Cheating and Incentives: Learning from a Policy Experiment

Cesar Martinelli, Susan W. Parker, Ana Cristina Pérez-Gea and Rodimiro Rodrigo

October 2015

Discussion Paper

Interdisciplinary Center for Economic Science
4400 University Drive, MSN 1B2, Fairfax, VA 22030
Tel: +1-703-993-4719 Fax: +1-703-993-4851
ICES Website: http://ices.gmu.edu
ICES RePEc Archive Online at: http://edirc.repec.org/data/icgmuus.html

CHEATING AND INCENTIVES: LEARNING FROM A POLICY EXPERIMENT[1]

CÉSAR MARTINELLI, SUSAN W. PARKER, ANA CRISTINA PÉREZ-GEA, AND RODIMIRO RODRIGO

ABSTRACT. We use a database generated by a policy intervention that incentivized learning as measured by standardized exams to investigate empirically the relationship between cheating by students and cash incentives to students and teachers. We adapt methods from the education measurement literature to calculate the extent of cheating, and show that cheating is more prevalent under treatments that provide monetary incentives to students (versus no incentives, or incentives only to teachers), both in the sense of a larger number of cheating students per classroom and in the sense of more cheating relations per classroom. We also provide evidence of learning to cheat, with both the number of cheating students per classroom and the average number of cheating relations increasing over the years under treatments that provide monetary incentives to students.

Date: February 26, 2015.

[1]The authors are grateful for financial support from Asociación Mexicana de Cultura. Martinelli: George Mason University, cmarti33@gmu.edu. Parker: CIDE, susan.parker@cide.edu. Pérez-Gea: Instituto Tecnológico Autónomo de México, acperezgea@gmail.com. Rodrigo: Secretaría de Hacienda y Crédito Público, México, rodimiro.rodrigo@hacienda.gob.mx. The authors alone are responsible for all contents of this paper.

1. INTRODUCTION

Anecdotal evidence, available to most anyone who has taken, or administered, written exams, indicates that cheating is common. This view is confirmed by self-reported evidence (Cizek 1999); in fact, an important body of education literature is devoted to the statistical detection of cheating in multiple choice exams.[1] There is, however, surprisingly little empirical analysis of the effects of incentives on cheating. Jacob and Levitt (2003) have documented, using data from Chicago public schools, that cheating is responsive to incentives provided to teachers. Jacob and Levitt took advantage of a policy reform by which Chicago schools were put on probation if not enough students performed at or above national levels in a standardized multiple choice achievement exam, with the subsequent danger that the school could be closed and the school staff dismissed or reassigned. Another piece of the policy reform was to require students to satisfy minimum standards on the same exam on reading and mathematics in order to be promoted to the next grade.

In this paper, we explore the effects of incentives to students and teachers on cheating using a data base generated by *Aligning Learning Incentives* [ALI], a policy intervention involving around 40,000 students that incentivized learning in mathematics in 88 high schools throughout Mexico. The Mexican intervention included explicit monetary incentives linked to performance in a multiple choice exam. Three treatment groups provided incentives to students alone, to teachers alone, and to students, teachers and school administrators. Though the policy intervention proved to be very successful in increasing performance, there was some evidence of cheating in the incentivized exams, as described by Behrman et al. (2014). Incentivized exams were monitored by staff from the Mexican Secretariat of Public Education, not by teachers. This feature is unlike the Chicago reform, where exams were left to be monitored by the incentivized teachers. Correspondingly, our evidence suggests a focus on the students as the main agents in breaking the rules. We analyze the extent of cheating in the incentivized exams, the impact of monetary incentives and other variables on cheating, and the evolution of cheating over the duration of the program.

To guide the analysis, we propose a simple model of cheating and incentives. We approach the decision to cheat as that of an illicit communication between two students, a *copier* and a *source*, in the course of an exam. Provided an opportunity for communication arises, the incentive for the copier to communicate is proportional to the benefits attached to the exam score and the expected increase in the score. The expected increase in score, in turn, is related to the knowledge or ability of the source. Acquiescence from the source student may be obtained via side payments, rewards in social status, or implicit threats. The cost of communication may vary with personal characteristics of the copier and with random events such as how carefully the test is being supervised, the physical distance between the copier and the source, etc. Thinking of illicit communication as a *directed link* and of the students as *nodes*, cheating in a classroom can then be described as a *directed network*. Each active component of this communication network consists of a source and one or several copiers. When there are substantial incentives linked to the exam score, potential copiers may seek better students as sources, so that several copiers may try to communicate with the same source, leading to larger components in the cheating network.

We start our empirical analysis by identifying cheating students extending methods borrowed from the education measurement literature. These methods rest on the statistical detection of pairs

[1] On the pervasiveness of copying in high school and college, see also Davis et al. (1992), Davis and Ludvigson (1995), Brandes (1986) and Schab (1991).

of students whose response patterns are unusually similar.[2] Ideally, these methods are designed with the aim of testing whether a particular pair of students on which there may be some suspicion have, in fact, engaged in illicit communication. In our setting, since we test for illicit communication every possible pair of students in each classroom, the probability of accusing falsely of cheating any student is much larger than that of accusing any given pair. To classify students into cheaters and non cheaters, we exploit the fact that the same exam was administered across classrooms, whilst cheating is likely confined to pairs of students in the same classroom. We test for cheating every pair of students in a school, and raise the threshold for accusing a given pair to the point where only 10% of students are accused of cheating because of an unusual similarity with a student in a different classroom. We then use this threshold to classify all pairs of students within each classroom as cheating or non cheating pairs, and identify a student as a cheater if the student belongs to a cheating pair.

The statistical classification confirms that the fraction of students involved in cheating is larger in the treatments that provide incentives to students than in those that provide incentives to teachers alone. For the control (i.e. no incentives) group, cheating ranges from 5% to 7.5% in different years. For the student incentives group, cheating jumps from about 11% the first year to about 27% the second year and an astonishing 30% the third year. For the teacher and student incentives group, cheating jumps from about 7% the first year to about 23.5% the second year and an equally astonishing 32% the third year. For the teacher only incentives group, per contra, cheating increases from about 7% the first year to nearly 10% the third year; statistically different than cheating in the control group but not very different in magnitude. Cheating is not only more widespread in incentivized settings, but also more intense. In particular, illicit communication networks identified in the data contain not only more active students but also are more densely connected. There is a large variance in the prevalence of cheating across schools and classrooms, with a few schools having a large percentage of cheaters in every or almost every classroom.

In recent years there has been a growing reliance on standardized testing to evaluate performance of different education institutions and to introduce accountability in public education. The No Child Left Behind Act of 2001, for instance, provides support for standards-based education reform in the United States. The introduction of the National Evaluation of Achievement in Schools [ENLACE] exams in Mexico in 2006 pursues similar measurement and accountability goals. Concomitantly, there has a growing interest in incentives programs that include incentives to students and teachers linked to performance in standardized tests. Recent teacher incentives program include Muralidharan and Sundararaman (2011) carried out in rural India, Glewwe, Ilias and Kremer (2010) in rural Kenya, and Springer et. al. (2010) in Nashville, Tennessee public schools. Recent examples of student incentives programs include Angrist and Lavy (2009) study of high school student incentives in Israel, Kremer, Miguel and Thornton (2009) of 6th grade girls in Kenya, and Fryer (2011) report on four different field experiments in Chicago, Dallas, New York City and Washington, DC. None of these studies analyzes the incidence of student cheating or if incentives resulted in an increase in student (or teacher) cheating.

To our knowledge, ours is the first research effort to analyze the incidence of student cheating in standardized testing in reaction to monetary incentives. In particular, our study is different from Jacob and Levitt's (2003) pioneer work in that the intervention we focus on provided explicit monetary incentives, and avoided cheating by teachers by employing other monitors for incentivized exams. Behrman et al. (2014), who report on the ALI intervention effects on learning, are careful

[2]See e.g. Wollack 1997, 2003 and 2006, Wesolowsky 2000, van der Linden and Sotaridona 2006, Romero et al. 2012, Zopluoglu and Davenport 2012.

to isolate the effects of cheating on inflating test scores, but do not elaborate on the determinants of cheating, the characteristics of cheating networks, or the evolution of cheating over time.

Other recent work on cheating has focused on peer effects and the use of external monitors. Carrell et al. (2008) use self-reported data from US military academies to show that peer honesty (as measured by the presence of high school cheaters in the classroom) result in a substantial increase in the probability that student will cheat. They interpret this social effect as an evolving social norm of toleration, which may be a mechanism operating behind the evolution of cheating networks in the policy intervention we study. Lucifora and Tonello (2012) use a dataset drawn from a national evaluation standardized test in Italy that including random monitoring by external inspectors to show that grades are inflated in the absence of such inspectors.[3]

Our results illustrate what has been dubbed (e.g. by Charness et al. 2013) "the dark side of incentives:" explicit rewards often have unintended consequences, as individuals attempt to game the system in ways that are sometimes detrimental to the objectives pursued by the rewards. Policy interventions that rely on incentivize exams should pay attention to these unintended consequences, and analyze the impact of incentives on gaming attempts.

We believe that cheating in an incentivized exam illustrates the tension between material incentives and ethical and social considerations. With few students in a classroom engaging in cheating, cheating may be an activity subject to stigma; as cheating becomes widespread, it may lose any such negative connotation (Benabou and Tirole 2006, 2011). In this sense, cheating in the classroom resembles illegal activities in the society at large. Glaeser, Sacerdote and Scheinkman's (1996) classical work on social interactions and crime interprets the high variance of crime rates across US cities as evidence of the importance of social interactions in crime, leading to interdependent decisions. In this line, Calvó-Armengol and Zenou (2004) and Ballester, Calvó-Armengol and Zenou (2010) offer models of crime decisions in which criminals benefit from friendship links, modeled explicitly as a network. They consider positive externalities at the local level, stemming from shared knowledge, but also competition between criminals at the aggregate level. In our setting, per contra, both local and aggregate interactions may have helped the spreading of cheating in incentivized settings.

The remainder of this paper is organized as follows. Section 2 provides a network model of cheating in a classroom. Section 3 describes the policy intervention on which our exploration is based, explains the statistical methods employed to detect cheating, and provides an estimate of the extent of cheating during the policy intervention for the different treatments. Section 4 explores empirically cheating at the student level, focusing on the impact of incentives and experience on the probability of cheating, and on the effects of cheating in exam performance. Section 5 explores empirically cheating at the classroom level, focusing on the percentage of cheaters and the characteristics of the cheating network for each treatment. Section 6 concludes.

2. A MODEL OF CHEATING AND INCENTIVES

2.1. **The decision to copy.** We propose in this section a simple model as a guide to the empirical analysis of cheating at the classroom level. Consider a classroom with N students, $i = 1,\dots,N$, who are about to take an exam. The exam has a reward for each student equal to b times the score obtained by the student in the exam, where $b > 0$ represents explicit monetary benefits, as those

[3]Figlio and Winicki (2005) show another gaming effect of school accountability based on high-stakes testing: school districts under an accountability system in the state of Virginia reacted by substantially increasing calories in their menus on testing days, apparently with some success in raising standardized test scores.

provided by the incentivized treatments, as well as the implicit satisfaction or amenity value of performing well in a test. Student i's expected score in the exam, if the student does not copy anybody else, is given by s_i. This score is supposed to reflect the ability and preparation of the student and is known to all students in the classroom. If student i copies another student j, then the expected score for student i is $\max\{s_i, s_j\}$, and the student must pay a communication cost $c_{ij} > 0$. This cost represents the effort and pain in illicitly communicating with another student, and may vary with the diligence of the person monitoring the exam, with the social norm of toleration for cheating in the classroom (and the school), with friendship or other affinity easing communication among the students involved, with the location and physical proximity of the students in the classroom, and with the cheating skills of the students involved. We assume the triangular inequality $c_{ij} + c_{jk} \geq c_{ik}$ for each triple of distinct students i, j, k, corresponding to the notion that mediated communication is feasible.

Note that we assume that costs and benefits of copying accrue only to the copier, that copying affects only the score of the student copying, and that a student can copy only from one source or from none. (Alternatively, we could introduce costs for the source student, and approach the strategic interaction between the students as a cooperative game, with side payments representing rewards in friendship or status.)

Under the preceding assumptions, a student i will find it optimal to copy if

$$\max_{j \neq i} \left\{ bs_j - c_{ij} \right\} \geq bs_i,$$

or equivalently if

$$\max_{j \neq i} \left\{ s_j - c_{ij}/b \right\} \geq s_i, \tag{1}$$

and will copy only from one of the students maximizing the expression on the left.

