoretical framework

We develop a simple model to investigate the mechanisms that may drive social interactions in cheating behavior. We consider the (endogenous) decision students face, when taking an exam, as to whether work individually or, alternatively when the monitoring technology is loose, engage in any kind of prohibited cooperation exchanging information with other students. We assume that students derive utility from achievement, which depends on own (costly) effort and on the effort of classmates[13]. Since own effort and peers' effort in the classroom are complementary inputs in the achievement function, students may decide to cheat choosing the optimal level of cooperative effort to be shared with their peers (Anderman et al., 2007). In this context, cheating originates a behavioral externality among individuals, who simultaneously choose their utility-maximizing level of effort taking into account peers' best response to each level of effort chosen (Brock and Durlauf, 2001). Note that the type of social externality that emerges from student's cheating behavior is different from the traditional peers' achievement externality (i.e. based on predetermined characteristics of the students, such as unobserved ability or 'quality') considered in the literature, since here individual decisions play an important role in shaping students' behavior which, in turn, originates the endogenous effects needed to determine the social multiplier (Sacerdote, 2001; Cooley, 2010a,b; De Giorgi and Pellizzari, 2011; Imberman et al. 2012; Lavy et al. 2012)[14]. In particular, in the present context, peers' achievement *per se* may not affect a student's achievement when cheating or any other form of behavioral interactions are absent[15].

---

[13] Notice that we assume no cost of cheating. This is consistent with the institutional setting (and the empirical application, i.e. SNV surveys) as in practice disciplinary measures or sanctions have never been applied to students and teachers who behave dishonestly.

[14] In terms of behavioral interactions, the literature on drug use, smoke habit or alcohol addiction provides a better illustration of a social mechanism through which group's behavior directly affects individual decision. In these cases, we have endogenous peer effects whenever the 'average behavior' of the reference group directly influences the individual behavior or choice. It is the group's decision to drink, smoke or use drugs that influences the individual decision to take some action, and both group and individual behaviors are directly captured by some quantifiable measures (alcoholic drinks per day, binary decision to smoke/not smoke, or cigarettes per day etc.) (Cooley, 2010b; Gaviria and Raphael, 2001; Sacerdote, 2001).

[15] The empirical literature on peer effects, traditionally, does not distinguish between the effect on test scores deriving from unobservable pre-determined characteristics of the students and their unobservable behavioral choices (Sacerdote, 2001; Imberman et al. 2012; Lavy et al. 2012). However, as noted by Cooley (2010b, p. 7) "*[…]Annual standardized exams are often the outcome of interest, and, in the absence of cheating, are not a group effort. Thus, peer achievement per se may not affect a student's achievement. In contrast, the decision of a teenager to smoke or drink alcohol might be readily affected by having peers that engage in these behaviors*". Examples of endogenous peer behavior on achievement are discussed in Lazear (2001) where peer disruptive behavior imposes negative externalities on other students in the classroom. Similarly, Figlio (2007), Lavy and Schlosser (2011) and Kinsler (2006) present empirical

We model students' achievement (y_i) as dependent on the following elements: x_i and x_j are, respectively, individual and peers' predetermined characteristics (i.e. gender, parental background, non-native status, etc), μ represents shared class-level factors (i.e. school and class environment, teacher's experience), while e_i and e_j are, respectively, unobservable individual and peers' endogenous behaviors where j indicates any other student different from i:

$$y_i = x_i'\pi_x + x_j'\tilde{\pi}_x + e_i + \tilde{\pi}_e e_j + \mu \qquad [1]$$

The above specification describes the Achievement Production Function (APF) suggesting that achievement is increasing in both the student and peers' unobservable behavior, such that individual achievement may improve when cheating externalities are present (i.e. $\tilde{\pi}_e > 0$). The parameter of interest here is $\tilde{\pi}_e$ which identifies the endogenous social interactions characterizing students' behavior in the presence of cheating. The other parameters π_x, and $\tilde{\pi}_x$ describe exogenous (or contextual) effects (Cooley, 2010b)[16]. To get further insights on how student's behave, we specify individual's utility (U_i) as a quadratic function that depends positively on achievement ($\beta_y \geq 0$) and it is concave in own effort ($\beta_e \geq 0$):

$$U_i\left(y_i, e_i, e_j, c\right) = \beta_y y_i - \frac{1}{2}\beta_e e_i^2 + \tilde{\beta}_e e_i e_j - \overline{c}(\mu) \qquad [2]$$

The component $\overline{c}(\mu)$ represents an exogenous cost due to teacher's monitoring activity during the test. All individuals have to bear this cost which is likely to depend on class-level characteristics (strictness in teacher monitoring, class physical dimension, desks allocations, etc.). Notice that peers' behavior matters as long as there are social interactions during the exam (i.e. loose or benign monitoring allows cheating) and students are willing to share their effort cooperating with other students (i.e. conforming to other students' cheating behavior): hence individual utility increases with peers' effort ($\tilde{\beta}_e > 0$). Students maximize utility choosing the level effort as best response to peers' (simultaneous) effort choices and subject to the achievement function (i.e. given by the structural APF):

---

evidence that disruptive peers may negatively affect achievement. In the robustness section we also test whether achievement peer effects play a role in our data.

[16] Note that human capital externalities still operate in the APF (eq. [1]) but, in some sense, they can be thought as being part of the individual and peers' predetermined characteristics and contribute to individual outcome as 'endowment effects' (i.e. exogenous effects incorporated in $\tilde{\pi}_x$). Given that we only want to estimate the endogenous component of the social interactions process due to cheating behavior during the test, and that students' quality is likely to be same in the two sub-populations used for the empirical estimation, we assume that x_i also includes unobserved predetermined individual characteristics.

$$Max_{e_i} \ U_i = \beta_y y_i - \frac{1}{2}\beta_e e_i^2 + \tilde{\beta}_e e_i e_j - \overline{c}(\mu)$$

$$s.t. \quad y_i = x_i'\pi_x + x_j'\tilde{\pi}_x + e_i + \tilde{\pi}_e e_j + \mu$$

[3]

Solving for e_i the first order condition yields the effort best response function:

$$e_i^{BR} = \frac{\beta_y}{\beta_e} + \frac{\tilde{\beta}_e}{\beta_e}e_j$$

[4]

The effort best response is a function of the marginal utility of effort relative to the cost and is increasing in the average effort of peers when cheating interactions occurs (i.e. cooperative peers' effort, $\tilde{\beta}_e > 0$). Given the assumption that achievement is monotonically increasing in cooperative effort behavior (e_i), the effort best response can be mapped into an achievement best response which is observable to the researcher[17]:

$$y_i^{BR} = \delta_0 + x_i'\delta_x + x_j'\tilde{\delta}_x + Jy_j + \mu'\tilde{\delta}_\mu$$

[5]

Given the linear-in-parameters form of the achievement best response, it can be shown that a unique Nash equilibrium exists ($y_i^*; y_j^*$) so that equation [5] can be rewritten as:

$$y_i^* = \delta_0 + x_i'\delta_x + x_j'\tilde{\delta}_x + Jy_j^* + \mu'\tilde{\delta}_\mu$$

[6]

Under the assumption that the achievement observed during any exam or test (i.e. grade, tests scores, etc.) originates from the described utility-maximising behavior - when cheating occurs - we can use peer achievement to proxy for peer cooperative behavior (effort) such that equation [6] expresses individual achievement as function of individual and peers' characteristics as well as peers' achievement. The parameter J corresponds to the 'unobserved endogenous social effects' and it is a measure of the endogenously determined effect of individual behavior on the reference group average behavior:

$$J = \frac{\tilde{\beta}_e + \beta_e\tilde{\pi}_e}{\beta_e + \tilde{\beta}_e\tilde{\pi}_e}$$

[7]

It is composed by three structural parameters: the marginal (dis)utility from own effort exerted in cheating activities (β_e), the marginal utility derived from peers' effort in cooperative cheating behavior ($\tilde{\beta}_e$), and the marginal effect of peers' effort exerted in cheating on individual achievement ($\tilde{\pi}_e$).

---

[17] See Appendix A for the detailed derivation.

The linear-in-means model in equation [6] requires $J < 1$ (i.e. a stability condition to ensure that a small change in cheating behavior will not determine a diverging response in aggregate), and this is true if two restrictions are imposed to the structural parameters: that is $\tilde{\beta}_e < \beta_e$ and $\tilde{\pi}_e < 1$ [18]. The first condition states that the utility from cooperative cheating behavior (i.e. peers' effort) must be smaller than the disutility from own effort; the second condition requires the marginal contribution of peers' effort on individual achievement to be smaller than own contribution (i.e. normalized to 1 in the APF, see equation [1]). Both conditions are rather intuitive and realistically met in our framework. Notice also that, when monitoring allows cheating to occur, the assumption of cooperative peer effort (i.e. $\tilde{\beta}_e > 0$), implies that J is always positive ($J \geq 0$) [19]. In other words, as we show in the descriptive evidence, when the monitoring technology prevents students to interact or cooperate during the test, their achievements (or test-scores) tend to be more dissimilar and exhibit a larger within-class variance as compared to the achievements of the non-monitored students where behavioral interactions are present.

2.1 The social multiplier

The simple model described above implies a social multiplier, such that any shock to individual behavior - *via* social interactions - determines relatively larger aggregate responses. To frame the model in a way suitable for empirical estimation, we need to retrieve an expression for the social multiplier. First, without loss of generality, we can rearrange equation [6] substituting average peers' characteristics and average peers' achievement:

$$y_{ic}^* = \delta_0 + x_i'\delta_x + \bar{x}_c'\tilde{\delta}_x + J\,\bar{y}_c^* + \mu'\tilde{\delta}_\mu \qquad [8]$$

Averaging within the reference group (i.e. the classroom) and solving for $\bar{y}_c$ yields [20]:

$$\bar{y}_c = \gamma\delta_0 + \bar{x}_c'\gamma(\delta_x + \tilde{\delta}_x) + \mu'\gamma\tilde{\delta}_\mu \qquad [9]$$

Where $\gamma = (1-J)^{-1}$ represents the social multiplier in students' cooperative efforts, when during the exam cheating can occur (Glaeser *et al.*, 2003). Substituting equation [9] into [8] we obtain the following reduced form model:

---

[18] $\pi_x > 0$ is without loss of generality, assuming that covariates are constructed accordingly.

[19] It is easy to show that, since we have assumed cooperative peer effects ($\tilde{\beta}_e > 0$), and given that $\beta_e > \tilde{\beta}_e$, this necessarily implies that $\beta_e > 0$ and also $\tilde{\pi}_e > 0$ thus ensuring that $J \geq 0$.

[20] In the social interactions literature the group whose (average) behavior influences the behavior of each individual is considered a "reference group", in our setting the classroom is the natural reference group to be considered in the empirical analysis.

$$y_i^* = \gamma\delta_0 + x_i'\delta_x + \gamma\tilde{\delta}_x\bar{x}_c' + (\gamma-1)\delta_x\bar{x}_c' + \gamma\delta_\mu\mu' \qquad\qquad [10]$$

The achievement best response takes the form of the classical linear-in-means model of social interaction *à la* Manski (1993). While this has been obtained at the cost of introducing some *ad hoc* linear functional forms, it has some clear advantages[21]: first, it highlights the mechanism through which cheating behavior may generate students' social interactions; second, it provides a specification that allows direct estimation of the social multiplier parameter (γ) using the Excess-Variance approach (Graham, 2008).

Note that, in some sense, the interpretation of the social interaction parameter as students' cooperative effort is specific to our model, since cheating is the only social externality we are modeling. However, while we think that cheating externalities are the main driving force in the estimation of our structural social multiplier, we cannot exclude that other social mechanisms may also play a role. We briefly discuss some alternative interpretations hereafter.

One hypothesis, also discussed in the literature (Jacob and Levitt, 2003a,b; Jacob, 2005; Lavy, 2009), is that our social multiplier parameter may originate also from explicit teacher cheating rather than students' cooperative efforts in exchanging information when monitoring is more benevolent or looser. Teacher cheating may take the form of suggesting the right answers to all students, or even altering students' answers sheets during the marking phase. Indeed, besides the ethical implications of such behavior, there are several reasons why teachers may want to alter students' outcomes: for example, they may wish to improve their students' results in the exams, alternatively teachers may dislike sharp differences in results across classes within the same school, or feel pressure because of monetary incentives linked to student performance, or because the allocation of resources to schools depends on students outcomes (Jacob, 2005; Lavy 2009). A second hypothesis is that students in classroom with an external inspector feel intimidated and are negatively affected in their performance during the test (Bertoni et al. 2012). Finally social effects may also derive from the presence of some ethical norms of behavior whose strength decreases with the extent of cheating itself (Algan et al. 2011; Myles and Naylor, 1996).