Because of the triangular inequality, a student will not find it optimal to have as a source another student who is in turn copying. It is possible, however, that several students copy from the same source. The following is immediate:

Proposition 1. *Increasing the benefits of the exam, while keeping fixed the expected scores and the structure of costs, (i) weakly increases the number of copiers in the classroom, and (ii) weakly increases the score of the source for each copier.*

The first part of the proposition follows from simple inspection of equation 1. The second relies on the fact that larger benefits make it attractive to incur larger communication costs.

2.2. **The cheating network.** We can represent cheating in the classroom as a *directed network*, in which each student in the classroom is a *node* and illicit communication is represented as a *directed link* from the copier to the source student.[4] A *path* is a sequence of links connecting a sequence of nodes (ignoring the direction of the links). A *component* of the network is a maximal subset of students in the classroom with the property that there is a path between each pair of students in the subset. Given our previous assumptions, each component of the cheating network is either an isolated student who is neither a copier nor a source, or is composed of one source and one or more copiers. That is, each component with active cheating has a star-shaped structure. We refer to copiers and sources as cheaters, and to a component with a single copier as a pair of cheaters.

[4]Definitions in this section are adapted from Jackson (2010).

Let us assume that there are no ties at the top score for the classroom. Denote N the student with the best expected score; that is, $s_N > s_i$ for all $i \neq N$, and define

$$\underline{b} = \min_{i,j:s_j>s_i} \left\{ \frac{c_{ij}}{s_j - s_i} \right\} \quad \text{and} \quad \overline{b} = \max_{i,j\neq N:s_j>s_i} \left\{ \frac{c_{i1} - c_{ij}}{s_1 - s_j} \right\}.$$

Suppose the ratio of communication costs to score differences is heterogenous across student pairs; then it follows from Proposition 1:

Proposition 2. *For $b \leq \underline{b}$, there is at most a pair of cheaters, and for $b \geq \overline{b}$, every student other than N either copies from N or does not copy.*

The following example illustrates the evolution of the cheating network in reaction to increased incentives from isolated pairs of cheaters to a unique component with active cheating. Suppose there are $N = 2n+1$ students, with expected scores

$$s_i = \begin{cases} 0 & \text{if } i \text{ is odd and different from } N, \\ 1 & \text{if } i \text{ is even}, \\ 2 & \text{if } i = N, \end{cases}$$

and copying costs

$$c_{ij} = \begin{cases} \varepsilon i & \text{if } i \text{ is odd and } j = i+1, \\ 1+\varepsilon i & \text{in all other cases.} \end{cases}$$

Intuitively, each student with a low expected grade has a "neighbor" with a middle expected grade. We have $\underline{b} = \varepsilon$ and $\overline{b} = 1$. If benefits are low ($b < 1$), students with low expected grades such that $i \leq b/\varepsilon$ prefer to copy from their neighbors, forming cheating pairs. If benefits are high ($b > 1$), students with low expected grades prefer to copy from the student with the top score even if this is costlier. Moreover, students with middle expected grades such that $i \leq (b-1)/\varepsilon$ also copy from the student with the highest expected grade. Copying in this case is a generalized affair.

Figure 1 depicts the evolution of the cheating network in the example as incentives linked to the exam increase. For low incentives, as in cases A and B, copying is a local affair occurring between students and their neighbors with better expected grades. For higher incentives, as in cases C, D, and E, students copy from a better if costlier source. The active components of the network are $\{1,2\}$ in case A, $\{1,2\}$ and $\{3,4\}$ in case B, $\{1,3,N\}$ in case C, $\{1,2,3,N\}$ in case D, and $\{1,2,3,4,N\}$ in case E. Note that when incentives are high enough in reaction to communication costs (case C), the best students are drafted into cheating, which may lead to a positive correlation between ability and cheating. For very high incentives (case E), such positive correlation is likely to get swamped by generalized cheating.

Like the cost of other illicit activities, part of the cost of copying may depend on social norms that stigmatize this behavior. These social norms may be eroded in a classroom if students who did not copy in an exam observe that other students were not so reticent. Thus, even if incentives are held constant, the cheating network may evolve over time in a pattern similar to that depicted by Figure 1.

A useful indicator for the level of activity in a network is the *density*, which is defined as the number of directed links in the network divided by the potential number of links, $N(N-1)$. From Proposition 1, the density of the cheating network is weakly increasing in the benefits associated to the exam. (The number of cheaters is not, as illustrated by going from case B to case C.) A commonly used indicator of activity at the individual node level is the *degree*, which is defined as the number of directed links stemming from a node; note that density is equal to the average

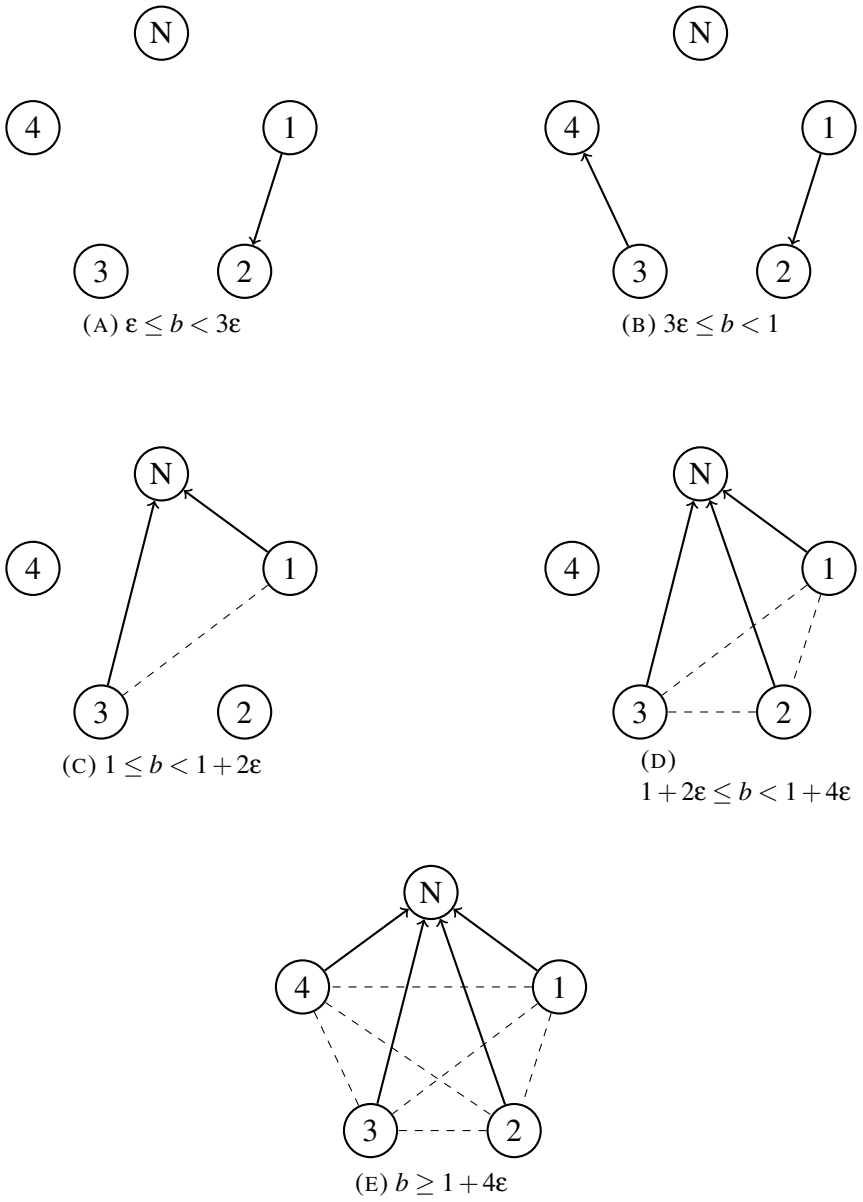

FIGURE 1. Cheating network for different incentive levels

degree divided by $N-1$. We return to both indicators later to describe likely illicit communication activity in our data.

3. INCIDENCE OF CHEATING IN THE ALI EXPERIMENT

3.1. **The ALI experiment.** The data used in this paper derive from the *Aligning Learning Incentives* [ALI] experiment carried out in Mexico, which began with the 2008-09 academic year and ended with the 2010-11 academic year (Behrman et al. 2014). A total of 88 high schools (*preparatorias*) participated in the experiment; Figure 2 illustrates the location of the schools. The schools

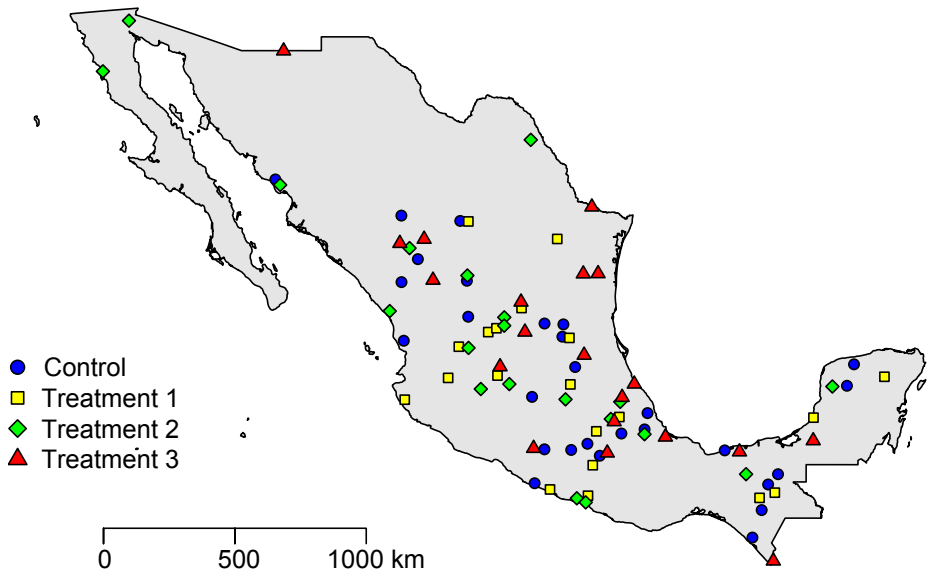

FIGURE 2. ALI schools

were randomly assigned to four different groups; 20 schools were assigned to each of three treatment schools, corresponding to different incentive schemes, and 28 schools were assigned to a control group with no incentives. Specifically, the four groups were:

(C) Control group: No payments.
(T1) Treatment group 1: Payments to students based on their own performance.
(T2) Treatment group 2: Payments to mathematics teachers based on the performance of the students in their classes.
(T3) Treatment group 3: Payments to students based on their own performance and on the performance of the other students in their class. Payments to mathematics teachers based on the performance of the students in their classes and on the performance of the students in all other mathematics classes. Payments to non-mathematics teachers and school administrators based on the performance of all of the students in the school.

Incentive payments were based on standardized curriculum-based mathematics exams in 10th, 11th and 12th grade given at the end of each academic year. Incentive payments were based on the amount of improvement in mathematics learning over the school year for 10th and 11th graders and on the final level of learning for 12th graders. The score on the 9th grade ENLACE, a Mexican national level exam in reading and mathematics skills, was used as the baseline for math achievement. For the purpose of determining incentive payments, performance on each exam was categorized, as in the 9th grade ENLACE, into four levels: Pre-Basic, Basic, Proficient, and Advanced. The exams were designed by CENEVAL (an independent and widely regarded Mexican education evaluation agency) based on the input of Mexican experts on high school mathematics. The monetary incentives for improving performance from one level to one or more levels above

fluctuated between 4,000 and 15,000 Mexican pesos (approximately 30 to 120 US dollars at the exchange rate at the time of the program); these are substantial incentives for Mexican high school students. The Appendix provides more details on the incentive payments of the ALI program.

Baseline and follow up questionnaires were applied to students and teachers at the beginning and at the end of each year. The student questionnaires provided (self reported) information on family background and personal characteristics. The incentivized exams were not administered or monitored by school personnel, but by representatives of the Secretariat of Public Education state offices, with one monitor assigned to each class and an overall supervisor assigned to the school. The same administrators collected the answer sheets and were required to account for all copies of the exams after administration of the exam to reduce the possibility of teaching to the test based on past tests.