In a later section we provide evidence to prove the robustness of our results to these alternative effects and their interpretation. Moreover, while we cannot exclude that some of the above effects is at work, it should be stressed that their presence does not invalidate our estimation procedure to provide a structural estimation of the social multiplier, while the randomized experiment in the data allows us to identify precisely behavioral interactions (i.e. cheating) during an exam.

---

[21] See Cooley (2010a) for an illustration of the general case.

3. Institutional context, data and descriptive statistics

The Protocol for the SNV survey entails the use of external inspectors for the administration of the tests, in a representative and random sample of classrooms. We define a 'sampled school' as a school where there are one or more 'monitored classrooms', and a 'monitored classrooms' (in a sampled school) as a classroom where an inspector is present during the test. Moreover, a 'non-monitored classroom in a sampled school' is a classroom in a sampled school where the inspector was not present. The natural experiment in SNV surveys administration determines a random variation in the type of classrooms subject to and not-subject to the external monitoring (monitored *versus* non-monitored classrooms) which is exploited to identify social spillovers due to students' cheating behaviors.

3.1 The National Survey of Students' Attainments

Starting from 2009-10 school-year the 'National Institute for the Evaluation of the Education System' (Invalsi, from now onwards), carries out a yearly evaluation of students' attainment and schools quality administering the SNV survey based on questionnaires and test scores evaluations[22]. SNV takes the form of an annual census, since it is compulsory for all schools and students attending the second and fifth grade, in primary schools, and the sixth and eighth grade, in junior high schools (about 500,000 students in each grade)[23]. Each student takes a test in Mathematics and Language in two different days in late May. Test administration and marking is carried out by school teachers, while Invalsi enforces a detailed Protocol (i.e. Invalsi, SNV Report 2010) for the administration and marking of the tests to reduce the possibility of teachers' cheating. For example, as often done in National Evaluation programs (Eurydice, 2009), the test is not administrated by the class teacher but by teachers of other classes and specialized in a different subject with respect to the one that is tested. All school teachers are simultaneously involved in the marking process, so that they cross-check each other during the marking, and the School-head - who is responsible for the correct implementation of the Protocol - supervises the whole process. Finally, an external specialized institution is charged to compute the test scores using an automatic procedure. However, what cannot be excluded *a priori* is that teachers adopt forms of soft monitoring. Teachers might

---

[22] Since 2005-06 school year a similar survey was carried out on a representative sample of schools, while all the other schools not in the survey sample were invited to participate on a voluntary basis.

[23] The choice of these grades corresponds to the requirement to test students' abilities at the beginning and at the end of the education path in primary and junior-high school levels. Formally, 8th grade test is part of the final exam at the end of the junior high school and follows different procedures and protocols. Pupils with disabilities are recognized by a team of specialists since the beginning of their schooling path, sit special formats of the tests and their results are not included in the official reports. In any case, it is not possible to change their 'disability' status during the school year.

simply adopt some form benevolent supervision because they allow students to exchange information or use prohibited material, or even because they are not able to implement a strict monitoring simply because of classrooms dimensions. Another kind of teachers' benevolent behavior which is not possible to control ex ante concerns the so called 'teaching to the test activity' (Lazear, 2006; Jacob, 2005; Kohn, 2007). For instance, since the beginning of the SNV surveys in 2008, it has become a common practice in many schools as teachers want to prepare students to test and quiz like the ones that they have to solve the day of the exam[24].

3.2 The randomized experiment in SNV data

External inspectors are sent to administrate and mark the SNV tests in a representative and random sample of classrooms both to validate the general results of the survey and give each school a 'certified' benchmark. In particular, inspectors are required to perform a number of tasks in the selected classrooms: (i) invigilate students during the tests, (ii) provide specific information on the test administration, (iii) compute the test scores and send results and documentation to Invalsi within a couple of days (Invalsi, 2010).

The allocation of inspectors to a random sample of classroom in the SNV data provides the ideal framework for our empirical strategy, for it introduces a random treatment with respect to the possibility of students to interact exchanging information or cooperating during the test − i.e. 'monitored classrooms' constitute the treated group of students, while 'non-monitored classrooms' are the control group. While, the possibility of any interactions among the students (cheating behavior) in the monitored classrooms is totally excluded and rigorously tested by Invalsi (Invalsi, 2010)[25], there is evidence that students in the non-monitored classroom received a more 'benevolent' supervision allowing the possibility of exchange of information and cooperative interactions. The latter is also confirmed by a number of studies which have used Invalsi data to investigate the extent of 'cheating bias' in test scores (Invalsi, 2010; Ferrer-Esteban, 2012; Bertoni et al. 2012; Castellano et al. 2009)[26]. Given that the choice of the monitored classrooms was random and done after classrooms formation, there is no sorting or matching between the treatment and school or classroom characteristics. The only exclusion criterion from the sample is constituted by

---

[24] A confirmation can be easily found looking at how text books have changed with the introduction of the SNV Program and started to include tests and quiz similar to the SNV exams structure.

[25] To test this Invalsi implemented sophisticated statistical techniques based on fuzzy-logic algorithms – i.e. see, Castellano et al. (2009) - and reported no evidence of cheating in the monitored classes.

[26] Bertoni et al. (2012) find that the presence of the external inspector reduces the average score (i.e. in terms of percent of correct answers) in the classroom by 5.5 to 8.5 per cent as compared to classrooms in schools with no external monitoring. They also find evidence of indirect effects on non-monitored classrooms in sampled schools, although the magnitude of the effect in this case is much smaller.

classrooms with less than 10 students[27]: this feature will require a careful analysis in the empirical estimations (see Section 5). On average, monitored students correspond to 7-8% of the total student population in each grade, while sampled classrooms corresponds to 6-7% of the total number of classrooms in each grade.

3.3 Data and descriptive statistics

In the empirical analysis we use the 2009-10 SNV data for sixth graders[28], for each student SNV data provides the test score for Math and Language and micro-data containing individual level information which are discussed in detail in Appendix B. Test scores are obtained as percentage of right answers for each subject and standardized with zero mean and unit standard deviation[29]. Individual characteristics cover information on gender, year and place of birth, Italian citizenship, grade retention, kindergarten attendance and school and class (anonymous) identifier.

<< Table 1 here >>

<< Table 2 here >>

Table 1 sums up the major characteristics of the dataset: number of schools, classes and students by each grade tested, average number of students per school and class while Table 2 shows that the two groups are not different in terms of observable characteristics. The only systematic difference is found in the presence of immigrant students who are oversampled. This feature suggests particular care when estimating the social multiplier (see Section 5). The two groups mainly differ because of the cheating behaviors of non-monitored students. Invalsi excludes the possibility of any interactions among students in monitored classes (SNV 2010 Report, Appendix 10, p.330) and provides statistical evidence of cheating behavior occurring in non-monitored classes by computing an index of 'cheating' (i.e. a class-level and subject-specific indicator ranging from 1, cheating is high, to 0, no cheating)[30]. The statistical method implemented by Invalsi highlights a high probability of cheating behaviors in non-monitored classrooms: the average cheating coefficients are .97 for Math and .92 for Language tests (Table 1). On the contrary, Invalsi Report shows that cheating coefficients for monitored classrooms are statistically not different from 0.

Finally, Table 3 provides statistical evidence on the differences in test score results between monitored and non-monitored students.

---

[27] In case in which a class with less than 10 students was selected, it was discarded and replaced with another class.

[28] We also repeat the analysis using SNV 5th and 2nd grade data in the robustness section.

[29] Students with special education needs take appropriate versions of the tests compatible with their physical or mental disability. Their results are not available due to privacy regulation restrictions.

[30] Invalsi uses these techniques to detect cheating behaviors also in other surveys and official national examinations. For further details about the "fuzzy c-means clustering" technique which is at the base of the indicator, see Castellano et al. (2009), Dunn (1973), Bezdek (1981).

<< Table 3 here >>

The mean and the median test score of non-monitored students is generally higher compared to monitored students, while the total variance is lower. The effect of the inspector's supervision becomes more clear when we decompose the total variance in its within- and between-class components[31]: within-class variance is greater in monitored classes while the between class variance is lower.

4. Empirical strategy

To identify the endogenous social multiplier effect originating from students' cheating behavior, we implement the Excess-Variance approach developed by Graham (2008). This approach, by relying only on the cross-group variation that originates from endogenous social effects, allows a direct estimation of the (structural) social multiplier - i.e. parameter γ in equations [9] and [10] as derived in the theoretical section. One advantage of this empirical strategy is that it bypasses most of the identification problems that characterize the classical reduced-form linear-in-means model[32]. For example, most studies in the social interactions literature (i.e. Gleaser et al. 1996, 2003; Entorf and Lauk, 2006; Entorf and Tatsi, 2009) have not been able to reliably separate the different sources of variability of individual and group level heterogeneity from the 'excess variability' genuinely originating from social interactions (Sacerdote, 2010)[33]. Moreover, the E-V approach has other notable advantages: first, it is robust to individual and group-level heterogeneity; second, the data requirements necessary to overcome the bias originating from standard omitted bias variable – i.e. which is a rather fundamental problem in social interactions setting due to the various sources of correlated effects - are very limited[34].

In practice, we observe N classrooms, each composed of N_c students. For each student we observe y_i, the outcome variable (test score), Z_c and Ψ_c, vectors containing group-level information, while

---

[31] The formula is corrected with appropriate weights to take into account the different size of the subgroups (i.e. classes) (see Ammermüller and Pischke, 2009)

[32] For example using proxy for peers' education level (Hanushek et al. 2003, 2009) or having to rely on specific exclusion restrictions (De Giorgi et al. 2011; Bramoullé et al. 2009).

[33] Some recent papers in the social interaction literature refer to the concept of social multiplier as the 'multiplicative effect due to social interactions' and derive the estimation of the multiplier indirectly (e.g. Maurin and Moschion, 2009; for female labour market participation decisions; Drago and Galbiati, 2012, for crime and recidivism).

[34] Durlauf and Tanaka (2008) discuss the advantages of the E-V approach compared to the regression approach and conclude that the former requires stronger assumptions on the variance covariance matrix which are not needed in the classical estimation of peer effects parameters from linear-in-means models. However, the authors suggest that E-V can be better justified whenever the sort of exclusion restriction needed on the variance covariance matrix of the outcomes can be substituted by appropriate prior information on the variance matrix structure. Our implementation of the EVA follows exactly this direction: we implement EVA exploiting the exclusion restriction which directly arises from the natural experiment in Invalsi SNV data.

individual-level (ε_i) and classroom-level heterogeneity (μ_c) are unobserved latent variables. Following Galbiati and Zanella (2012), we can rewrite the reduced form model from equations [10] and [9] in variance-components: let the classroom-level heterogeneity be, $\mu_c = \tilde{\delta}_x \bar{x}_c + \delta_\mu \mu + \delta_0$; the individual-level heterogeneity, $\varepsilon_i = \delta_x x_i$; and the classroom-level average of individual heterogeneity, $\varepsilon_c = \delta_x \bar{x}_c$. This transformation yields the following behavioral equations:

$$y_{ic} = \varepsilon_i + (\gamma - 1)\varepsilon_c + \gamma \mu_c \qquad [11]$$

$$\bar{y}_c = \gamma \varepsilon_c + \gamma \mu_c \qquad [12]$$

The social multiplier parameter to be estimated is, γ (with $\gamma \geq 1$), which captures the equilibrium social effect on individual achievement (i.e. test score) due students' cheating cooperative behavior during the exam. Equation [12] shows that the social multiplier is related to both the average of classroom-level (individual) heterogeneity, ε_c, as well as to the classroom-level heterogeneity, μ_c, such that - as implied by the theoretical model – exogenous shocks to contextual factors can also contribute (feeding-back through individual behaviors) to amplify the effects social externalities[35].