3.2. **Statistical detection of cheating.** A number of statistical indices for the detection of answer copying in an exam have been developed by the education measurement literature, including the ω index (Wollack 1997), the Generalized Binomial Test [GBT] index (van der Linden and Sotaridona 2006), the K index (Holland 1996, Sotaridona and Meijer 2002), and the S1 and S2 indices (Sotaridona and Meijer 2003). There is, however, a relatively small literature comparing the performance of the different indices with real data. It is known that the indices perform better when tests have a larger number of questions and a larger sample (Wollack 2003 and 2006). We focus the analysis on the ω index and the GBT, which looked most promising given the recent education literature.[5]

Both the ω index and the GBT index use the similarity in both correct and incorrect answers for each ordered pair of students to assess whether a student copied from the other. Figure 3 illustrates the distribution of the number of exact matches for sampled pairs of students in the same classroom for each treatment in the 12th grade exam the last year of the program (2010-11). This is the cohort that went through the program all three years, and is the focus of our empirical analysis. The distribution of exact matches for the teacher incentive treatment is almost identical (mostly overlapping) to that for the control treatment. The distributions for the student and the teacher and student incentive treatments, however, are markedly different, with both of them exhibiting first order stochastic dominance over the distribution for the control. Note that the distribution of exact matches for the teacher and student incentive treatment, in particular, is bimodal, with the second mode at about 75 exact matches, out of 112 questions. This is not, in itself, evidence of more cheating. Part of it reflects increased achievement. A necessary building block for determining the extent of cheating is a model to determine the probability that an individual chooses a given answer to a multiple-choice question if the individual is not copying.

Both the ω index and the GBT index calculate the probability of each answer if an individual is not cheating using the Nominal Response Model (NRM) proposed by Bock (1972). In particular, the probability that individual i chooses answer m to a given question q is taken to be

$$P_{iq}(m) = \frac{\exp(\zeta_m + \lambda_m \theta_i)}{\sum_{m' \in M_q} \exp(\zeta_{m'} + \lambda_{m'} \theta_i)},$$

where M_q indicates the set of answers to question q. Intuitively, the parameters ζ_m, λ_m for $m \in M_q$ capture the difficulty of question q or the distractors associated to different possible answers,

[5]We explored the other indices with our data. The K index, which only uses information from wrong answers, did not show much difference between the different groups. In line with the real-data test provided by Wollack (2003), this seems to reflect a poor performance of the K index detecting cheating. Zopluoglu and Davenport (2012) find little difference in the statistical power of the ω index and the GBT in the context of a simulation study, but we do not know of previous comparisons between the ω index and the GBT using real data.

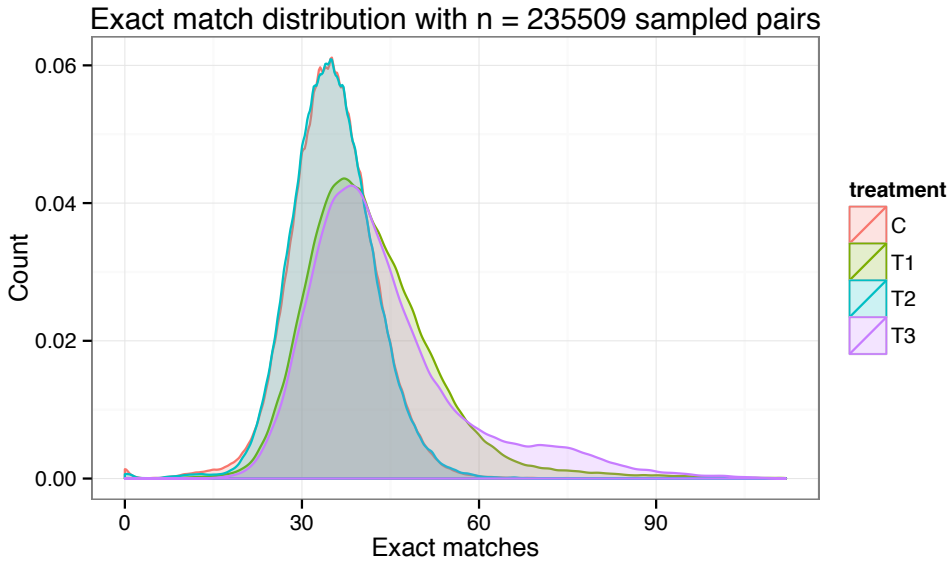

FIGURE 3. Distribution of exact matches for pairs of students

and the parameter θ_i captures the ability of individual i. Since $P_{iq}(m)$ is invariant to translations of the vector of $\zeta_m + \lambda_m \theta_i$, arbitrary linear restrictions on the parameters such as $\sum_{m \in M_q} \zeta_m = 0$ and $\sum_{m \in M_q} \lambda_m = 0$ allow to normalize the denominator of $P_{iq}(m)$ to one. The parameters can be estimated jointly by maximum likelihood using all the answers of all individuals taking a test.

The ω index then identifies copiers by computing the standardized difference between the number of answer matches between the pair of students and the number predicted by chance, conditional on the answers by the potential source, the estimated ability for the potential copier, and the estimated difficulty for each item. That is, if m_{jq} is the answer of student j to question q, and h_{ij} is the number of matches between the answers of student i and the answers of student j, the index ω_{ij} for the ordered pair (i, j) (where i is the potential copier and j the potential source) is given by

$$\omega_{ij} = \frac{h_{ij} - \sum_q P_{iq}(m_{jq})}{\sqrt{\sum_q (P_{iq}(m_{jq})(1 - P_{iq}(m_{jq})))}}.$$

The idea behind classification is that ω is approximately standard normal, a postulate based on the central limit theorem. Thus, i is classified as having copied from j, with α probability of accusing an innocent pair, if $1 - \Phi(\omega_{ij}) \leq \alpha$, where Φ is the standard normal distribution function. Student i is classified as a cheater if there is some other student j in the same classroom such that $1 - \Phi(\omega_{ij}) \leq \alpha$ or $1 - \Phi(\omega_{ji}) \leq \alpha$.

The GBT approach, instead, starts with the observation that the probability of an exact match between the answer of student i and the answer of student j to a given question q, under the null hypothesis that neither student has cheated, is equal to

$$P_{ijq} = \sum_{m \in M_q} P_{iq}(m) P_{jq}(m).$$

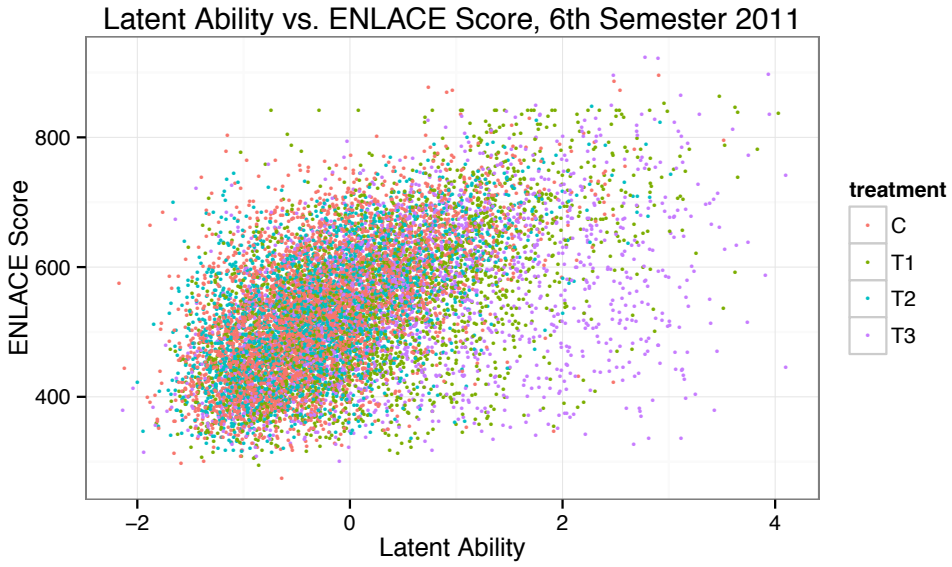

FIGURE 4. Latent ability vs. preprogram score

Then the index GBT_{ij} is computed as the probability of h_{ij} matches or more given that the set of questions in the exam is Q:

$$\mathrm{GBT}_{ij} = \sum_{\{Q' \subseteq Q : |Q'| \geq h_{ij}\}} \left(\prod_{q \in Q'} P_{ijq} \prod_{q' \notin Q'} (1 - P_{ijq'}) \right).$$

Students i and j are classified as having one of them copied from the other, with α probability of accusing an innocent pair, if $\mathrm{GBT}_{ij} \leq \alpha$. Note that, unlike ω, GBT treats both students symmetrically.

As an illustration, Figure 4 plots latent ability for 12th grade students in 2010-11, according to an NRM estimation, against their pre-program, 9th grade ENLACE score. It provides some external validity to the NRM estimation that the estimated ability is in fact well correlated with the score in a baseline exam carried out preprogram.

An important variable is the statistical threshold employed to detect cheating. Since each student belongs to many pairs as a potential source and as a potential copier, the probability of a false positive increases substantially at the student level. The results of the tests applied to different potential pairs for one given student are possibly not independent, however, so it is not useful to simply compound α. To handle this problem, we exploited the fact that the same exam was administered across classrooms, whilst answer copying is (precluding the use of electronic devises) likely confined to pairs of students in the same classroom. We tested for cheating every pair of students in the same school, and raised the threshold for accusing a given pair to the point where only 10% of students are accused of cheating in the control group because of an unusual similarity with a student in a different classroom. We consider only possible pairs within the same school for computational reasons.

Figure 5 illustrates our procedure for choosing α. In the top figure we consider all 12th grade students in the teacher and students incentive group in 2010-11. The horizontal axis shows decreasing values of α. For the different values of α, we depict the percentage of students who are accused of cheating because of an unusual similarity with another student in the same classroom,

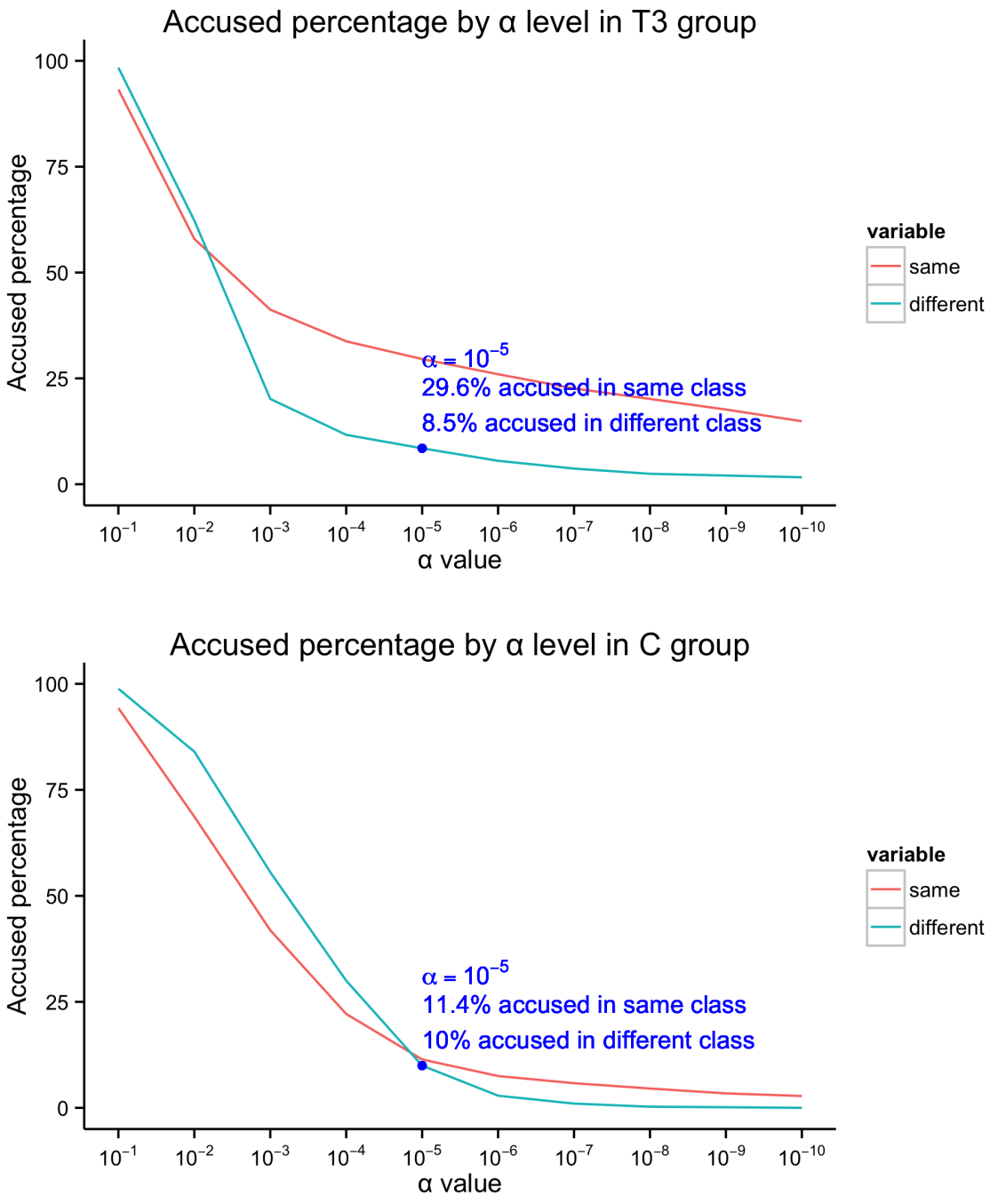

FIGURE 5. Estimated probability of cheating in all incentivized group and in control group, for different α values for the ω index