4.1 The Excess-Variance approach

A simplified notation for the conditional variances and covariance of individual and group-level heterogeneity is given hereafter: let $\sigma_\varepsilon^2(Z_c, \Psi_c) = \sigma_\varepsilon^2$ be the conditional variance (i.e. on Z_c and Ψ_c) of individual-level heterogeneity; $\sigma_{\varepsilon\varepsilon}(Z_c, \Psi_c) = \sigma_{\varepsilon\varepsilon}$ the conditional covariance of across individuals heterogeneity; $\sigma_\mu^2(Z_c, \Psi_c) = \sigma_\mu^2$ the conditional variance of group-level heterogeneity; $\sigma_{\mu\varepsilon}(Z_c, \Psi_c) = \sigma_{\mu\varepsilon}$ the conditional covariance of group-level heterogeneity with individual heterogeneity; while $V_c^w(Z_c, \Psi_c) = V_c^w$ and $V_c^b(Z_c, \Psi_c) = V_c^b$ are, respectively, the within-group and the between-groups conditional variance. Notice that: $\sigma_{\varepsilon\varepsilon}$ can be considered a measure of the degree of student sorting across classrooms; while σ_μ^2 represents the variance of unobserved teachers' characteristics, such as experience, strictness, ability and effectiveness, as well as the variance of all other unobserved characteristics that are common to all students in a classroom; $\sigma_{\mu\varepsilon}$ is a measure of 'matching' between these characteristics and the students. The latter is non-zero any

---

[35] Note that Graham (2008) defines the social multiplier parameter as a combination of both endogenous and exogenous peer effects - as group level heterogeneity there is obtained through group level averages of both observable individual characteristics and unobservable behaviors –, while in our setting it incorporates only the endogenous part of the cheating interactions.

time teachers (or classroom characteristics) and students are not randomly allocated – i.e. student can choose the school or, within each school, the classroom in which enrol. Then assuming that: $\sigma_{\varepsilon\varepsilon}(Z_c, \Psi_c) = \sigma_{\varepsilon\varepsilon}$, $\sigma_\mu^2(Z_c, \Psi_c) = \sigma_\mu^2$, $\sigma_{\mu\varepsilon}(Z_c, \Psi_c) = \sigma_{\mu\varepsilon}$ are independent of Z_c; and that the portion of the between-group variance independent from the within-group variance can be approximated by a linear function, such as:

$$\gamma^2 \left[\sigma_\mu^2(\Psi_c) + 2\sigma_{\mu\varepsilon}(\Psi_c) + \sigma_{\varepsilon\varepsilon}(\Psi_c) \right] = \pi \Psi_c \qquad [13]$$

Graham (2008) shows that V_c^w and V_c^b can be rewritten as follows:

$$V_c^w = \mathrm{E}\left[\frac{\sigma_\varepsilon^2(Z_c, \Psi_c) - \sigma_{\varepsilon\varepsilon}(\Psi_c)}{N_c} \,|\, Z_c, \Psi_c \right] \qquad [14]$$

$$V_c^b = \gamma^2 \left[\sigma_\mu^2(\Psi_c) + 2\sigma_{\mu\varepsilon}(\Psi_c) + \sigma_{\varepsilon\varepsilon}(\Psi_c) + V_c^w \right] \qquad [15]$$

where, substituting expression [13] into [14], it yields:

$$V_c^b = \pi \Psi_c + \gamma^2 V_c^w \qquad [16]$$

It is easy to show that the within-group variance of students' achievements in classroom c (denoted V_c^w in equation [14] above) is independent of social interactions and classroom-level heterogeneity. Note that, within-classroom differences in individual cheating behavior, when teachers are not sufficiently scrupulous in supervising students during exams such that cheating occurs, cannot be attributed to social externalities – since in our model, by definition, are the same for all students - but only to differences in individual characteristics and the covariances arising from students' sorting. Conversely, the between-group variance (denoted V_c^b in equation [15] above) depends on classroom heterogeneity and, when students' cheating behavior occurs, is magnified by social externalities. In this case, part of the variability in students' achievement between two different classrooms, one in which teachers do not strictly supervise students and another where strict monitoring is efficiently enforced, must necessarily depend on supervision. Then, since students' achievement in a classroom is also driven by cheating interactions, the cross-classroom variation will be affected. Cheating interactions introduce a wedge between the variance of students' achievements (measured by test scores) at different levels of aggregation, which is what we exploit to identify the social multiplier.

Expressing the conditional variances, as in [14] and [15] above, as conditional expectations of the relative within- (G_c^w) and between-classroom (G_c^b) statistics, namely: $V_c^w = \mathrm{E}\left[G_c^w \mid Z_c, \Psi_c\right]$ and $V_c^b = \mathrm{E}\left[G_c^b \mid Z_c, \Psi_c\right]$, we can rewrite equation 17[16] as[36]:

$$\mathrm{E}\left(G_c^b \mid Z_c, \Psi_c\right) = \pi\Psi_c + \gamma^2\left[\mathrm{E}\left(G_c^w \mid Z_c, \Psi_c\right)\right] \qquad [17]$$

which implies the following conditional and unconditional moment restrictions, respectively:

$$\mathrm{E}\left[G_c^b - \pi\Psi_c - G_g^w \mid Z_c, \Psi_c\right] = 0 \qquad [18]$$

$$\mathrm{E}\left[\begin{pmatrix} Z_c \\ \Psi_c \end{pmatrix}\left(G_c^b - \pi\Psi_c - \gamma^2 G_c^w\right)\right] = 0 \qquad [19]$$

Equation [19] delivers the appropriate specification to estimate (i.e. by GMM) the social multiplier, γ^2, using Z_c as instrumental variable.

4.2 The identifying assumption

The randomized experiment in Invalsi SNV data provides the ideal setting for identification. We observe two classrooms with, otherwise identical, students interacting in different ways: in one classroom achievement can also be attained by student cooperative behavior (i.e. control group); in another classroom external monitoring limits students' possibilities to interact, such that achievement is only based on individual effort (i.e. treatment group)[37]. Given the perfect randomization in treatment assignment, both individual and group level heterogeneity are likely to be the same across the two classrooms, such that the only difference in achievement between the two is the one originating from social externalities in students' cheating behavior: which are present only in the control group. Notice, that the presence of an inspector, by virtue of randomization, has no effect on the allocation of students and teachers to classroom, nor any effect on matching and sorting process of students' characteristics. According to our main identifying assumption (i.e. see equations [14] and [15]), Z_c generates an exogenous variation that affects the between-classroom variance in students' achievement only via the effect that cheating interactions have on the within-

---

[36] For each class c we observe the outcome for a (random) sample of students ($n_c \leq N_c$) given by all students who sit both Language and Math test scores. For this reason we rewrite expressions [14] and [15] using the appropriate statistics containing correction terms to take into account the difference between the sample and the population means. See Galbiati and Zanella (2012) web supplement for a formal derivation of conditional expectations.

[37] We may also expect that supervision is more efficient in treated-group classroom simply because of the joint presence of the inspector and a school teacher rather than just one teacher as in the control-group classroom. Note that in this case, the test score incorporates both the 'endowment type' peer effects (i.e. ability) and the 'behavioral peer effects' due to students' cheating interactions (see Section 6.3).

classroom variance. That is, by comparing the conditional variance of individual behavior within and between classrooms that we can identify the contribution due to endogenous social interactions only. In practice, we define a dummy variable identifying classrooms with external monitoring, ($Z_c=1$), and classrooms without external monitoring ($Z_c=0$)[38]. The standard rank condition for Z_c to be a valid instrument can be easily assessed empirically: $\mathrm{E}\left(G_c^w \mid Z_c = 1, \Psi_c\right) \neq \mathrm{E}\left(G_g^w \mid Z_c = 0, \Psi_c\right)$.

Since the model is just-identified, we can simply estimate it by two-stage least squares and given that the instrument, Z_c, is a dummy variable, the estimator of the social multiplier takes the form of a Wald estimator:

$$\gamma^2 = \frac{\mathrm{E}\left(G_c^b \mid Z_c = 1\right) - \mathrm{E}\left(G_c^b \mid Z_c = 0\right)}{\mathrm{E}\left(G_c^w \mid Z_c = 1\right) - \mathrm{E}\left(G_c^w \mid Z_c = 0\right)} \qquad [20]$$

The numerator is a contrast of observed (or actual) between-classroom variance in student achievement across treatment states (i.e. $Z_c=1$ versus $Z_c=0$). As discussed above, under perfect randomization, this contrast is purged of the influence of teacher heterogeneity, matching, and sorting; thus it solely reflects differences in the variance of achievements across the above treatment states as amplified by the cheating interactions. The denominator also equals the difference in the variance of achievements across the treatment states, but unaffected by social interactions (Graham, 2008; Sacerdote, 2010).

Finally, the feasible estimator requires an estimate of the conditional expectation of students' achievement $\mathrm{E}\left(y_{ic} \mid Z_c, \Psi_c\right)$ which we obtain from a regression of y_{ic} on Z_c and Ψ_c. We then use the residuals to replace G_c^b with $\hat{G}_c^b = (\hat{y}_c - Z_c' \hat{\pi}_1 - \Psi_c' \hat{\pi}_2)^2$, where $\hat{\pi}_1$ and $\hat{\pi}_2$ are least squares estimates. Randomization also implies that (in principle) we do not need to include any variable in the vector Ψ_c to control for sorting or matching of students with respect to assignment to treatment, Z_c, and class characteristics. Descriptive evidence provided in Section 3 shows that the two subgroups constitute a representative and random sample of the students population for the sixth grade (see also Appendix B). There are, however, a couple of matters for concern: first, we may need to control for the share of immigrant students as they appeared to be slightly oversampled in treated classroom (see Table 2); second, there may also be spill-over effects of external monitoring in treated classroom on non-monitored classrooms of sampled schools, which we need to control for (Bertoni et al. 2012). For these reasons, we include two additional controls: a dummy variable

---

[38] Graham posits that identification relies on: "[…] two subpopulations of social groups where assignment to groups is as if random" (Graham, 2008, p. 658). In his paper, Graham identifies a social multiplier arising from differences in peer quality across groups, in our setting however peer quality is homogeneous across groups the only source of excess variation being cheating behavior.

indicating whether a classroom is a 'non-monitored class in a sampled school', and a dummy variable indicating whether there is a 'high share' of immigrant students in the classroom (i.e. takes value 1 if the immigrant share is greater than the 75th or the 90th percentile of the immigrant class share distribution). We discuss further extensions in the following section.

5. Results

The estimates of the social multiplier are obtained through two-stages least squares where we regress the feasible estimator for the between-groups variance ($\hat{G}_c^b$) on the additional controls (Ψ_c), and on the within-groups variance, $\hat{G}_c^w$, instrumented by the class type indicator (Z_c). We first report our estimates of equation [20], without including any control variable (i.e. baseline social multiplier), then we progressively add other control variables to the vector Ψ_c to account for selected features of randomization, or test the existence of spill-over effects. Social externalities exist if the social multiplier is different from one (eq. [11] and [12]), thus we test the null that $\gamma^2=1$ and report the correspondent p-value in each table. To allow for the comparability of the results across subjects, we focus on all students who sit both Language and Math test scores. In fact, given that the tests were in two different, although subsequent, days there are students who sit just one of the two tests and students who do not sit none of them because they are absent in both days. The percentage of absent students in 6th grade is about 0.6%. As previously discussed, the only criterion Invalsi used in the randomized experiment to drop, *a priori*, some classrooms from receiving the treatment (i.e. external monitoring) was classroom size – i.e. less than 10 students (723 classes for corresponding to 2.7% of the total number). For this reason we conduct the analysis dropping classes with less than 10 students, while robustness checks to the inclusion of these classes are tested in the next section.

Note, that the E-V approach leaves the sign of the social multiplier (γ^2), in principle, undetermined (since we estimate its square). Hence, the sign has to be inferred from the underlying theoretical model which, in our case, posits a positive effect of social multiplier (γ and $J >0$) due to the assumption of students' cooperative effort, such that cheating interactions among students during the exam are likely to increase each student's achievement and the class average performance. Next, we explore the heterogeneous effect of social interactions comparing sub-populations with a different degree of heterogeneity according to a set of selected (exogenous) characteristics.

5.1 Baseline estimates

<< Table 4 here >>

First stage F-statistics reported in Table 4 show that instruments are not weak and the standard rank condition is always satisfied (the coefficient of the excluded instrument is always positive and statistically different from zero at 1% confidence level). First-stage results, not reported in the tables, indicate that in monitored classrooms the variance of the tests scores is higher compared to non-monitored classrooms. This reflects the larger dispersion of individual heterogeneity in test scores when behavioral interactions are not at work and students cannot exchange information or engage in any cooperative effort.

From our baseline specification we obtain an estimate γ^2 of 5.13 for Math and 4.18 for Language. Progressively adding the control variables described above does not alter the results: estimates for Math range between 5.13 and 5.89, while estimates for Language range between 4.18 and 4.77. This confirms that the two subgroups are (almost) identical in terms of observable characteristics, and that adding control variables (included in the Ψ_c vector) only has a negligible effect on the estimated social multiplier. All estimates are significantly different from 1 at 1% confidence level: this means that we can strongly reject the null of 'no social interactions' (i.e. that $\gamma=1$, Graham, 2008).

Our results imply a strong amplifying role played by social interactions within students in the classroom. The above estimates correspond to values for γ ranging between 2.26 and 2.43 for Math, and between 2.05 and 2.18 for Language, and values for J ranging between 0.56 and 0.59 for Math, and are slightly lower for Language (0.51 - 0.54)[39]. In terms of our structural parameters, a cheating social multiplier close to two (i.e. $\gamma \in [2.05;2.43]$) means that cooperative behaviors, when external monitoring is loose or benevolent, may generate a change in the equilibrium of students' achievements that is twice as big as the class average achievement without behavioral interactions (equation [9]). In terms of individual test score, the estimates for J (i.e. $J \in [.51;.59]$) imply that the marginal contribution due to cheating increases individual test score by almost a half of the standard deviation (equation [10]), which corresponds to almost 10 points in Math and 8 points in Language tests[40].

For what concerns the general pattern of the results with respect to the two subjects, the magnitude of the estimated social multiplier is slightly larger in Math with respect to Language. This small difference can be explained considering that cheating behavior may be easier for mathematics,

---

[39] Standard errors for the model parameters (γ, J) are obtained using the delta method. The delta method expands a function of a random variable (i.e. the estimated parameters) about its mean with a one-step Taylor approximation. Then, it computes the variance to obtain an estimate of the standard errors (see Davidson and MacKinnon 2004, chap. 5.6).

[40] According to the corresponding values of J, the cheating marginal contribution for Math ranges between 10.1 – 10.7 points. For Language it is sensibly smaller (between 7.8 – 8.2 points).

which are based on closed answers and quiz, rather than in language since text comprehension exercises require more effort and longer time to get through the text, to interpret it and derive the answers. This result is also in line with educational psychology literature which finds that cheating occurs more frequently in the hard sciences compared to the arts and social sciences (Miller *et al.*, 2007).

Our results, although not directly comparable, confirm in general the evidence available from other studies in the social interactions literature (see Carrel et al. 2008; Glaeser et al., 1996; Drago and Galbiati, 2012; Maurin and Moschion, 2009) which find social multipliers between 2 and 3 in order of magnitude. A more direct comparison can be done with those studies that use the E-V approach to recover an estimate of the social multiplier. In his analysis of students' peer effects in class learning activities, using Project STAR data, Graham (2008) reports an estimate for the social multiplier of approximately 1.9 for Math, and 2.29 for Reading. Galbiati and Zanella (2012) estimate a social multiplier arising from congestion externalities in tax cheating between 3.1 and 3.2. In other words, in all the above settings an exogenous shock altering the variable subject to social interactions (respectively, school achievement and concealed income) produces an equilibrium variation that is between two and three times the initial response. Note, however, that when comparing the results reported in Graham (2008) and Galbiati and Zanella (2012) to our own, some important differences should be born in mind. First, while we exploit the identifying restriction given by the natural experiment in Invalsi SNV data, both Graham (2008) and Galbiati and Zanella (2012) identify the social multiplier through exogenous variations in the size of the reference group. As standard in this literature (Sacerdote, 2001; Imberman et al. 2012), Graham's social multiplier due to peer interactions in achievement embeds both exogenous and endogenous effects[41]. Galbiati and Zanella (2012) provide a structural interpretation of the social multiplier generated by externalities in concealed income due to tax congestion within Local Tax Authorities so that their social multiplier represents an upper bound of the long run effects of the endogenous effects of tax cheating.

5.2 Heterogeneous effects

We exploit the richness of individual-level information in the SNV data to explore different dimensions of students' characteristics which may give rise to heterogeneous effects in cheating behavior. In practice, we test whether the social multiplier differs across selected subpopulations of classrooms characterized by large amounts of heterogeneity in some observed students' attributes,

---

[41] Graham (2008) points out that the estimated structural parameter for the social multiplier should be referred to an explicit structural model to highlight the underlying social mechanisms which originate the peer effects.

with respect to a subpopulation of classrooms with low heterogeneity (Graham, 2008). Since cheating requires some cooperative effort between students within each classroom, one may expect that classrooms in which students are more homogeneous with respect to some exogenous attributes[42] exhibit stronger social interactions as compared to classrooms in which students are more heterogeneous. This corresponds to test whether there is complementarity or substitutability between the intensity of cheating behaviour, due to looser external monitoring, and the strength of classroom social ties. With complementarity, moving a group of students with more homogeneous characteristics and stronger social ties (i.e. low heterogeneity subpopulations) to a non-monitored classroom should, in addition to increase average test scores, reduce its variance more than for a comparable group of students with less homogeneous characteristics (i.e. high heterogeneity subpopulations). Thus, if external monitoring and classroom heterogeneity are complementary, the social multiplier estimated on the low heterogeneity subpopulations should be greater compared to the one calculated on the high heterogeneity subpopulations. If they are substitutes, the opposite will occur.

In particular, we select the following attributes for the subpopulations: number of books at home, sport practice (outside school), participation to outside school activities (other than sport, e.g. music, arts and foreign languages courses) and time spent playing with friends (outside school)[43]. In all the above cases, we split the sample of classrooms into two groups characterized by high and low degrees of heterogeneity. We refer to the number of books that students have at home as a proxy for heterogeneity of parental background in terms of education and socio-economic status (Ammermüller and Pischke, 2009). In this case, the high (low) heterogeneity group is defined as the subpopulation of classrooms having a standard deviation higher or equal (lower) to the median standard deviation observed in the entire classroom population. The 'sport', the 'outside school activities' and the 'time spent playing with friends' variables are themselves a proxy of the strength of the social links within each classroom, measured as the amount of time classmates meet and spend time together outside the school. Classrooms in which social ties, proxied by the above variables, are below the median level of the whole population belong to the high heterogeneity group. In other words, classrooms above the median level of these variables encompass situations in which a lot of students interact more outside school (sport, music, arts, playing with friends) thus showing stronger social ties. An opposite reasoning is true for classrooms below the median levels.

<< Table 5 here >>

---

[42] Note that all the attributes are considered exogenously pre-determined with respect to students' achievement during the exam.
[43] See Appendix B for further details on the definition of the variables.

Table 5 shows the main results: for each selected attribute we report γ^2 - the square of the social multiplier - for the group of classrooms with high and low heterogeneity, respectively, and test the null of no differences (p-values reported)[44]. We exclude from the analysis students with missing values in any of the four variables used and drop classrooms with less than ten students because of the above discussions[45]. First stage F-statistics show that the effect is always strongly identified. We find that the social multiplier is larger in the subpopulation of classroom with low heterogeneity with respect to parental background characteristics and students' outside school activities both in Language and Math. For Language, the same result holds also for the sport variable. No statistically significant difference is detected with respect to time spent playing with friends. This suggests that higher strength of social ties and more homogeneous classrooms in terms of family socio-economic background favour social interactions in cheating behavior[46].

In general, we find support for the hypothesis that cooperative efforts in cheating interactions require a more homogeneous pool of classmates and deliver a greater social multiplier. In particular, the results for the sport practice and the outside school activities variables, seem to suggest that practicing sport with classmates outside school and doing other leisure activities such as arts and music courses are to be considered complementary to the social links that are useful to support cheating.

6. Robustness

We test the robustness of the empirical results taking into account different forms of social mechanisms that could affect our estimates of the cheating social multiplier. For example, we investigate whether teachers' cheating in non-monitored classes, or stress induced by the presence of an external inspector in monitored classes may explain (part of) the gap in performance between monitored and non-monitored students, as opposed to students' cheating. Next, we replicate on our data Graham's empirical analysis of achievement peer effects in the Tennessee Schools STAR Project. All the robustness checks support our identification strategy and show that estimated values of the cheating multiplier are not affected by alternative mechanisms that could bias the results. There alternatives are discussed hereafter.

---

[44] We test the null, H_0: $\gamma^2_H = \gamma^2_L$, using the Sargan-Hansen test of over-identification associated with the estimates of the combined sample where the binary instrument (monitored/non-monitored classrooms) and its interaction with the high heterogeneity dummy serve as excluded instruments (Graham, 2008).
[45] The same pattern of results holds keeping classrooms with less than 10 students. Dropping these classrooms slightly improves p-values for the 'books at home' variable.
[46] We also computed different effects for high and low heterogeneity in 'ability' (i.e. proxied by the variance in teachers marks for each classrooms at the end of the first semester). We find that the cheating social multiplier is higher the more the class is homogeneous in terms of teachers marks (ability). Results are not reported but are available upon request.

6.1 Teachers' cheating

Several forms of teachers' cheating are discussed in the literature. There could be totally illicit activities, so called 'explicit cheating', such as changing student responses on answer sheets, providing correct answers to students, or obtaining copies of an exam illegitimately prior to the test date and teaching students using knowledge of the precise exam questions. There is 'hidden cheating' in which educators attempt to raise a school overall performance profile by retaining low-scoring students in grade, classifying more students as 'special needs' in order to exclude their scores from school averages, or lavishing attention on students who are close to passing, and ignoring those who are sure to do well and those likely to fail (Kohn, 2007). Additionally, there could also be 'soft' forms of teacher cheating such as 'teaching to the test'.

One reason why teachers' cheating should not play a significant role in the Italian schools is due to the fact that the career of teachers follows a simple experience-age rule and is not linked in any way to students' performance. In fact, teachers' cheating has been found to be a substantial problem when high-stakes testing programs are introduced in the school system (Jacob, 2005; Jacob and Levitt, 2003a,b). Moreover, Invalsi controls that the SNV Protocol is strictly followed by school teachers and School-heads are responsible for any illicit behavior of the school staff. However, teachers may be induced in illicit behavior because, for example, they simply dislike sharp differences in results across classes within the same school (Bertoni et al. 2012). Anytime teachers help students in suggesting the right answers or changing their answers while marking the test, the estimates for the social multiplier will also include this component and be upward biased.

Ferrer-Esteban (2012) and Bertoni et al. (2012) analyse the effects of monitoring on students test scores using SNV data and show that external monitoring has a negative effects on students' test scores. Bertoni et al. (2012) use Math tests of elementary school students (5th grade) and argue that the better performance of classes without the external inspector is due to the manipulation of tests by students and/or teachers. The authors do not distinguish between students and teachers' cheating so that they interpret the performance gap between monitored and non-monitored classrooms as a measure of the average intensity of (generalized) cheating taking place in non-monitored classrooms. They also show that spill-overs effects are present in non-monitored classrooms of sampled schools. This fact also justifies the inclusion of the 'non-monitored classroom in sampled school' indicator variable in the vector of controls. Ferrer-Esteban (2012) uses data both from elementary schools (2nd to 5th grades) and junior high schools (6th to 8th grades) in the 2009-10 SNV to build an individual level cheating indicator. Similarly to Jacob and Levitt (2003 a,b), a student is suspected of cheating if the entire path of the answers of the test - item by item, independently of

whether answers are right or wrong - is equal to the one of a class-mate. He shows that the distribution of 'suspected cheaters' conditional on the result in the tests is sharply different across grades. In the elementary schools 'suspected cheaters' are all distributed in the upper tail of the test score performance distribution while in the junior high schools 'suspected cheaters' are normally distributed along the test score performance range of results. The author interprets this evidence as teachers' cheating playing a substantive role especially in the elementary schools, as 'suspected cheaters' always give right answers as if they are suggested by teachers and not by each other copying or cheating. Taken together, Bertoni et al. (2012) and Ferrer-Esteban (2012) studies suggest that teachers' cheating - if any - is particularly concentrated in elementary schools and less in the junior high schools.

As robustness check, we replicate the analysis on elementary school students in 5th and 2nd grades who sit the 2009-10 SNV test (see Appendix B for details). Grounding on the aforementioned studies, we expect cheating social multiplier to be higher in magnitude than the 6th grade as it potentially includes bias given by teachers' cheating which is likely to increase class average test scores.