and the percentage of students who are accused of cheating because of an unusual similarity with another student in a different classroom in the same school, according to the ω index. Note that some students may be in both categories, so the two lines add to more than 100% of the percentage of students for relatively high values of α. The two lines are very close for α above 0.01, but they differ sharply for α below 0.001. At $\alpha = 0.00001$, the percentage of students accused because an

unusual similarity with a student in the same classroom is 29.6%, while the percentage of students accused exclusively because an unusual similarity with a student in a different classroom is 8.5%. Further lowering α does not help in distinguishing between the two categories. In the figure on bottom we repeat the exercise but considering the control group. As expected, since there are no incentives for cheating, the two lines are much closer. At $\alpha = 0.00001$, the percentage of students accused because an unusual similarity with a student in the same classroom is 11.4%, while the percentage of students accused exclusively because an unusual similarity with a student in a different classroom is 10%. We take the latter percentage as an approximation to the percentage of students erroneously accused of cheating. (Note that the figure on top suggests a lower probability of error of this kind.)

A potential difficulty for the estimation of the parameters of the NRM was the fact that exam conditions were different for the control and the treatment groups, given the different incentive schemes. To assess this problem, we re-estimated the parameters ζ_j, λ_j using only the control group, and used these estimates to recalculate latent ability for the students in the treatment groups. We then computed the statistical indices ω and GBT, and compared the classification of students between cheaters and non cheaters for different α values with those obtained when all parameters are estimated jointly. The results came out very similar, classifying almost exactly the same individuals as cheaters. Thus, we use both the control and treatment groups to estimate jointly item parameters and the latent ability of individuals. We observed little difference overall between using the ω index and using the GBT index. We settled in using the ω index with a threshold value of 0.00001, and calculating item response parameters and individual talent using both the control group and the treatment groups.[6]

For most pairs of students, it made little difference who was assumed to be the copier and who was assumed to be the source. As discussed above, the statistical detection of copying relies on the existence of an unusual similarity between the answers to the exam of a given student, the putative copier, and those of another student in the same classroom, the putative source, given the former's latent ability, as estimated using the student's exam answers. Unfortunately, if a student copies thoroughly from another student, the estimated latent ability for both will be very similar, and the ω test will report that they are copying from each other. For similar reasons, if several students are copying from the same source, and they are more or less thorough, the ω test will report that they are copying from each other. As a consequence of this, we can identify (likely) cheaters in the data, but we cannot distinguish copiers from sources. Similarly, we can identify sets of students who are likely to be active cheating components of the classroom network, but we cannot reasonably ascertain who is the source for each component. Note that we can use the baseline ENLACE exam to make an initial classification of cheaters into sources and copiers for isolated pairs of cheaters, but we cannot do that for larger connected components of the cheating network.

3.3. **Cheating in the ALI experiment.** Table 1 provides aggregate results from the statistical cheating analysis. The table reports the percentage of students who were members of at least one cheating pair for the cohort that went three years through the program. This cohort began in 10th grade in the ALI program in the 2008-09 year and completed 12th grade in the ALI program in the 2010-11 year. As already mentioned, this is the only cohort we observe during all three years of the program and thus have longitudinal data on tests scores and cheating over this time. The estimated percentage of cheaters in the control group varied between 5% and 7.5%, depending on the year.

[6]In Behrman et al. (2014) the method used to identify cheaters was Wesolowsky (2000). The results of that analysis are consistent with ours in terms of the extent of cheating and its distribution across schools and classrooms.

TABLE 1. Cheating in ALI by treatment group

Fraction of students involved in cheating			
	1st year 10th grade	2nd year 11th grade	3rd year 12th grade
No incentives	7.223	5.135	7.506
Student incentives	11.323	26.687	29.693
Difference in means	−4.099***	−21.551***	−22.187***
	(0.227)	(0.318)	(0.362)
Teacher incentives	7.356	8.969	9.599
Difference in means	−0.133	−3.834***	−2.093***
	(0.180)	(0.167)	(0.232)
All incentivized	7.347	23.510	31.878
Difference in means	−0.123	−18.375***	−24.372***
	(0.177)	(0.378)	(0.439)
Observations	11,530	11,530	11,530

Note: Statistical differences in means are with respect to No incentives group. Standard errors in parentheses; $^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$.

In the student incentives group, the estimated percentage of cheaters increased from about 11.3% the first year to nearly 30% the third. Similarly, in the group that received incentives for teachers and students, the estimated percentage of cheaters increased from around 7.4% to nearly 32%. The estimated percentage of cheaters in the teacher incentive group, per contra, barely increased from around 7.4% the first year to about 9.6% the third.

The aggregate estimates in Table 1 are reliable insofar as both types of errors, accusing innocent students and not accusing cheating students, roughly cancel out. Recall that our statistical analysis aims at a probability of accusing falsely a student of approximately 10%. A more conservative stance would entail disregarding the possibility of not accusing cheating students, and recalculating Table 1 taking into account the probability of accusing falsely a student. If the fraction of accused individuals for a given treatment and year is z, a conservative estimate $\underline{z}$ can be derived from $\underline{z}+0.1(1-\underline{z})=z$ for $z>0.1$ and $\underline{z}=0$ for $z\leq 0.1$, yielding $\underline{z}=\max\{0,(z-0.1)/0.9\}$. From this conservative viewpoint, cheating in the control group and in the teacher incentives group may have been close to zero through out the program. Cheating in the student incentive group, per contra, went from around 1.5% the first year to nearly 19% the second year to nearly 22% the third year. Similarly, cheating in the teacher and student incentives group went from close to nil to about 15% the second year to about 24% the third year. That is, even from this viewpoint, incentives had a very large effect on cheating, and the effect compounded with experience in the program.

4. EMPIRICAL ANALYSIS OF INDIVIDUAL CHEATING BEHAVIOR

4.1. **Descriptive analysis.** In Table 2 we compare the characteristics of cheaters with those of other students in the cohort that was three years in the program, from 10th to 12th grade, where we define as a cheater any student who was a member of at least one cheating pair in any year through

TABLE 2. Descriptive Analysis

	Cheaters	Non cheaters	All students
Age	15.265	15.365	15.335
	(0.867)	(0.891)	(0.885)
Gender (female = 1)	0.483	0.474	0.477
	(0.500)	(0.499)	(0.499)
Family members	3.906	3.976	3.966
	(1.116)	(1.149)	(1.144)
Monthly household income: 2000 to 4000 pesos	0.322	0.306	0.311
	(0.467)	(0.461)	(0.463)
Monthly household income: 4000 to 8000 pesos	0.169	0.137	0.147
	(0.374)	(0.343)	(0.354)
Monthly household income: more than 8000 pesos	0.110	0.071	0.083
	(0.313)	(0.256)	(0.276)
Have internet	0.204	0.128	0.152
	(0.403)	(0.334)	(0.359)
Three or four books at home	0.283	0.248	0.259
	(0.451)	(0.432)	(0.438)
Five or more books at home	0.119	0.084	0.095
	(0.324)	(0.278)	(0.294)
Mother with secondary education	0.313	0.318	0.316
	(0.464)	(0.466)	(0.465)
Mother with high school or technical education	0.208	0.155	0.171
	(0.406)	(0.362)	(0.377)
Mother with college or more	0.120	0.071	0.086
	(0.324)	(0.256)	(0.280)
Father with secondary education	0.284	0.291	0.289
	(0.451)	(0.454)	(0.453)
Father with high school or technical education	0.203	0.172	0.182
	(0.402)	(0.378)	(0.386)
Father with college or more	0.185	0.103	0.129
	(0.388)	(0.304)	(0.335)
I have a positive attitude[1]	0.915	0.898	0.903
	(0.279)	(0.303)	(0.296)
I believe I am a failure[1]	0.081	0.086	0.084
	(0.273)	(0.280)	(0.278)
I am at least as capable as most people[1]	0.930	0.924	0.926
	(0.256)	(0.264)	(0.262)
Hang out with friends often	0.090	0.073	0.078
	(0.286)	(0.260)	(0.269)
Have scholarship	0.284	0.347	0.328
	(0.451)	(0.476)	(0.469)
Baseline score	0.190	−0.081	0.000
	(1.068)	(0.958)	(1.000)
ALI math score	47.008	41.807	43.373
	(12.675)	(11.199)	(11.904)
Observations	3,471	8,059	11,530

Note: Cheaters defined as those students that cheated at least once during the three years. Standard errors in parenthesis. [1]Dummy variable where 1 means agree and 0 disagree with the statement.

TABLE 3. Determinants of cheating

	Marginal effects from logit of cheating					
	10th graders		11th graders		12th graders	
	1	2	3	4	5	6
Student incentives	0.031***	0.025***	0.221***	0.214***	0.223***	0.217***
	(0.007)	(0.007)	(0.010)	(0.010)	(0.010)	(0.010)
Teacher incentives	−0.001	−0.002	0.073***	0.073***	0.037***	0.034***
	(0.008)	(0.008)	(0.012)	(0.012)	(0.013)	(0.013)
All incentivized	−0.006	−0.007	0.201***	0.203***	0.238***	0.240***
	(0.008)	(0.008)	(0.011)	(0.011)	(0.010)	(0.010)
Baseline score	0.018***	0.003	0.033***	0.028***	0.024***	0.014
	(0.003)	(0.005)	(0.003)	(0.009)	(0.004)	(0.009)
Score×student incentives		0.028***		0.025**		0.036***
		(0.007)		(0.011)		(0.011)
Score×teacher incentives		0.007		−0.007		0.027**
		(0.008)		(0.013)		(0.014)
Score×All incentivized		0.014*		−0.016		−0.023**
		(0.008)		(0.011)		(0.011)
Age	−0.024	−0.023	−0.017	−0.013	−0.049**	−0.044**
	(0.015)	(0.015)	(0.019)	(0.019)	(0.019)	(0.019)
Age squared	0.002	0.002	0.002	0.001	0.006**	0.005**
	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)
Gender (female = 1)	−0.012**	−0.012**	0.010	0.010	−0.001	−0.002
	(0.005)	(0.005)	(0.007)	(0.007)	(0.007)	(0.007)
Family members	−0.001	−0.001	−0.002**	−0.002**	−0.002**	−0.002**
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
Monthly household income:	−0.003	−0.003	0.012	0.013	0.019**	0.020**
2000 to 4000 pesos	(0.007)	(0.007)	(0.009)	(0.009)	(0.009)	(0.009)
Monthly household income:	−0.009	−0.010	0.031***	0.032***	0.005	0.005
4000 to 8000 pesos	(0.009)	(0.009)	(0.011)	(0.011)	(0.012)	(0.012)
Monthly household income:	0.009	0.007	0.015	0.015	0.013	0.013
more than 8000 pesos	(0.011)	(0.011)	(0.014)	(0.014)	(0.015)	(0.015)
Have internet	0.003	0.003	0.018*	0.019*	0.017	0.019*
	(0.008)	(0.008)	(0.010)	(0.010)	(0.011)	(0.011)
3 or 4 books at home	0.005	0.005	0.017**	0.017**	0.000	0.000
	(0.007)	(0.007)	(0.008)	(0.008)	(0.009)	(0.009)
5 or more books at home	−0.001	−0.003	0.009	0.007	0.027**	0.026**
	(0.010)	(0.010)	(0.012)	(0.012)	(0.013)	(0.013)
Mother with secondary	0.014*	0.014*	0.009	0.009	0.003	0.003
education	(0.007)	(0.007)	(0.009)	(0.009)	(0.010)	(0.009)
Mother with high school	0.029***	0.029***	0.030***	0.031***	0.008	0.01
or technical education	(0.009)	(0.009)	(0.011)	(0.011)	(0.012)	(0.012)
Mother with college	0.033***	0.034***	0.030**	0.032**	0.026*	0.027*
or more	(0.011)	(0.011)	(0.015)	(0.015)	(0.016)	(0.016)
Father with secondary	0.021***	0.020***	0.012	0.011	0.002	0.001
education	(0.008)	(0.008)	(0.009)	(0.009)	(0.010)	(0.010)
Father with high school	0.022**	0.021**	0.009	0.007	0.041***	0.039***
or technical education	(0.009)	(0.009)	(0.011)	(0.011)	(0.012)	(0.012)
Father with college	0.030***	0.029***	0.037***	0.034***	0.066***	0.063***
or more	(0.011)	(0.011)	(0.013)	(0.013)	(0.014)	(0.014)
I have a positive attitude[1]	0.000	−0.001	0.016	0.016	0.007	0.007
	(0.010)	(0.010)	(0.013)	(0.013)	(0.014)	(0.013)

continues next page

TABLE 3 continues:

I believe I am a failure[1]	−0.022*	−0.021*	0.007	0.007	−0.010	−0.011
	(0.012)	(0.012)	(0.013)	(0.013)	(0.014)	(0.014)
I am at least as capable	0.020	0.020	0.009	0.010	−0.007	−0.005
as most people[1]	(0.013)	(0.012)	(0.014)	(0.014)	(0.015)	(0.015)
Hang out with friends often	0.014	0.015	0.021*	0.023*	0.012	0.015
	(0.010)	(0.010)	(0.012)	(0.012)	(0.014)	(0.014)
Have scholarship	−0.011*	−0.01	−0.01	−0.01	−0.011	−0.011
	(0.006)	(0.006)	(0.008)	(0.008)	(0.008)	(0.008)
Observations	11,530	11,530	11,530	11,530	11,530	11,530

Notes: Standard errors in parentheses; $^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$. [1]Dummy variable where 1 means agree and 0 disagree with the statement.

the program. Table 2 suggests that those students who cheat are somewhat better off economically than their non-cheating counterparts and more likely to live in smaller families. Cheaters are also more likely to report hanging out with friends frequently, and have in average a higher baseline score (i.e. the ENLACE 9th grade math score). Overall, this initial picture suggests that those engaging in cheating have a somewhat more privileged family background and greater social networks.

4.2. **Determinants of cheating behavior.** In Table 3, we report marginal effects (percentage points) of the different treatments, of the baseline score, and of the personal characteristics and family background of the student, on the probability of the student being detected as a cheater, obtained from logit regressions. The marginal effects are taken with respect to the control treatment (no incentives). We restrict our analysis to students who remain in the cohort of reference the three years of the program. Columns 1, 3 and 5 report the results for this cohort as it went through 10th, 11th, and 12th grade. Columns 2, 4 and 6 report again the results for this cohort, but including interaction terms between the baseline score and the different incentive treatments.

We find that student incentives significantly increase the probability of cheating. The marginal effect of student incentives on the probability of cheating is 3 percentage points in the first year in the program and around 22 percentage points in the second and third year, an increase that likely reflects the accumulation of experience in cheating as well as the credibility of the incentive rewards associated to the program. Similarly, the marginal effect of student and teacher incentives is nearly nil the first year of the program but in the range of 20 to 24 percentage points in the second and third year. Teacher incentives alone have a much lower marginal effect of cheating, nil the first year and between 3 and 7 percentage points the second and third year.

Recall that the baseline score is included as a proxy of ability. A higher baseline score also reduces incentives to copy (see Table 8 in the Appendix), so any positive impact of this score on the probability of cheating is likely due to sources in cheating. We find that the effect of the baseline score is positive and statistically significant, for the score itself and especially for the interaction of the score with student incentives. This last result supports the hypothesis that better students are drafted as sources when incentives are high. The interaction term between baseline score and incentives for students and teachers is near zero or has the wrong sign, which suggests even lower costs of communication for this treatment. Even though the exam was proctored by staff from the Secretariat of Public Education, the fact that the school administrators and teachers had a stake in the performance of the students may have had an impact on the cost of communication, perhaps via a reduced stigma of copying or a reduction of expected costs of unrelated to monitoring.

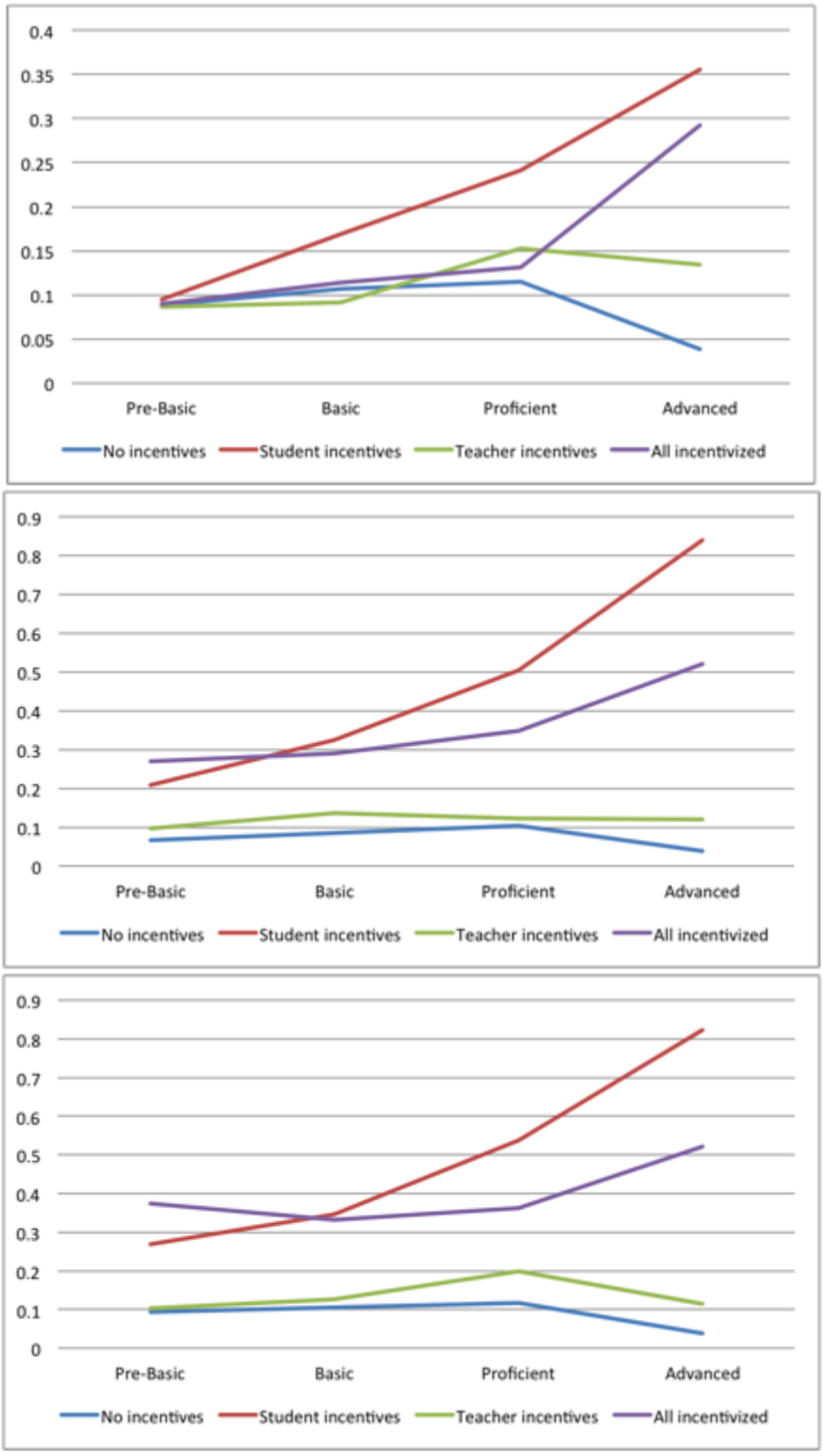

FIGURE 6. Probability of cheating by pre-program achievement, from top to bottom: 10th graders, 11th graders and 12th graders

TABLE 4. Persistence of cheating

	Marginal effects from logit of cheating	
	11th graders	12th graders
Cheating prior school year	0.208***	0.224***
	(0.020)	(0.022)
Student incentives	0.223***	0.169***
	(0.012)	(0.011)
Teacher incentives	0.078***	0.035***
	(0.014)	(0.013)
All incentivized	0.210***	0.184***
	(0.012)	(0.011)
Cheating prior school year×student incentives	−0.058**	−0.039
	(0.025)	(0.025)
Cheating prior school year×teacher incentives	−0.021	−0.071**
	(0.029)	(0.031)
Cheating prior school year×all incentivized	−0.047*	−0.014
	(0.027)	(0.026)
Baseline score	0.028***	0.013***
	(0.003)	(0.003)
Observations	11,530	11,530

Notes: Standard errors in parentheses; $^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$. Control variables include the same personal and family characteristics included in Table 2.

In the first year of the program, when there is less experience cheating, some personal characteristics seem to be influential in the decision to cheat, including being male, and having high self-esteem (as measured by answers to attitudinal questions). Personal characteristics have a smaller impact on the probability of cheating the second and third year, except for parental education which may be related to ability.

To illustrate further the relationship between ability, incentives, and cheating, we plot in Figure 6 the relationship between the baseline score and the probability of engaging in cheating by treatment group in 10th grade, 11th grade and 12th grade. It is noteworthy that the relationship is basically flat with a slightly downward slope for the control group and for the teacher incentives group, indicating that higher levels of pre-program achievement are associated with constant or lower probabilities of cheating. For the student incentive group and for the teacher and student incentive group, instead, the relationship is strongly positive. That is less so for the teacher and student incentive group in 12th grade, but for this incentive treatment, cheating in 12th grade seem to include a large fraction of students. As illustrated by the example in the previous section, incentives appear to draw higher ability students as sources for cheaters.

4.3. **Persistence of cheating.** We also look at the persistence of cheating and the relationship with incentives. In Table 4 we report marginal effects (percentage points) of the different determinants of cheating, including now cheating the previous year for the second and third year in the program of the cohort that began in 10th grade in the ALI program in the 2008-2009 year. We include as controls the same personal and family characteristics from Tables 2 and 3. Table 4 shows that there is substantial correlation in cheating over time although there is still movement in and out of cheating over time. We include in the estimation interaction terms between cheating the previous year and incentive treatments; these generally come up as negative, reflecting a much

TABLE 5. Classrooms by year and treatment

	Number of classrooms (and students per classroom)		
	10th graders	11th graders	12th graders
No incentives	169 (33)	164 (28)	156 (26)
Student incentives	125 (35)	122 (32)	122 (28)
Teacher incentives	120 (33)	110 (30)	105 (28)
All incentivized	112 (32)	107 (29)	99 (27)

larger increase in cheating among students who did not previously cheat in incentivized treatments than in the control group.

5. EMPIRICAL ANALYSIS OF CLASSROOM CHEATING NETWORKS

5.1. **Cheating by school and classroom.** In this section we change our perspective from a focus on individual students to a focus on classroom networks. Networks provide us a picture of the structure of illicit communication in the classroom, beyond the number and characteristics of those involved. Table 5 provides information about the number of classrooms and (in parenthesis) the average number of students per classroom in the cohort by treatment and year. There are more students leaving than new arrivals every year, and schools sometimes reduce the number of classrooms and reassign remaining students.

Figures 7 to 10 illustrate the percentage of cheaters by classroom for the different treatment groups for the cohort of interest in each year in the program. Classrooms are grouped by school. Schools are ordered from left to right in decreasing order with respect to the prevalence of cheating in the last year in the program; classrooms within each school are ordered from left to right in decreasing order with respect to the prevalence of cheating in the corresponding year. The height of the bar indicates the percentage of cheaters in a given classroom, while the absence of a bar indicates no cheating in a classroom. Consistent with the individual data, there is more cheating in classrooms under the student incentive group and the teacher and student incentive group. Moreover, cheating in those groups increases considerably in going from the first to the second year in the program, while cheating in the other two groups declines or remains constant.

The figures show a large variance in the prevalence of cheating across schools within each treatment group, even in the no incentives group. For instance, under student incentives, the last year in the program, cheating goes from being more than 50% for every classroom in a school in one extreme to being zero in a couple of schools in the opposite extreme. There is also some diversity in the prevalence of cheating for different classrooms within the same school. For instance, in the fourth school from the right, under student incentives, the last year in the program, cheating goes from more than 75% in one classroom to zero in another. Overall, the data suggests the existence of a "cheating culture" in a few schools and in some classrooms.

5.2. **Cheating networks by treatment.** In Table 6, we report on detected communication activity within each classroom. For the calculation, we use the directed links detected using the ω test as described in Section 3. We consider two indicators: network density and average degree. Network density is defined as the percentage of directed edges that are active in the classroom, that is the number of directed links detected using the ω test divided by the number of possible links, that is $N(N-1)$ if the school size is N. We report the density averaged over classrooms for each treatment and year. The degree of a student is a defined as the number of students he or she is found to have

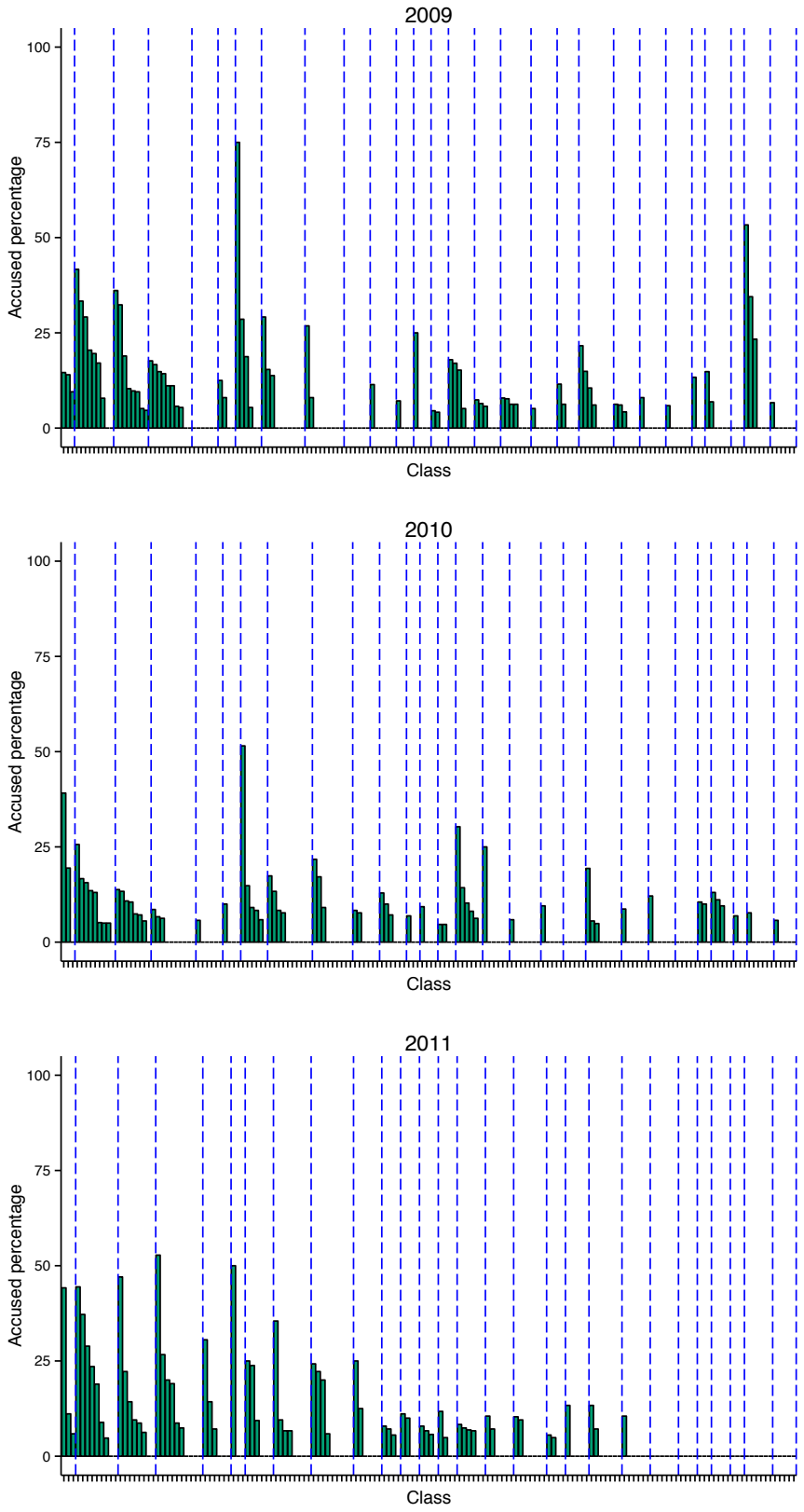

FIGURE 7. Percentage of cheaters by classroom and school: control group

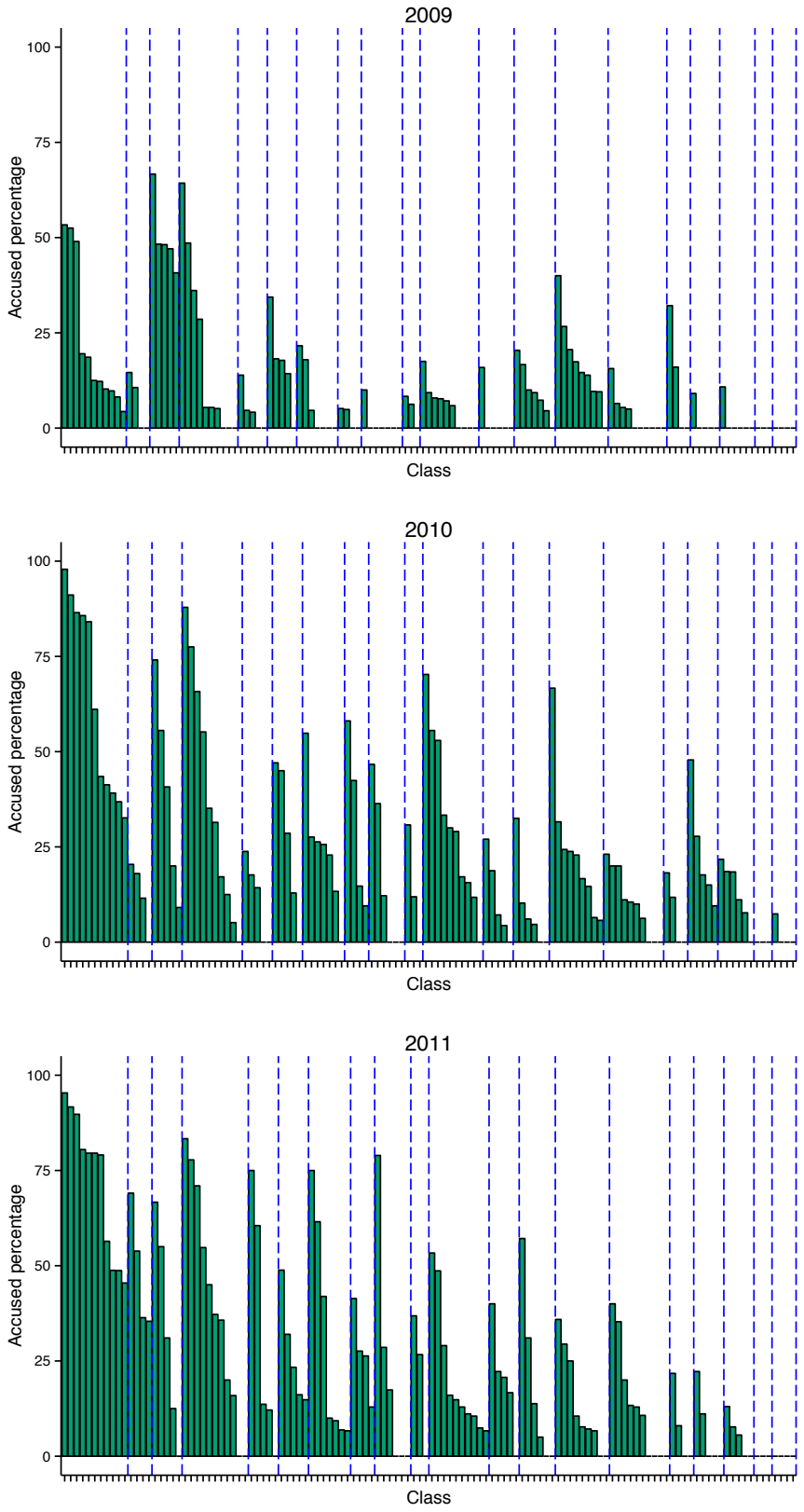

FIGURE 8. Percentage of cheaters by classroom and school: student incentives treatment

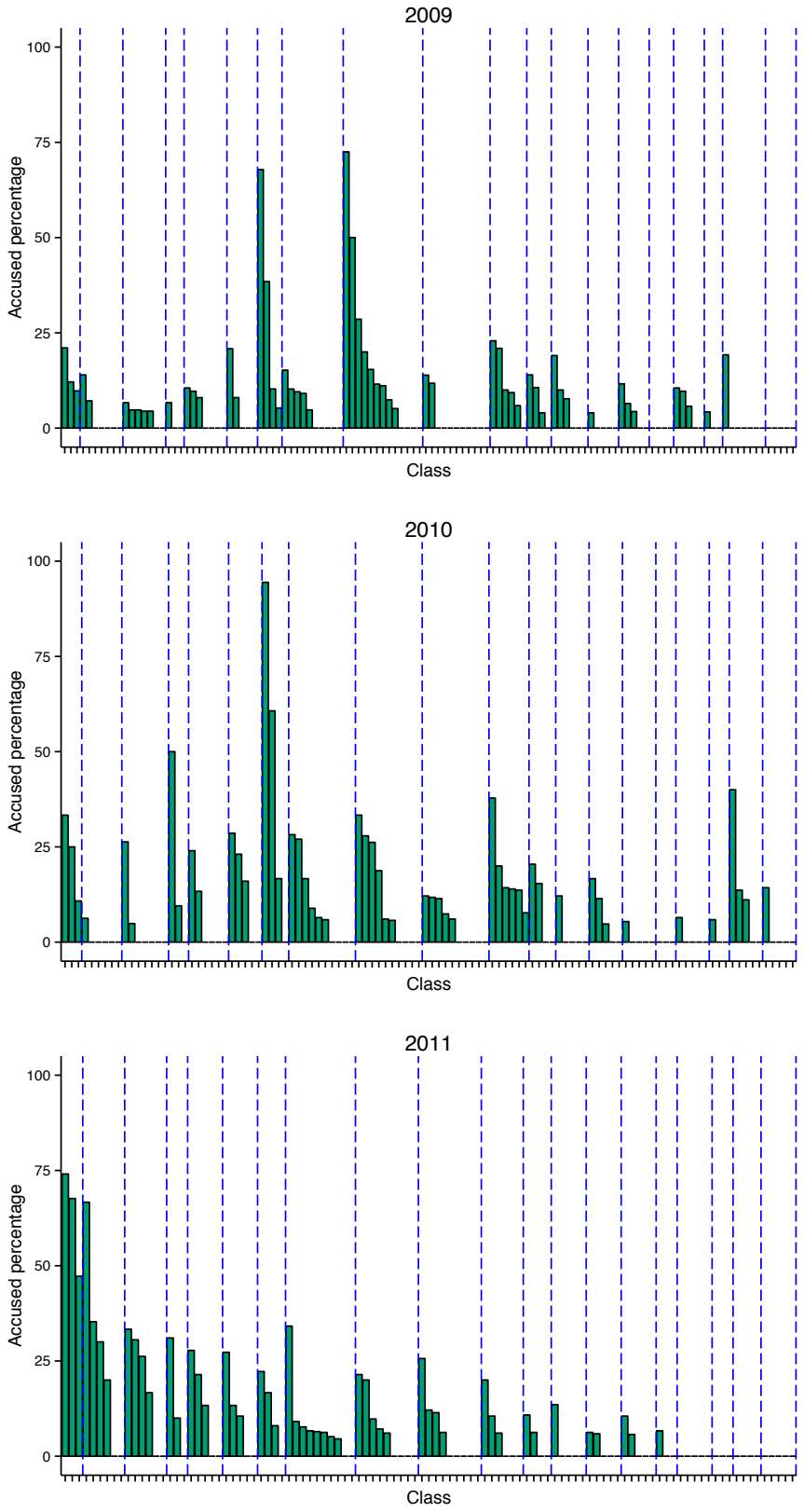

FIGURE 9. Percentage of cheaters by classroom and school: teacher incentives treatment