<< Table 6 here >>

Table 6 shows descriptive evidence on test score means and variances across grades. It is easy to notice that the gap between mean test scores of monitored and non-monitored students is much higher in the elementary grades compared to 6th grade (Language test score gap between sixth grade monitored and non-monitored students is not even statistically different). The same is true for the total variances, while the variance within classes is always higher in monitored classes.

<< Table 7 here >>

The estimates of the cheating social multiplier for 5th and 2nd grade students do not show significant differences in terms of strength of identification and statistical significance, but they are always higher in magnitude (Table 7)[47]. This confirms that teachers' monitoring is looser in elementary schools as compared to junior high schools. Restricting our main analysis to 6th grade students, thus minimizes possible bias due to teachers' cheating behavior.

6.2 Stress induced by external monitoring

The presence of an external inspector in the classroom during the test (under the external monitoring regime) may exert psychological pressure or induce stress among student, which might

---

[47] Because of the differences in the test structure, 2nd grade Language results are not directly comparable across grades. School and family background information are not provided for 2nd graders as students do not have to fill in the 'Student Questionnaire'. The estimates obtained without dropping classes with less than 10 students (not included in the text) do not change the overall pattern of the results and confirm their robustness.

alter their performance and lower the average test score in monitored classroom. In this case, the observed gap in test scores between monitored and non-monitored classrooms might incorporate a component that is due to psychological stress. We use the SNV 'Student Questionnaire' (see Appendix B), which contains a set of motivational questions that students have to answer immediately after taking the test, to ascertain the emotional feelings and psychological pressures that students experience while taking the test or preparing for it[48].

<< Figure 1 here >>

We find no difference in the answers to the motivational questions between monitored and non-monitored students (Figure 1), which leads us to exclude that our estimates might be biased (upward) due to the stress induced by external monitoring. Exploiting the same variables for elementary schools (5th grade), Bertoni et al. (2012) discuss in detail the possibility that young students under-perform as a consequence of the distraction induced by the presence of a stranger in the class and find no evidence that being in a classroom with an external inspector increases anxiety or nervousness.

6.3 Achievement peer effects and class-size

Since randomization ensures that students' quality across monitored and non-monitored classrooms is the same, social interactions can only arise from students' cheating behavior. In this section, we test this proposition and investigate whether a more conventional 'peer effects in achievement' may also influence the social multiplier we estimate. Peer effects may work either *via* peers' characteristics (contextual effects such as aptitude to learn, readiness, ability to focus), or *via* alternative endogenous social interactions (such as information gathering, endogenous preference formation, congestion externalities) (Sacerdote, 2001). We replicate the empirical strategy proposed by Graham (2008), which relies on classroom-size variation as instrument, estimating our model separately on for monitored and non-monitored classrooms (Lazear, 2001; Graham, 2008; Carrel et al. 2009; Cooley, 2010a,b)[49]. The key assumption, in this case, is absence of sorting and unobserved heterogeneity across small and large classrooms. Since, general rules for class size formation in junior high schools are considerably influenced at the school-district level by the availaby of tenured versus non-tenured teachers and the allocation of resources across schools in the same district, we include in our baseline specification the usual classrooms level controls (Ψ_c), as well as

---

[48] Students are asked whether they totally agree / partially agree / partially disagree / totally disagree with the following statements: 'I already was worried before taking the tests'; 'I was so nervous I could not find the answers'; 'While taking the test I was calm'.

[49] Group-size is a good instrument for the E-V approach because, provided that group-level heterogeneity is the same across the two subpopulations (small vs. large classes), the dispersion of individual heterogeneity typically is not the same.

school-district fixed effects (i.e. 110 dummies corresponding to Italian provinces, NUTS 5 level). Specifically, we run the analysis separately for monitored and non-monitored classrooms and calculate an achievement (squared) social multiplier, that we label γ_a^2 to keep it distinguished from the usual cheating social multiplier[50]. We expect the value of the social multiplier estimated for the group of non-monitored classrooms ($\gamma_a^2|Z_c=0$) to be larger than the social multiplier estimated for the group of the monitored classrooms ($\gamma_a^2|Z_c=1$), since the former is likely to be inflated by cheating interactions while the latter is not.

<< Table 8 here >>

The instrument we use is a dummy for 'small class size' that takes value 1 if class size is below the median class size. Table 8 contains the different estimates for the (squared) social multiplier. The standard rank condition is satisfied, as the coefficient of the excluded instrument – not reported in the Table - is always positive and significantly different from zero at 1 per cent confidence level, and the first-stage F-statistics show that the effect is always strongly identified. The positive sign in the first stage regressions confirms that small class size tends to increase individual-level heterogeneity. Interestingly, estimates of the (squared) social multiplier are found to be not statistically different from 1 in the subgroup of monitored classes - where only interactions in achievement may have taken place -, while the estimates show up statistically different from 1, ranging between 2.08 and 3.21 (close to Graham's estimations), in the subgroup of non-monitored classes[51]. In other words, since we cannot reject the null of 'no-achievement social interactions' in the monitored classrooms (both for Language and Math), while we find sizable social interactions in the non-monitored classes, it seems reasonable to expect any effect of 'achievement social interactions' to be negligible as compared to the effect of 'cheating social interactions'.

<< Table 9 here >>

A final concern with respect to class-size might arise with respect to classrooms with less than 10 students which were dropped from the main analysis to meet the only ex-ante selectivity criteria implemented by Invalsi in the random selection of the monitored classrooms. To assess whether this threshold introduced some selectivity in the sample of treated versus control classrooms, we repeated the analysis also including all the classrooms with less of 10 students (723 classes for

---

[50] We run the analysis on the whole population as well as excluding classes with less than 10 students. Results do not change.

[51] Graham (2008) finds a (squared) social multiplier of 2.33 for Math and 2.11 for the Reading test scores in the complete specification. However, while we exploit junior high school students, Graham (2008) focuses on kindergarten students.

grade 6 corresponding to 2.7% of the total number)[52]. Results reported in Table 9 show no significant differences with respect to the baseline estimates.

7. Discussion and conclusions

There is abundant evidence showing that students' cheating has worsened over the last few decades, becoming a widespread practice in schools, college and high-ranked universities (Dee and Jacob, 2012). Experts say that cheating has grown hand in hand with high-stakes testing systems, such as the No-Child-Left-Behind-Act (2001) in the U.S. (Jacob, 2005), and it has become easier and more widely tolerated, as both schools and parents fail to give students clear messages about what is allowed and what is prohibited (*The New York Times*, September 7, 2012). In this paper we provide evidence on the social interactions which are generated when students' cheat - either exchanging information and cooperating with other students, or using any prohibited materials - while taking an exam. We develop a simple theoretical model describing the mechanisms that drive social interactions in cheating behavior, and show that students optimally decide whether or not to cheat taking into account other students' best response. We estimate the social multiplier generated by cheating behaviors using data from a randomized experiment in a national evaluation tests. Our findings suggest a strong amplifying role played by cheating social interactions in the classroom, which increases in the strength of social ties. The value of the social multiplier implied by students' cheating behaviors is estimated to be between 2 and 3 in all the specifications, suggesting that cooperative behaviors, when a strict external monitoring is missing, may generate a change in the equilibrium of students' achievements that is twice as big as the class average achievement. In terms of individual test score, the marginal contribution of cheating interactions increases individual test score by almost half of the standard deviation (i.e. between 7 and 10 points). Heterogeneous effects show that the strength of social ties in the classroom is a complementary input to cheating behaviors such that the effect is larger the more the classroom is homogeneous. Several sensitivity checks confirm the overall robustness of our results.

Our findings have a number of relevant policy implications. First, we show that tolerating cheating behavior, as it is often done in schools, is a very dangerous practice, since the social multiplier magnifies the negative effects on both students' performance and on the signaling role of education in the labor market. McCabe (2005) documents that a large share of college students considers cheating and other forms of illicit collaboration with classmates as a minor offence or no offence at

---

[52] Due to students absence on the day of the test, we do find classrooms with less than 10 students also in the treated group. This, of course, was not known ex-ante, and absent students re-sit in September.

all. He also finds that most high school teachers and college professors fail to report and pursue most of the violations that are detected. Moreover, commitment to academic integrity and sanctions to violations are still not adequately considered: few schools place any meaningful emphasis on academic integrity, and colleges are even more indifferent than high schools[53]. Our estimates also show that tolerating such behaviors is particularly relevant as cheating is likely to feedback onto social norms thus raising collective indulgence with respect to various forms of dishonest practices. In other words, ethical or honor codes of behavior in schools should be strictly enforced and students' cheating behavior reported and sanctioned. Second, given that increasing competition in the job market and high-stakes testing systems are likely to exert considerable pressure on students to perform well in exams, it should be recognized that where (and when) the pressure is higher, more resources should be devoted to monitoring activities in order to avoid cheating interactions to become widespread. In this sense, the social multiplier mechanism would also magnify the effects of policies directed to stricter monitoring and sanctioning of cheaters. From the policymaker perspective a commitment to rigorous monitoring and sanctioning - by changing the individual's private incentives to cheat, would deliver significantly larger social effects (Durlauf and Cohen-Cole, 2004). Our results also show that strong social links among classmates are likely to facilitate social interactions and cheating behaviors. In this context, a rather inexpensive way to reduce students' illicit behaviors would consist in a random reshuffling of students and teachers across classrooms, within any given school, so to reduce students' tendency to conform to other students' behavior. Finally, the presence of spill-over effects of monitoring in non-monitored classrooms of sampled schools, suggests that another rather inexpensive intervention to contrast cheating would be to spread the inspectors on more schools as non-monitored classrooms in sampled schools show a significantly lower degree of cheating interactions.

---

[53] Michael Josephson, president of the Institute for Academic Integrity, *The New York Times*, September 7, 2012.

Appendix A - Derivation of the achievement best response

Under the assumption that achievement is monotonically increasing in cooperative cheating effort behavior (henceforth simply referred to as *effort*), we can solve from the APF (eq. [1]) for the unobservable effort. Thus, for individual i and j - where j represents any i's classmate peer - we have:

$$e_i = y_i - x_i \pi_x - x_j \tilde{\pi}_x - e_j \tilde{\pi}_e - \mu \qquad [A.1]$$

$$e_j = y_j - x_j \pi_x - x_i \tilde{\pi}_x - e_i \tilde{\pi}_e - \mu \qquad [A.2]$$

Plugging the expression for e_j from equation [A.2] into [A.1] and solving for e_i, we obtain:

$$e_i = \left(\frac{1}{1-\tilde{\pi}_e}\right)\left[y_i - y_j \tilde{\pi}_e - x_i(\pi_x - \tilde{\pi}_x \tilde{\pi}_e) - x_j(\tilde{\pi}_x - \pi_x \tilde{\pi}_e) + \mu(\tilde{\pi}_e - 1)\right] \qquad [A.3]$$

Similarly, for individual j:

$$e_j = \left(\frac{1}{1-\tilde{\pi}_e}\right)\left[y_j - y_i \tilde{\pi}_e - x_j(\pi_x - \tilde{\pi}_x \tilde{\pi}_e) - x_i(\tilde{\pi}_x - \pi_x \tilde{\pi}_e) + \mu(\tilde{\pi}_e - 1)\right] \qquad [A.4]$$

Substituting equation [A.4] into the effort best response of individual i from equation [4] yields:

$$e_i^{BR} = \frac{\beta_y}{\beta_e} + \frac{\tilde{\beta}_e}{\beta_e}\left(\frac{1}{1-\tilde{\pi}_e}\right)\left[y_j - y_i \tilde{\pi}_e - x_j(\pi_x - \tilde{\pi}_x \tilde{\pi}_e) - x_i(\tilde{\pi}_x - \pi_x \tilde{\pi}_e) + \mu(\tilde{\pi}_e - 1)\right] \qquad [A.5]$$

Finally, substituting the LHS of equation [A.3] with the equation of the best response effort function from equation [A.5] and rearranging we obtain the expression of the achievement best response function (Cooley, 2010b):

$$\frac{\beta_y}{\beta_e} + \frac{\tilde{\beta}_e}{\beta_e}\left(\frac{1}{1-\tilde{\pi}_e}\right)\left[y_j - y_i \tilde{\pi}_e - x_j(\pi_x - \tilde{\pi}_x \tilde{\pi}_e) - x_i(\tilde{\pi}_x - \pi_x \tilde{\pi}_e) + \mu(\tilde{\pi}_e - 1)\right] =$$

$$= \left(\frac{1}{1-\tilde{\pi}_e}\right)\left[y_i - y_j \tilde{\pi}_e - x_i(\pi_x - \tilde{\pi}_x \tilde{\pi}_e) - x_j(\tilde{\pi}_x - \pi_x \tilde{\pi}_e) + \mu(\tilde{\pi}_e - 1)\right] \qquad [A.6]$$

$$y_i = \frac{\beta_y(1-\tilde{\pi}_e^2)}{\beta_e + \tilde{\beta}_e\tilde{\pi}_e} + x_i\left[\pi_x - \frac{\tilde{\beta}_e + \beta_e\tilde{\pi}_e}{\beta_e + \tilde{\beta}_e\tilde{\pi}_e}\tilde{\pi}_e\right] + x_j\left[\tilde{\pi}_x - \frac{\tilde{\beta}_e + \beta_e\tilde{\pi}_e}{\beta_e + \tilde{\beta}_e\tilde{\pi}_e}\pi_e\right] +$$

$$+ y_j\left(\frac{\tilde{\beta}_e + \beta_e\tilde{\pi}_e}{\beta_e + \tilde{\beta}_e\tilde{\pi}_e}\right) + \mu\left(1 - \frac{\tilde{\beta}_e + \beta_e\tilde{\pi}_e}{\beta_e + \tilde{\beta}_e\tilde{\pi}_e}\right);$$

$$= \frac{\beta_y(1-\tilde{\pi}_e^2)}{\beta_e + \tilde{\beta}_e\tilde{\pi}_e} + x_i\left[\pi_x - J\tilde{\pi}_e\right] + x_j\left[\tilde{\pi}_x - J\pi_e\right] + y_jJ + \mu(1-J);$$

$$= \delta_0 + x_i\delta_x + x_j\tilde{\delta}_x + y_jJ + \mu\delta_\mu$$

[A.7]

Appendix B - Invalsi SNV data description

School system in Italy starts with 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). These two form the 'first cycle' of the educational system which is compulsory and identical for all students, while secondary education lasts three to five years depending on the path chosen (vocational, technical, academic). Children enrol in the first grade of the primary school the year they turn six, and start the junior high school when they turn eleven. Primary and junior high schools are quite different in terms of organization and types of teaching activities. In primary schools pupils spend almost all school time with two teachers, one teaches Language, History, Geography and the other teaches Math and Science. The two 'reference teachers' usually follow the pupils from the first to the fifth grade establishing a strong personal link. Junior high school is more similar to high school. Students experience a kind of more rigorous teaching, with several professors, one for each subject, and acquire a wide range of core skills necessary to succeed in high schools.

Invalsi SNV data contain test scores results and individual level information. Individual level information are gathered in the dataset from three different sources: (i) students' general information from school administrative records compiled directly from school administrative staff on each student's answer sheet; (ii) family background information collected through a 'Family Questionnaire' sent to each family some days before the test; (iii) additional individual information on family, school and environmental characteristics collected through a 'Student Questionnaire' taken by each 5th and 6th grade students the same day of one of the test (after finishing the exam). They are collected by the school administrative staff on the same answer sheets of the students' test and are taken from the administrative register data which are given by the families at the moment of the child's enrolment (at the beginning of each school year in September). Other parental background information are available and cover mother's and father's place of birth (Italy, EU, European but non-EU, other non-European country), occupation and education level.

Some variables (e.g. kindergarten and pre-kindergarten attendance; parental occupation and education) which are not administrative data records kept by the school but rather provided by the families filling in the 'Family Questionnaire' suffer from a relevant problem of missing information (from 9% to 30% depending on the grade and variable). This problem may be considerably mitigated for 5th and 6th grade students exploiting additional information about school characteristics and family background contained the 'Student Questionnaire' which is filled by each student in the class the first day of the test and does not entail problems of missing data. Second grade students did not have to fill such additional information. The 'Student Questionnaire' is different for 6 and 5 graders, but the more relevant variables are common to both. From these sources we obtain variables that are commonly used as proxy for socio-economic background and family information in international programs of students' attainments testing (e.g. PISA, TIMSS) and applied research. For instance, students have to answer questions such as "How many books have you at home?", "Which language do you usually speak at home?"; "Do you currently speak dialect at home?".

In the heterogeneous effects analysis (Section 5.1) we exploit some variables taken from the 'Student Questionnaire'. The 'number of books at home' is a categorical variable with 5 levels (0-10 books; 11-25; 26-100; 101-200; more than 200). The 'sport' variable asks students how many times per week he/she practices sport activities outside school (never, 1 or 2, 3 or 4, more than 4). Similarly, the 'outside school activities' variable asks students how many times per week he/she takes part to leisure activities outside the school time (e.g. music, arts, theatre or foreign language courses)(never, 1 or 2, 3 or 4, more than 4). Finally, the variable indicating the time spent each day playing with friends outside school takes the following values: never, 1 hour or less, 1 or 2 hours, more than 2 hours.

B.1 The sampling procedure: randomness and representation

Invalsi exploits a simple random computer routine that ensures the representation of the sampled group of students, classes and schools. First, for each of the 20 Italian regions they randomly choose a representative sample of primary schools for grade 2 and 5, and a sample of junior high schools for grade 6. Then, within each school they randomly picked up one or two classrooms for each grade. The sampling procedure starts at the regional level, so that the final sampled group is representative of the whole student population at the national and regional level. However, also the province dimension (NUT5) was implicitly taken into account so that the final sample can be considered also representative at the province level (Invalsi, 2010). The number of units to be sampled within each region to ensure the sample representation was calculated on the basis of past SNV surveys using the Neyman procedure which is able to generate a sample size in such a way

that both the dimension and the variability of the phenomenon under study are correctly mirrored in the sampled units (Invalsi, 2010). Sampled schools could not refuse to receive the inspectors and were informed just a couple of weeks before the test was taken.

To test the effective goodness of these subsamples, we repeat the same analysis as in Table 2 using two other subsamples which are defined according to whether a school is a monitored school or not. Thus, the group of students in monitored schools contains the subgroup of the treated, but is larger because it also contains students in non-monitored classrooms of a sampled school.

<< Table B.1 here >>

Results are shown in Table B.1. Although now the group of the students in monitored schools is much larger (more than 20% of the population) the t-test for the comparison of the means are statistically significant for almost all the observable characteristics we observe in our dataset. We take this piece of evidence as a further confirmation about the goodness of representation of the two subsamples given by 'monitored' and 'non-monitored' classrooms. We can conclude that the treatment randomly splits the students population of each grade into two equally representative subgroups. Finally, notice that the same analysis performed on elementary schools data used in the robustness checks (2nd and 5th grade students) confirm the same results.

References

Algan, Yann, Pierre Cahuc and Andrei Shleifer (2011), "Teaching Practices and Social Capital", IZA DP. No. 6052.

Ammermüller, Andreas and Pischke Jörn-Steffen (2009), "Peer Effects in European Primary Schools: Evidence from PIRLS", Journal of Labour Economics, 27 (3). pp. 315-348.

Anderman, Eric M. and Tamera B. Murdock (2007), "The psychology of academic cheating", Academic Press Inc.

Bertoni, Marco, Brunello, Giorgio and Rocco, Lorenzo (2012), "When the Cat Is Near, the Mice Won't Play: The Effect of External Examiners in Italian Schools", IZA DP. 6629, Institute for the Study of Labor (IZA).

Bramoullé, Yann, Habiba Djebbari, Bernard Fortin (2009), "Identification of peer effects through social networks", Journal of Econometrics 150 (2009) 41:55.

Brock,W. and Durlauf, S. (2001a), "Discrete Choice with Social Interaction", Review of Economic Studies, 68, 235-260.

Brock, W. and S. Durlauf. (2001b), "Interactions-Based Models", in Handbook of Econometrics 5, J. Heckman and E. Leamer eds., Elsevier, 3297-3380.

Calvo-Armengol, Antoni, Eleonora Patacchini, Yves Zenou (2009), "Peer Effects and Social Networks in Education", Review of Economic Studies 76:4, 1239-1267.

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

Carrel, S. E., Fullerton, R., and West J. E. (2009), "Does your cohort matter? Measuring peer effects in college achievement", Journal of Labour Economics, 27(3): 439-464.

Castellano R., Longobardi S., Quintano C. (2009), "A fuzzy clustering approach to improve the accuracy of Italian student data. An experimental procedure to correct the impact of outliers on assessment test scores", StatisticaApplicazioni, Vol. VII, July-December 2009.

Cizek, G. J. (2003), "Detecting and preventing classroom cheating: Promoting integrity in assessment", Thousand Oaks, CA: Corwin Press.

Clark A.E. and Y. Loheac (2007), "It wasn't me, it was them!" Social influence in risky behavior by adolescents", Journal of Health Economics, 26 (2007), pp. 763–784.

Cooley, Jane (2010a), "Desegregation and the Achievement Gap: Do Diverse Peers Help?", University of Wisconsin-Madison, Economics Department, working paper.

Cooley, Jane (2010b), "Can Achievement Peer Effect Estimates Inform Policy? A View from Inside the Black Box", University of Wisconsin-Madison, Economics Department, working paper.

Cutler, David M. and Edward L. Glaeser (2010), "Social Interactions and Smoking", NBER Chapters, in: Research Findings in the Economics of Aging, pages 123-141 National Bureau of Economic Research, Inc.

Davezies, Laurent, Xavier D'Haultfoeuille and Denis Fougère (2009), "Identification of peer effects using group size variation", Econometrics Journal, volume 12, pp. 397–413.

Davidson, R. and MacKinnon, J. (2004), "Econometric theory and methods", Oxford University press, New York.

Davis, Stephen F., Patrick F. Drinan and Tricia Bertram Gallant (2009), "Cheating in school", Wiley-Blackwell, U.K..

De Giorgi Giacomo, Michele Pellizzari and Silvia Redaelli (2010), "Identification of Social Interactions through Partially Overlapping Peer Groups", American Economic Journal: Applied Economics, 2(2): 241-75.

De Giorgi, Giacomo and Michele Pellizzari (2011), "Understanding Social Interactions: Evidence from the Classroom", IZA DP No. 5624.

Dee, Thomas S. and Brian A. Jacob (2012), "Rational Ignorance in Education: A Field Experiment in Student Plagiarism", Journal of Human Resources, University of Wisconsin Press, vol. 47(2), pp. 397-434.

Drago, Francesco and Roberto Galbiati (2012), "Indirect Effects of a Policy Altering Criminal Behaviour: Evidence from the Italian Prison Experiment", American Economic Journal: Applied Economics, 4(2): 199-218.

Durlauf, S., and H. Tanaka (2008). "Regression Versus Variance Tests for Social Interactions." Economic Inquiry, 46(1), 25-28.

Durlauf, Steven N. and Ethan Cohen-Cole (2004), "Social interactions models", in "The Encyclopedia of Social Measurement", K. Lempf-Leonard eds., Academic Press.

Entorf, Horst and Martina Lauk (2006), "Peer Effects, Social Multipliers and Migrants at School: An International Comparison", IZA DP No. 2182.

Entorf, Horst and Tatsi Eirini (2009), "Migrants at School: Educational Inequality and Social Interaction in the UK and Germany", IZA DP No. 4175.

Epple D. and Richard D. Romano (2011), "Peer effects in education: a survey of the theory and evidence", Handbook of Social Economics, Vol. 1B, Ch. 20.

Eurydice (2009), "National testing of pupils in Europe: objectives, organization and use of the results", Education, Audiovisual and Culture Executive Agency (European Commission).

Ferrer-Esteban, G. (2012), "Cheating to the test in the Italian standardized assessment system: rationale and incentives", FGA Working Paper, (in press). Giovanni Agnelli Foundation (Turin).

Figlio, D. (2007), "Boys Named Sue: Disruptive Children and Their Peers", Education Finance and Policy Vol. 2, No. 4, Pages 376-394.

Fortin B., Lacroix G. and Villeval M. C. (2007), "Tax evasion and social interactions", Journal of Public Economics, 91: 2089–2112

Foster, Gigi (2006), "It's not your peers and it's not your friends: Some progress towards understanding the educational peer effect mechanism", Journal of Public Economics 90, 1455 − 1475.