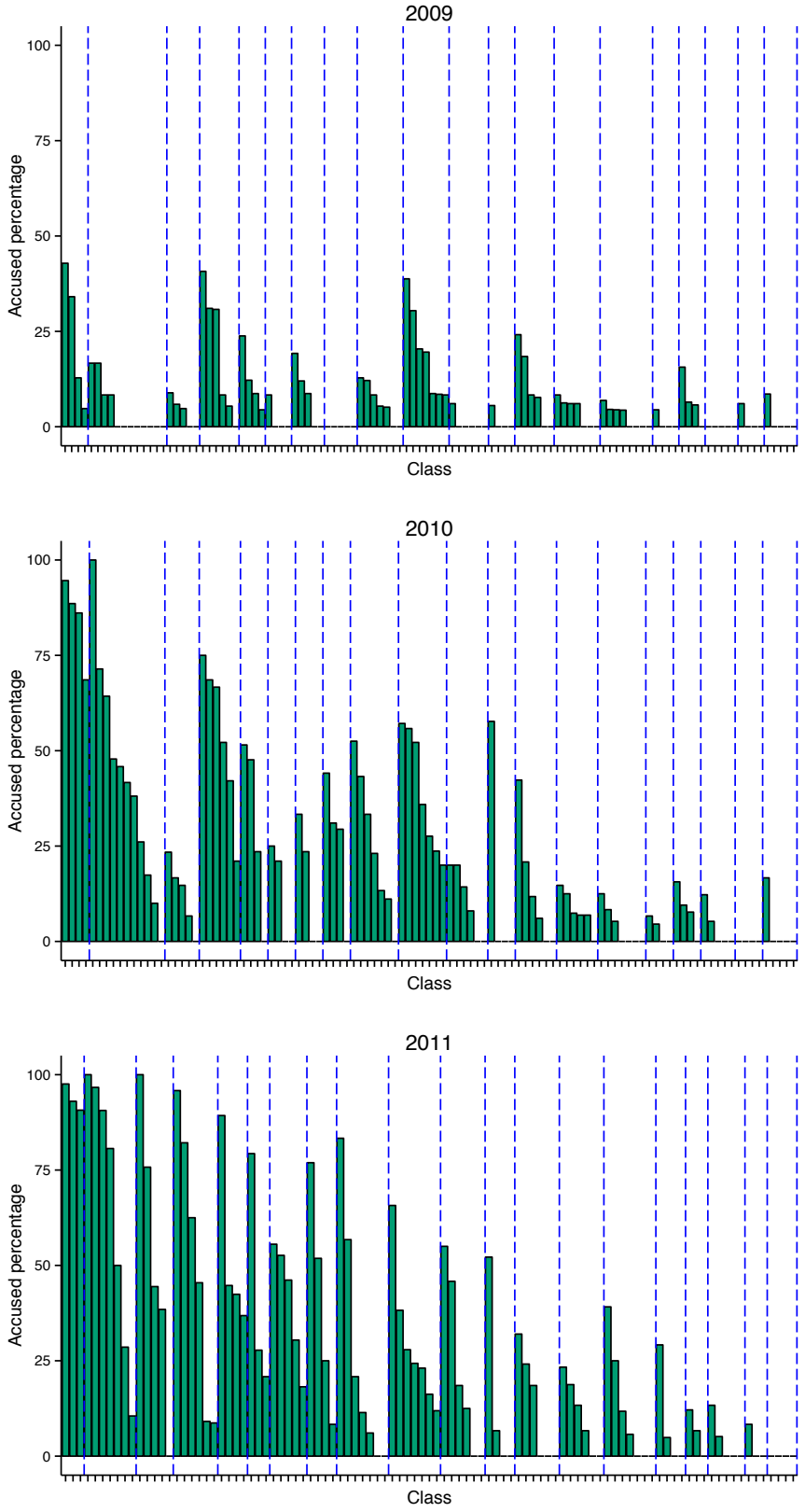

FIGURE 10. Percentage of cheaters by classroom and school: all incentivized treatment

TABLE 6. Communication activity by treatment and year

	Network density		
	10th graders	11th graders	12th graders
No incentives	0.00427	0.00244	0.00319
Student incentives	0.00494	0.02343	0.03965
Teacher incentives	0.00404	0.00728	0.00614
All incentivized	0.00283	0.02743	0.05836

	Average degree		
	10th graders	11th graders	12th graders
No incentives	0.1385	0.0821	0.0991
Student incentives	0.1730	0.9744	1.5107
Teacher incentives	0.1345	0.1737	0.1634
All incentivized	0.1074	0.8805	2.1679

TABLE 7. Communication groups by treatment and year

	Average size of active component		
	10th graders	11th graders	12th graders
No incentives	2.320	1.421	1.923
Student incentives	3.832	8.336	8.361
Teacher incentives	2.275	2.727	2.619
All incentivized	2.339	6.626	8.495

copied from. We report the degree averaged over students for each treatment and year. The two indicators are closely related by definition; however, since we average density over classrooms and not over students, density gives relatively more weight to students in smaller classrooms.

Consistent with the individual level data, there is comparatively little variation over time in network density and average degree for the control group, and some increase for the teacher incentives group in between the first and second year of the program. Per contra, network density in the student incentives group has a fivefold increase in the second year, and almost doubles again in the third year. Similarly, network density in the student and teacher incentive group raises almost tenfold in the second year, and doubles again in the third year. Average degree tells a similar story.

As discussed in Section 3, the density and average degree derived from pairwise statistical tests reflects not only "true" copier-to-source links but also overall unusual similarities between students in the same component probably due to the fact that copying becomes not only more widespread, but also more intensive in reaction to incentives.

We have also calculated the average size of active components of the cheating network by treatment and year, and report the results in Table 7. An active component is defined as a maximal subset of students in the classroom with the property that every pair of students in the subset is connected directly or indirectly by cheating according to the ω test. Again, we consider the cohort that underwent three years in the program, and consider all students in each classroom. We code zero as the average active component of a classroom when there is no detected cheating, and otherwise we calculate the average component for each classroom weighing each active component

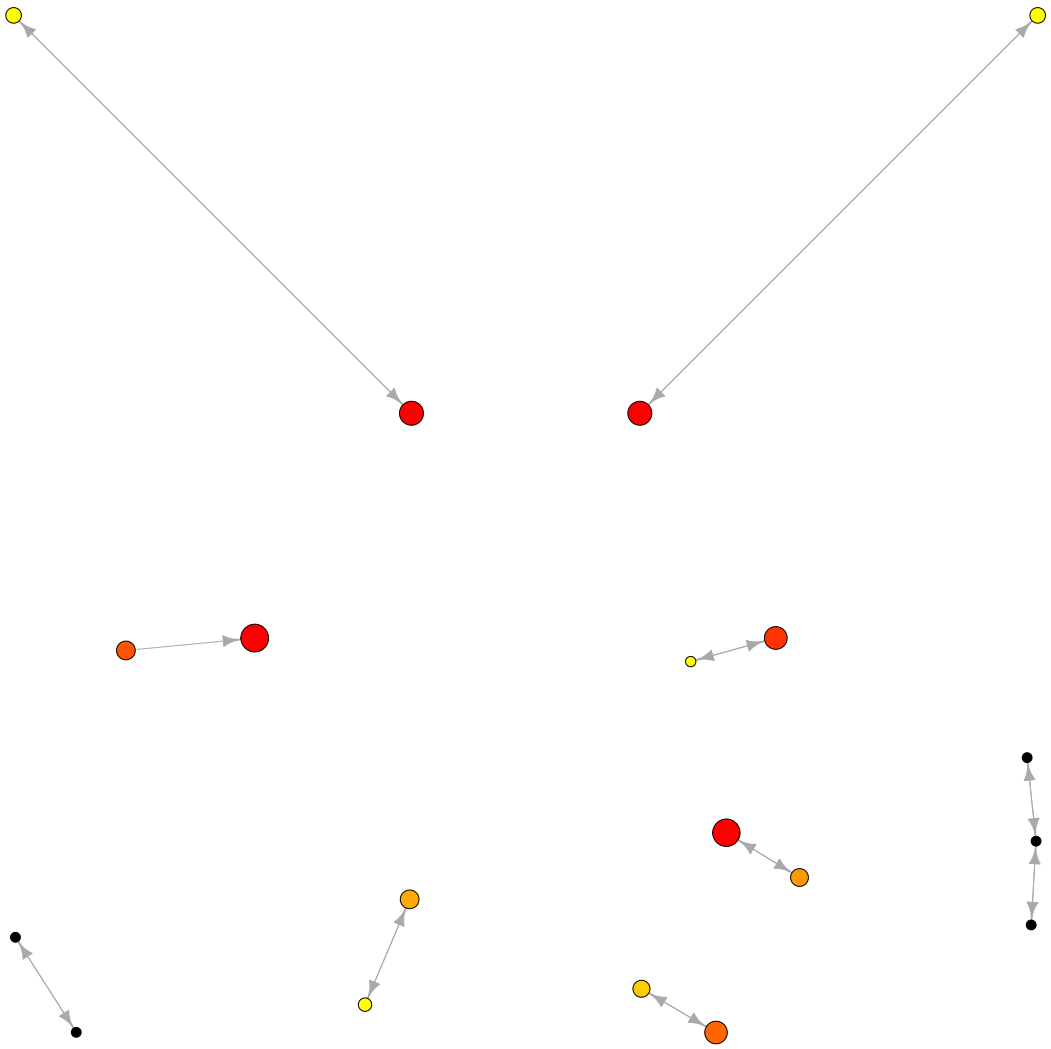

FIGURE 11. Small cheating networks in a school in the all incentivized treatment

size by the corresponding number of cheaters. Then we take the simple average over classrooms by treatment and year.

As transpires from Table 7, cheating in the no incentives and the teacher incentives groups is mainly the result of the activity of isolated cheating pairs; per contra, in the student and in the all incentivized groups, some cheating is performed in large (relative to classroom size) groups of students from the second year of the program on.

In Figures 11 and 12, we reproduce the cheating networks in two out of the twenty schools in the all incentivized group the third year into the program. Each node is an individual who was detected cheating. Each edge points in the direction in which the cheating was detected; for the reasons discussed in Section 2, most edges point both ways. The color and size of each vertex is

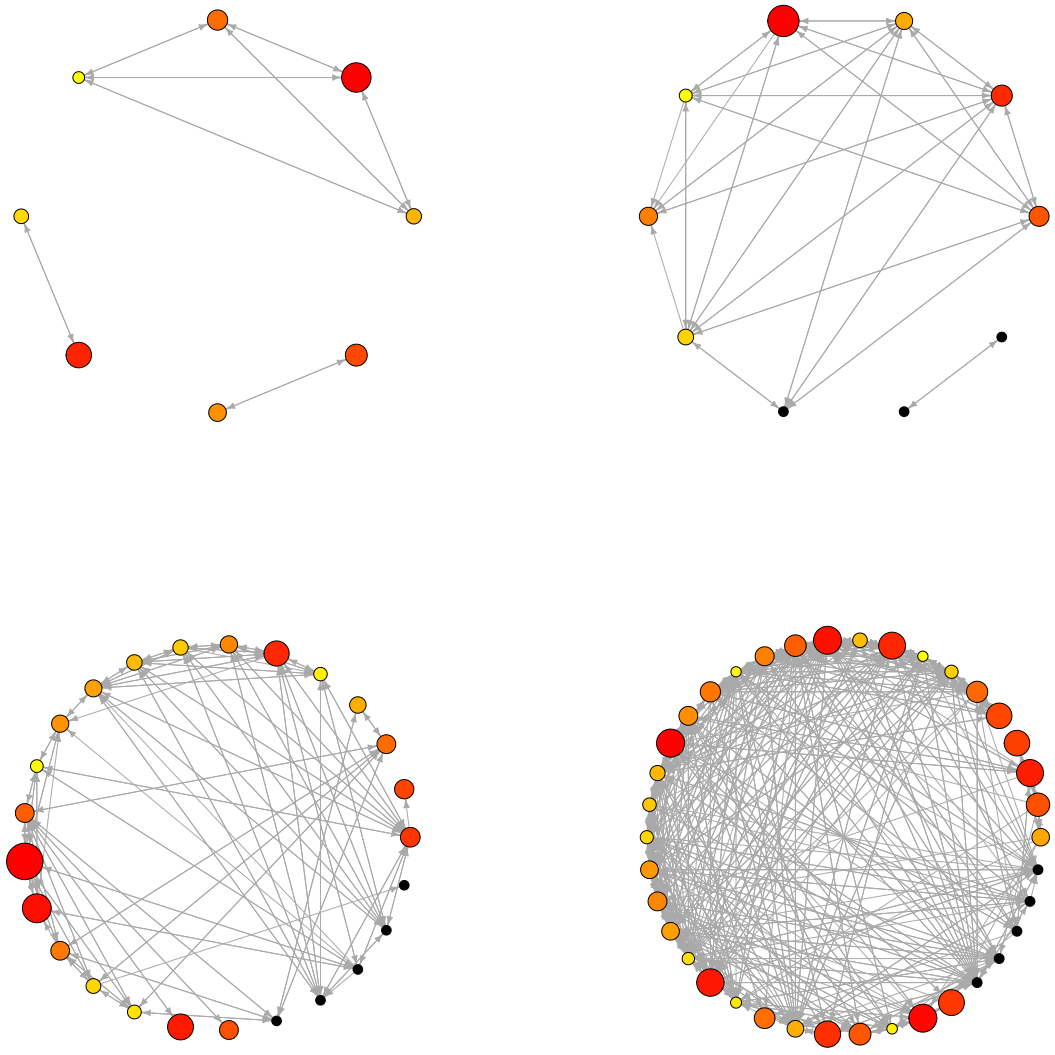

FIGURE 12. Large cheating networks in a school in the all incentivized treatment

determined by the baseline scores previous to the ALI tests. The color hue goes from yellow to red and the size from small to big, which corresponds to bad and good scores correspondingly. The node size is determined by a linear transformation of the baseline scores. The node color, instead, was done separately for each classroom so that yellow is the minimum score in the classroom and red is the highest. When the baseline score was not available for a student, the node is a small black point.