Galbiati, Roberto and Zanella, Giulio (2012), "The tax evasion social multiplier. Evidence from Italy", Journal of Public Economics, vol. 96(5): 485-494.

Gaviria, A. and Raphael S. (2001), "School-Based Peer Effects and Juvenile Behavior", The Review of Economics and Statistics, Vol. 83, No. 2, Pages 257-268.

Glaeser, E., B. Sacerdote, and J. Scheinkman (1996), "Crime and Social Interactions", The Quarterly Journal of Economics, 111, 507–548.

Glaeser, E., B. Sacerdote, and J. Scheinkman (2003), "The Social Multiplier", Journal of the European Economic Association, 1, 2-3, 345-353.

Graham, B. (2008). "Identifying Social Interactions through Conditional Variance Restrictions." Econometrica, 76(3), 643-660

Grimes, Paul W. and Jon P. Rezek (2005), "The Determinants of Cheating by High School Economics Students: A Comparative Study of Academic Dishonesty in the Transitional Economies", International Review of Economics Education, volume 4, issue 2, pp. 23-45.

Hanushek, Eric A., John F. Kain and Steven G. Rivkin (2009), "New Evidence About Brown V. Board of Education: The Complex Effects of School Racial Composition on Achievement", Journal of Labor Economics, 2009, vol. 27, no. 3.

Hanushek, Eric A., John F. Kain, Jacob M. Martin and Steven G. Rivkin (2003), "Does peer ability affect student achievement?", Journal of Applied Econometrics, 2003, 18: 527-544.

Imberman, S. A., Kugler, A. and Sacerdote B. I. (2012), "Katrina's children: evidence on the structure of peer effcts from hurricane evaquees", American Economic Review, 102(5), 2048-2082.

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

Jacob B. and S. D. Levitt (2003b), "Catching Cheating Teachers: The Results of an Unusual Experiment in Implementing Theory", NBER Working Paper No. 9414.

Jacob Brian A. (2005), "Accountability, incentives and behavior: the impact of high-stakes testing in the Chicago Public Schools", Journal of Public Economics, 89:761-796.

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

Kleven, Henrik Jacobsen , Knudsen, Martin B., Thustrup Kreiner, Claus, Pedersen, Søren, and Saez, Emmanuel, (2011), "Unwilling or unable to cheat? Evidence from a tax audit experiment in Denmark", Econometrica, Vol. 79, No. 3, 651–692.

Lavy, Victor (2009), "Performance Pay and Teachers' Effort, Productivity, and Grading Ethics", American Economic Review, Vol. 99, No. 5: Pages 1979-2011.

Lavy, Victor and Analia Schlosser (2011), "Mechanisms and Impacts of Gender Peer Effects at School", American Economic Journal: Applied Economics", vol. 3(2), pages 1-33, April.

Lavy, Victor, Silva Olmo and Weinhardt Felix (2012), "The good, the bad and the average: evidence on the scale and the nature of ability peer effects at school", Journal of Labor Economics, Vol. 30, No. 2, pp. 367-414.

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

Lyle, David S. (2007), "Estimating and Interpreting Peer and Role Model Effects from Randomly Assigned Social Groups at West Point", Review of Economics and Statistics 89(2), 289 – 299.

Manski, Charles F. (1993) "Identification of Endogenous Social Effects: The Reflection Problem", Review of Economic Studies 60, 531 -542.

Maurin, Eric and Julie Moschion (2009), "The Social Multiplier and Labor Market Participation of Mothers", American Economic Journal: Applied Economics 2009, 1:1, 251–272.

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

McCabe, Donald L., and Linda Klebe Trevino (1993), "Academic Dishonesty: Honour Codes and Other Contextual Influences", Journal of Higher Education 64(5):522–38.

McCabe, Donald L., and Linda Klebe Trevino (1997), "Individual and contextual influences on academic dishonesty: a multicampus investigation", Research in Higher Education, 38(3): 379-376.

Miller, A., Murdock, T., Anderman, E. M., and Poindexter, A. L., (2007), "Who are all these cheaters? Characteristics of academically dishonest students". In E. M. Anderman and T. B. Murdock (Eds.), "Psychology of Academic Cheating", Elsevier, Boston.

Murdock, T. B., and Anderman, E. M. (2006), "Motivational perspectives on student cheating: Towards an integrated model of academic dishonesty", Educational Psychologist, 41(3), 129–145.

Murdock, T. B., Hale, N. M., and Weber, M. J. (2001), "Predictors of cheating among early adolescents: Academic and social motivations", Contemporary Educational Psychology, 26(1), 96-115.

Patacchini, E. and Y. Zenou (2012), "Juvenile delinquency and conformism", *Journal of the Law, Economics, and Organization*, 28, 1-13.

Powell B., Manish G.P. and Nair M. (2010), "Corruption, crime and economic growth", in "Handbook on the Economics of Crime", Benson B. L. and Zimmerman P. R. eds., Edward Elgar Publishing, Northampton.

Sacerdote Bruce (2010), "Peer effects in education: how might they work, how big are they, and how much do we know so far?", in "Handbook of the Economics of Education", Volume 3, Ch. 4, Eric A. Hanushek, Stephen Machin, Ludger Woessmann Eds.

Sacerdote, Bruce (2001) "Peer Effects with Random Assignment: Results for Dartmouth Roommates" Quarterly Journal of Economics, 681-704.

Shepard E. M and Blackely P. R. (2010), "Economics of crime and drugs: prohibition and public policies for illicit drug control", in "Handbook on the Economics of Crime", Benson B. L. and Zimmerman P. R. eds., Edward Elgar Publishing, Northampton.

Stanard, Charles I., and William J. Bowers (1970), "The College Fraternity as an Opportunity Structure for Meeting Academic Demands", Social Problems 17(3):371–90.

Stinebrickner, Todd R. and Ralph Stinebrickner (2006) "What Can Be Learned About Peer Effects Using College Roommates? Evidence from New Survey Data and Students from Disadvantaged Backgrounds", Journal of Public Economics, vol. 6(8-9):1435-1454.

Stock James and Yogo Motohiro, (2005), "Testing for Weak Instruments in Linear IV Regression", In "Identification and Inference for Econometric Models Essays in Honor of Thomas Rothenberg" Edited by Donald W. K. Andrews and James H. Stock, Cambridge University Press.

U.S. 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". U.S. Department of Education, Office of Elementary and Secondary Education, Washington, D.C. 20202. Available at: http://www2.ed.gov/policy/elsec/guid/saaprguidance.doc

Figures

Figure 1. Students stress while taking the test. Comparison between monitored and non-monitored students' answers to motivational questions (1=totally disagree; 2=partially disagree; 3=partially agree; 4=totally agree).
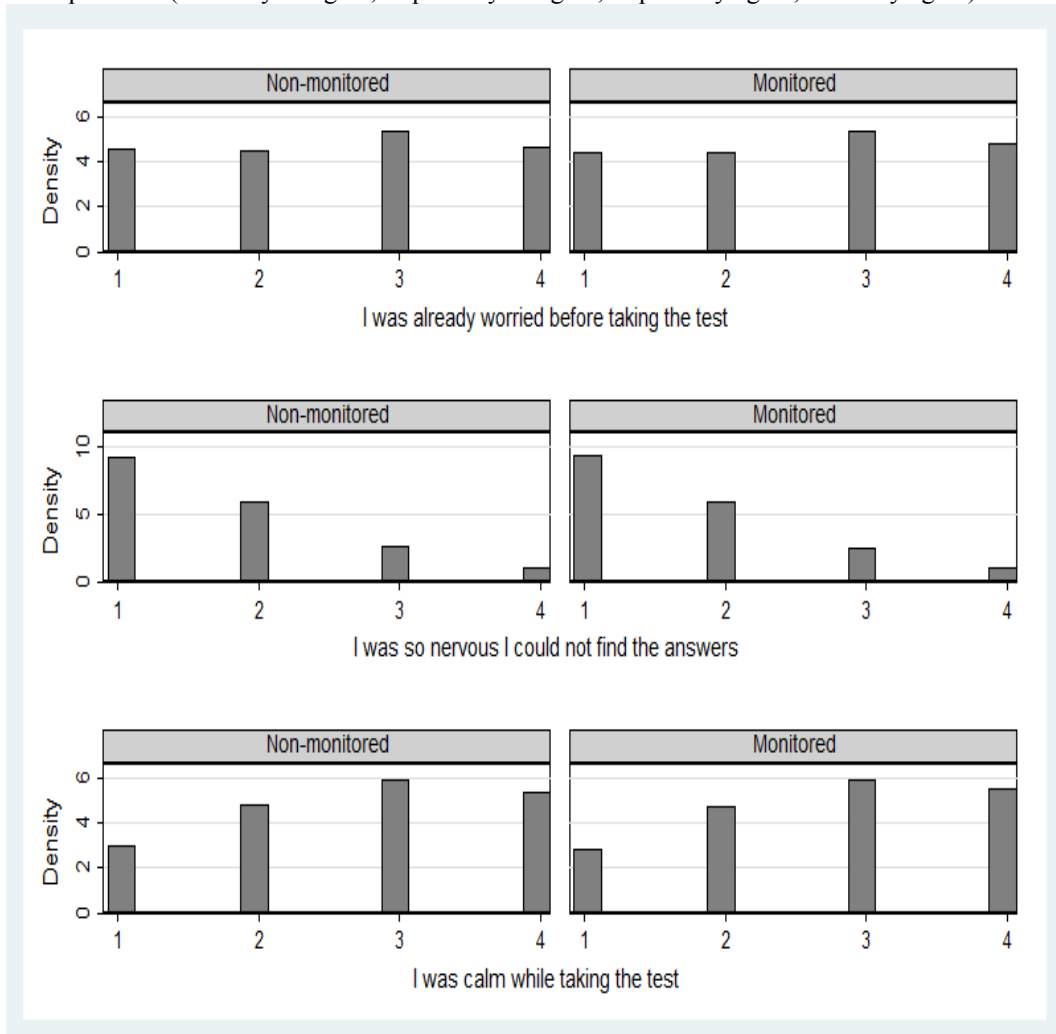
Notes. Students are asked whether they totally/partially agree/disagree with the following statements: 'I was already worried before taking the tests' (top); 'I was so nervous I could not find the answers' (center); 'I was calm while taking the test' (bottom).

Tables

Table 1. Descriptive statistics.

% Sampled Schools	22.48
% Monitored classes	7.78
% Monitored students (*)	8.01
% Non-monitored class in sampled school	13.07
% Absent students	0.71
Average school size	131.75
Average class size	20.58
Average cheating coefficient in non-monitored classrooms: Math	0.97
Average cheating coefficient in non-monitored classrooms: Language	0.91
Total no. schools	5,824
Total no. classrooms	26,707
Total no. students	522,655

Notes. (*) the percentage of 'monitored students' is calculated over the total number of students excluding absents. A student is considered 'absent' if he/she does not sit either Math or Language test, or both. Average class and school size refer to the average number of students in the class or school; the total no. of classrooms includes 25 classrooms with missing values in test scores results which are excluded from the empirical analysis. **Source**: SNV Invalsi 2009-10, 6th grade.

Table 2. Mean comparison of observable individual characteristics.

	Monitored students	Non-monitored students	Δ	Missing (% over total)
Female	48.3	48.34	0.04	1.3
Retained	7.31	7.09	-0.22	1.47
Immigrant	10.26	9.94	-0.32**	1.68
First gen. immigrants	6.59	6.54	-0.05	3.47
Second gen. immigrants	4	3.68	-0.32**	6.32
Kindergarten attendance	96.83	96.82	-0.01	22.6
Speak dialect at home	16.93	17.08	0.15	5.13
N (% over total)	41,550 (8.01)	477,395 (91.99)		

Notes. Absent students are excluded: a student is considered 'absent' if he/she does not sit either Math or Language test, or both. Δ indicates the difference between mean characteristics in the two groups; asterisks indicate whether the difference is statistically significant at 0.01 (***), 0.05 (**), 0.1 (*) significance levels. **Source**: SNV Invalsi 2009-10, 6th grade.

Table 3. Descriptive statistics: test scores mean, median and variance decomposition.

	Language			Math		
	All Pop.	Monitored	Non-monitored	All Pop.	Monitored	Non-monitored
Mean	61.44	61.39	61.45	51.95	51.42	51.99
Median	63.79	63.80	63.79	50.00	50.00	50.00
Total Var.	232.21	235.75	231.91	329.03	329.78	328.94
Var. Between Classrooms	48.79	42.88	49.30	76.80	69.44	77.41
Var. Within Classrooms	183.43	192.86	182.60	252.24	260.34	251.53

Notes. The formula is corrected with appropriate weights to take into account the different size of the subgroups (i.e. classrooms) (Ammermüller and Pischke, 2009, p.323). **Source:** SNV Invalsi 2009-10, 6th grade.

Table 4. Baseline estimates of the social multiplier.

	MATH					
γ^2	5.135	5.136	5.889	5.390	5.926	5.401
	(0.211)	(0.211)	(0.301)	(0.244)	(0.291)	(0.240)
P-value (H$_0$: γ^2=1)	0.00	0.00	0.00	0.00	0.00	0.00
Model Parameters						
γ	2.266	2.266	2.427	2.322	2.434	2.324
	(0.047)	(0.047)	(0.062)	(0.053)	(0.060)	(0.052)
J	0.559	0.559	0.588	0.569	0.589	0.570
	(0.009)	(0.009)	(0.011)	(0.010)	(0.010)	(0.010)
First Stage F-Statistic	10772.44	10772.02	3231.63	6161.20	3820.05	6838.82
	LANGUAGE					
γ^2	4.189	4.182	4.713	4.370	4.774	4.383
	(0.169)	(0.169)	(0.241)	(0.198)	(0.234)	(0.195)
P-value (H$_0$: γ^2=1)	0.00	0.00	0.00	0.00	0.00	0.00
Model Parameters						
γ	2.047	2.045	2.171	2.090	2.185	2.094
	(0.041)	(0.041)	(0.056)	(0.047)	(0.053)	(0.047)
J	0.511	0.511	0.539	0.522	0.542	0.522
	(0.010)	(0.010)	(0.012)	(0.011)	(0.011)	(0.011)
First Stage F-Stat	8290.73	8290.42	2868.01	5172.91	3315.70	5641.34
No. Classrooms	25959	25959	25959	25959	25959	25959
Additional controls (Ψ_c)						
Non-monitored class in sampled school		yes			yes	yes
High immigrant share (>P75)			yes		yes	
High immigrant share (>P90)				yes		yes

Notes. Classes with less than 10 students are dropped from the sample. Additional controls (Ψ_c) include the dummy for non-monitored classrooms in sampled school and the dummy for high share of immigrants, respectively, for immigrant class shares greater than P75 or P90. **Source:** SNV Invalsi 2009-10, 6th grade.

Table 5. Heterogeneous effects in the cheating social multiplier.

PANEL A	MATH			
	High heterogeneity	Low heterogeneity	High heterogeneity	Low heterogeneity
	Books at home		*Outside-school activities*	
γ^2	4.9199	5.8522	4.7388	6.1082
	(0.3203)	(0.3781)	(0.2932)	(0.4092)
First Stage F-Statistic	2876.05	3635.15	3270.83	3187.27
P-value H0: $\gamma^2_H=\gamma^2_L$	0.06		0.01	
No. Classrooms	12031	12107	12040	12098
	Play with friends		*Sport practice*	
γ^2	5.4594	5.2770	5.5002	5.0890
	(0.4135)	(0.3044)	(0.3907)	(0.3151)
First Stage F-Statistic	2633.23	3811.77	2517.54	4066.95
P-value H0: $\gamma \gamma^2_H=\gamma^2_L$	0.72		0.41	
No. Classrooms	11897	12241	11903	12235
PANEL B	LANGUAGE			
	High heterogeneity	Low heterogeneity	High heterogeneity	Low heterogeneity
	Books at home		*Outside-school activities*	
γ^2	4.0339	4.7029	3.8383	4.9494
	(0.2606)	(0.2817)	(0.2410)	(0.3025)
First Stage F-Statistic	2679.83	2875.71	2881.15	2682.26
P-value H0: $\gamma^2_H=\gamma^2_L$	0.08		0.00	
No. Classrooms	12031	12107	12040	12098
	Play with friends		*Sport practice*	
γ^2	4.4492	4.2456	3.8128	4.6742
	(0.3320)	(0.2303)	(0.2577)	(0.2737)
First Stage F-Statistic	2133.98	3435.55	2244.61	3547.14
P-value H0: $\gamma^2_H=\gamma^2_L$	0.61		0.02	
No. Classrooms	11897	12241	11903	12235
Additional controls (Ψ_c)	yes	yes	yes	yes

Notes. Classes with less than 10 students and with missing values in the relevant variables are dropped from the sample. Additional controls (Ψ_c) include the dummy for non-monitored classrooms in sampled school and the dummy for high share of immigrants (immigrant class share greater than P90). **Source:** SNV Invalsi 2009-10, 6th grade.

Table 6. Robustness checks. Descriptive statistics elementary school students.

	LANGUAGE			MATH		
	All Pop.	Monitored	Non monitored	All Pop.	Monitored	Non monitored
5th grade						
Mean	70.23	67.54	70.44	64.76	61.89	65.38
Median	73.91	71.01	73.91	65.91	61.36	65.91
Var. Tot.	144.27	146.62	143.81	65.39	62.5	65.44
Var. Between Classes	45.90	35.03	46.46	27.05	19.7	27.45
Var. Within Classes	98.38	111.59	97.35	38.34	42.8	38
N (students)	475,343	34,554	440,789	475,343	34,554	440,789
2nd grade						
Mean	65.94	62.05	66.24	62.52	57.17	62.94
Median	69.23	65.38	69.23	60.71	57.14	64.28
Var. Tot.	34.85	35.5	34.71	30.81	27.2	30.89
Var. Between Classes	10.39	7.51	10.52	14.52	8.37	14.8
Var. Within Classes	24.46	27.99	24.18	16.29	18.83	16.09
N (students)	466,536	34,201	432,335	466,536	34,201	432,335

Source: SNV Invalsi 2009-10, 5th and 2nd grade.

Table 7. Robustness checks. Social multiplier estimates for elementary school students.

Panel A: 5th grade						
	MATH					
γ^2	7.482	7.471	7.812	7.562	8.027	7.590
	(0.365)	(0.364)	(0.513)	(0.408)	(0.484)	(0.399)
P-value H_0: $\gamma^2=1$	0.00	0.00	0.00	0.00	0.00	0.00
First Stage F-Statistic	7113.20	7112.93	2508.88	4840.38	3158.73	5393.57
	LANGUAGE					
γ^2	5.245	5.227	5.402	5.267	5.524	5.280
	(0.323)	(0.326)	(0.457)	(0.368)	(0.438)	(0.361)
P-value H_0: $\gamma^2=1$	0.00	0.00	0.00	0.00	0.00	0.00
First Stage F-Statistic	6504.80	6504.56	2429.27	4440.85	3004.18	4911.43
No. Classrooms	26942	26942	26942	26942	26942	26942
Panel B: 2nd grade						
	MATH					
γ^2	7.419	7.364	6.816	7.228	7.112	7.297
	(0.379)	(0.375)	(0.502)	(0.418)	(0.478)	(0.411)
P-value H_0: $\gamma^2=1$	0.00	0.00	0.00	0.00	0.00	0.00
First Stage F-Statistic	7867.09	7866.80	2987.77	5115.35	3557.85	5648.61
	LANGUAGE					
γ^2	4.437	4.385	4.246	4.273	4.348	4.296
	(0.201)	(0.198)	(0.269)	(0.219)	(0.257)	(0.215)
P-value H_0: $\gamma^2=1$	0.00	0.00	0.00	0.00	0.00	0.00
First Stage F-Statistic	8809.94	8809.61	3236.33	5608.41	3865.24	6218.09
No. Classrooms	26850	26850	26850	26850	26850	26850
Additional controls (Ψ_c)						
Non-monitored class in sampled school		yes			yes	yes
High immigrant share (>P75)			yes		yes	
High immigrant share (>P90)				yes		yes

Notes. Classes with less than 10 students are dropped from the sample. Additional controls (Ψ_c) include the dummy for non-monitored classrooms in sampled school and the dummy for high share of immigrants, respectively, for immigrant class shares greater than P75 or P90. **Source:** SNV Invalsi 2009-10, 5th and 2nd grade.

Table 8. Robustness checks. Achievement peer effects using an alternative instrument (class size).

PANEL A	Non-monitored classrooms			
	MATH		LANGUAGE	
γ_a^2	5.9350	3.2160	5.0901	2.0785
	(0.1059)	(0.4398)	(0.0976)	(0.3375)
P-value H_0: $\gamma_a^2=1$	0.00	0.00	0.00	0.00
First Stage F-Statistic	49602.35	3333.62	48037.98	3914.55
No. Classrooms	23901	23901	23901	23901
PANEL B	Monitored classrooms			
	MATH		LANGUAGE	
γ_a^2	5.3363	1.7149	4.2599	1.5793
	(0.2956)	(1.2544)	(0.2333)	(0.6559)
P-value H_0: $\gamma_a^2=1$	0.01	0.57	0.00	0.38
First Stage F-Statistic	5457.98	257.31	4767.00	309.41
No. Classrooms	2058	2058	2058	2058
Additional controls				
Class level variables (Ψ_c)	yes	yes	yes	yes
School-district fixed effects		yes		yes

Notes. Class level variables (Ψ_c) include the dummy for non-monitored classrooms in sampled school and the dummy for high share of immigrants (immigrant class share greater than P90). School-districts fixed effects correspond to 110 dummies. Classes with less than 10 students are dropped from the sample. **Source:** SNV Invalsi 2009-10, 6th grade.

Table 9. Robustness checks. Baseline results including classrooms with less than 10 students.

	MATH					
γ^2	5.172	5.173	5.965	5.447	5.994	5.454
	(0.210)	(0.210)	(0.305)	(0.245)	(0.293)	(0.241)
P-value ($H_0: \gamma^2=1$)	0.00	0.00	0.00	0.00	0.00	0.00
Model Parameters						
γ	2.274	2.274	2.442	2.334	2.448	2.335
	(0.046)	(0.046)	(0.062)	(0.053)	(0.060)	(0.052)
J	0.560	0.560	0.591	0.572	0.592	0.572
	(0.009)	(0.009)	(0.010)	(0.010)	(0.010)	(0.009)
First Stage F-Statistic	8267.11	8266.80	2538.80	4739.77	2960.45	5200.40
	LANGUAGE					
γ^2	4.272	4.265	4.819	4.456	4.879	4.468
	(0.175)	(0.176)	(0.255)	(0.208)	(0.247)	(0.205)
P-value ($H_0: \gamma^2=1$)	0.00	0.00	0.00	0.00	0.00	0.00
Model Parameters						
γ	2.067	2.065	2.195	2.111	2.209	2.114
	(0.042)	(0.043)	(0.058)	(0.049)	(0.056)	(0.049)
J	0.516	0.516	0.544	0.526	0.547	0.527
	(0.010)	(0.010)	(0.012)	(0.011)	(0.011)	(0.011)
First Stage F-Stat	7353.47	7353.20	2519.57	4558.58	2901.45	4956.61
No. Classrooms	26682	26682	26682	26628	26628	26628
Additional controls (Ψ_c)						
Non-monitored class in sampled school		yes			yes	yes
High immigrant share (>P75)			yes		yes	
High immigrant share (>P90)				yes		yes

Notes. Additional controls (Ψ_c) include the dummy for non-monitored classrooms in sampled school and the dummy for high share of immigrants, respectively, for immigrant class shares greater than P75 or P90. **Source:** SNV Invalsi 2009-10, 6th grade.

Table B.1. Mean comparison between students in sampled and non-sampled schools.

	Students in sampled schools	Students in non-sampled schools	Δ
Female	48.35	48.34	-0.01
Retained	7.27	7.07	-0.2**
Immigrant	10.36	9.85	-0.51***
First gen. immigrants	6.89	6.46	-0.43***
Second gen immigrants	3.83	3.67	-0.16***
Kindergarten attendance	96.61	96.87	0.26***
Speak dialect at home	16.06	17.35	1.29***
N (% over total)	111,497 (21.48)	407,448 (78.52)	

Notes. Absent students are excluded: a student is considered 'absent' if he/she does not sit either Math or Language test, or both. Δ indicates the difference between mean characteristics in the two groups; asterisks indicate whether the difference is statistically significant at 0.01 (***), 0.05 (**), 0.1 (*) confidence levels. **Source**: SNV Invalsi 2009-10, 6th grade.