The school depicted in Figure 11 has relatively little cheating; besides the four classrooms depicted, three others had no detected cheating activity. Classroom sizes are 35 and 17 in the upper row and 24 and 23 in the lower row. (The classrooms with no cheaters have 19, 20 and 31 students.) Almost all detected cheating occurs in isolated pairs, conforming most closely to the model of one

copier and one source underlying the statistical detection methods. Note that in the classroom to the right in the second row more than a third of the students is involved in cheating, but the density of the cheating network is still fairly small.

The school depicted in Figure 12, per contra, is a heavy offender. Still, besides the four classrooms depicted, there is another one with no detected cheating activity. Classroom sizes are 18 and 26 in the upper row and 33 and 38 in the lower row. (The classroom with no cheaters has 26 students.) Cheating in the classroom depicted to the left of the upper row occurs in three different components; given the relatively small size of the classroom, more than 40% of the students are involved. Cheating in the school to the right occurs, in two components, one of them with eight students connected by multiple edges. The latter school has a slightly smaller percentage of cheaters but a larger cheating density. Since it is hard to imagine eight or more students comparing answers pairwise, these and other heavily connected classrooms seem to corroborate that unusual similarities are due to copying from the same sources. Cheating in the classrooms in the lower row is even more egregious, involving three quarters of the class in one case and the entire class in the other. Such densely connected classrooms occur occasionally in other schools in the student incentive and the teacher and student incentive groups.

6. FINAL REMARKS

In recent years, there has been a movement toward greater accountability in education provision by the introduction of nationwide standardized testing in many countries and even across countries, as exemplified by the OECD Programme for International Student Assessment (PISA) testing. Reliance on standardized testing has led naturally to increases in the stakes in the tests both for teachers and students, through the policy consequences for the schools or through the consequences for student promotion. Policy experiments in a few countries have gone further, linking explicitly test scores to financial incentives for teachers and students in the hopes of improving learning. These interventions have often been careful in avoiding manipulation of test scores via "teaching to the test" (Kremer et al. 2009) or have engaged external monitors to avoid cheating being orchestrated by teachers. To the extent of our knowledge, however, there has been no previous systematic study on the effect of the high stakes on answer copying.

In this paper, we explore cheating in the context of an intervention conducted in a sample of Mexican high schools which included different incentive schemes associated to performance in standardized exams. We adapt methods from the education measurement literature designed with the intention to test whether a particular pair of students have engaged in copying in a multiple-choice exam. Those methods generally set a threshold for the similarity between the answers provided by two students such that if the threshold is exceeded the pair becomes suspect. In the intervention we study, the same exam was applied to students in different classrooms; we exploit this feature of the intervention to set a threshold such that the probability of accusing a student because of similarity with a student in a different classroom would be 10%, which we take to be the probability of accusing an innocent student. Financial incentives to students, combined with repeated participation in the program, seem to have had a large impact on cheating. Under conservative assumptions regarding classification errors, by the third year in the program, around 20% of students may have engaged in cheating in treatments that provided financial incentives to students, while the corresponding percentage without incentives or in a treatment that provided incentives to teachers may have been negligible. Cheating is not distributed homogeneously over the sample of high schools in each treatment; there are schools with widespread cheating and schools with no detected cheating in every treatment.

We also look at the network defined by excessive similarity in the answers to the exam within each classroom. For classrooms in the no incentives or in the teacher only incentive groups, groups of students connected by cheating are generally isolated pairs. For classrooms in the treatments that provide incentives to students, however, especially in the second and third year in the program, groups of students connected by excessive similarity can be quite large, encompassing for a couple of schools a large percentage of each classroom. The combination of high incentives and experience in the program seem to have given rise to extensive collaboration in illicit communication in these classrooms.

Our evidence indicates that an immediate worry for policy interventions relying on explicit monetary incentives is to be aware of the participants' attempts to improve measured rather than actual performance. Extensive cheating, as detected in some classrooms, is pernicious because it blunts the incentives for learning created by the program, at least for potential copiers. This is because copying and learning are to some extent substitute activities, and copying is likely facilitated if it is deemed acceptable by classmates. If monetary incentives undermine the role of stigma and other moral considerations in deterring copying, it may be that *smaller* incentives are actually more effective for learning, depending on the different elasticities of cheating and effort in learning to cash incentives. There is a pointed contrast here with recent literature stressing the point that monetary incentives may crow out moral restraints on behavior in some environments; in those environments, unlike ours, it may be optimal to provide either no incentives or large enough monetary incentives (Sneezy and Rustichini 2000, Gneezy et al. 2011). In both cases, it is possible that monetary incentives have non monotonic effects due to the interaction with moral motivations. At any rate, every policy intervention has the potential to trigger unintended consequences, chiefly among them the participants' attempts to game the rules. A careful research of those attempts must be an important ingredient of every intervention with an aspiration to influence policy.

REFERENCES

Joshua Angrist and Victor Lavy. The effects of high stakes high school achievements awards: Evidence from a randomized trial. *American Economic Review*, 99:1384–1414, 2009.

Coralio Ballester, Antoni Calvó-Armengol, and Yves Zenou. Delinquent networks. *Journal of the European Economic Association*, 8:34–61, 2010.

Jere Behrman, Susan Parker, Petra Todd, and Ken Wolpin. Aligning learning incentives of students and teachers: Results from a social experiment in Mexican high schools. *Journal of Political Economy*, forthcoming, 2014.

Roland Benabou and Jean Tirole. Incentives and prosocial behavior. *American Economic Review*, 96(5):1652–1678, 2006.

Roland Benabou and Jean Tirole. Laws and norms. NBER Working Paper #17579, 2011.

R. Darrell Bock. Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika*, 37:29–51, 1972.

Barbara Brandes. *Academic honesty: A special study of California students*. California State Department of Education, Bureau of Publications, Sacramento, CA, 1986.

Antoni Calvó-Armengol and Yves Zenou. Social networks and crime decisions: The role of social structure in facilitating delinquent behavior. *International Economic Review*, 45:939–958, 2004.

Scott Carrell, Frederick V. Malmstrom, and James E. West. Peer effects in academic cheating. *Journal of Human Resources*, 43:173–207, 2008.

Gary Charness, David Masclet, and Marie-Claire Villeval. The dark side of competition for status. *Management Science*, 60(1):38–55, 2013.

Gregory Cizek. *Cheating on Tests: How to do it, Detect it and Prevent it.* Lawrence Erlbaum, Mahwah, NJ, 1999.

Gregory Cizek, Michael B. Bunch, and Heather Koons. Setting performance standards: Contemporary methods. *Educational Measurement: Issues and Practice*, 23:31–62, 2004.

Allan S. Cohen and James Wollack. Test administration, security, scoring, and reporting. In R.L. Brennan, editor, *Educational Measurement*, pages 356–386. Praeger Publishers, 4th edition, 2006.

Stephen F. Davis and H. Wayne Ludvigson. Additional data on academic dishonesty and a proposal for remediation. *Teaching of Psychology*, 22:119–121, 1995.

Stephen F. Davis, Cathy A. Grover, Angela H. Becker, and Loretta N. McGregor. Academic dishonesty: Prevalence, determinants, techniques, and punishments. *Teaching of Psychology*, 19:16–20, 1992.

David Figlio and Joshua Winicki. Food for thought: the effects of school accountability plans on school nutrition. *Journal of Public Economics*, 89:381–394, 2005.

Roland Fryer. Financial incentives and student achievement: Evidence from randomized trials. *Quarterly Journal of Economics*, 126:1755–1798, 2011.

Edward L. Glaeser, Bruce Sacerdote, and José A. Scheinkman. Crime and social interactions. *Quarterly Journal of Economics*, 111:507–548, 1996.

Paul Glewwe, Nauman Ilias, and Michael Kremer. Teacher incentives. *American Economic Journal: Applied Economics*, 2:205–227, 2010.

Uri Gneezy and Aldo Rustichini. Pay enough or don't pay at all. *Quarterly Journal of Economics*, 115:791–810, 2000.

Uri Gneezy, Stephan Meier, and Pedro Rey-Biel. When and why incentives (don't) work to modify behavior. *Journal of Economic Perspectives*, 25:191–210, 2011.

Paul W. Holland. Assessing unusual agreement between the incorrect answers of two examinees: Using the K-index. Education Testing Service, Princeton, NJ, 1996.

C. Kirabo Jackson. A little now for a lot later: A look at a Texas advanced placement incentive program. *Journal of Human Resources*, 45:591–639, 2010.

Matthew Jackson. *Social and Economic Networks*. Princeton University Press, 2010.

Brian Jacob and Steven Levitt. Rotten apples: An investigation of the prevalence and predictors of teacher cheating. *Quarterly Journal of Economics*, 118:843–877, 2003.

Michael Kremer, Edward Miguel, and Rebecca Thornton. Incentives to learn. *Review of Economics and Statistics*, 91:437–456, 2009.

Steven D. Levitt, John A. List, and Sally Sadoff. The effect of performance-based incentives on educational achievement: Evidence from a randomized experiment. University of Chicago, 2010.

Claudio Lucifora and Marco Tonello. Students' cheating as a social interaction: Evidence from a randomized experiment in a national evaluation program. IZA Discussion Papers 6967, Bonn, Germany, 2012.

Jan R. Magnus, Victor M. Polterovich, Dmitri L. Danilov, and Alexei V. Savvateev. Tolerance of cheating: An analysis across countries. *Journal of Economic Education*, 3:125–135, 2002.

Karthik Muralidharan and Venkatesh Sundararaman. Teacher incentives in developing countries: Experimental evidence from India. *Journal of Political Economy*, 119:39–77, 2011.

Manuel Reif. *mcIRT: IRT models for multiple choice items*, 2014. URL `https://github.com/manuelreif/mcIRT`.

Mauricio Romero, Alvaro Riascos, and Diego Jara. Answer-copying index comparison and massive cheating detection. Universidad de los Andes, unpublished manuscript, 2012.

Fred Schab. Schooling without learning: Thirty years of cheating in high school. *Adolescence*, 26: 839–847, 1991.

Leonardo S. Sotaridona and Rob R. Meijer. Statistical properties of the K-index for detecting answer copying. *Journal of Educational Measurement*, 39:115–132, 2002.

Leonardo S. Sotaridona and Rob R. Meijer. Two new statistics to detect answer copying. *Journal of Educational Measurement*, 40:53–69, 2003.

Matthew G. Springer, Dale Ballou, Laura Hamilton, Vi-Nhuan Le, J.R. Lockwood, Daniel F. McCaffrey, Matthew Pepper, and Brian M. Stecher. *Teacher Pay for Performance: Experimental Evidence from the Project on Incentives in Teaching*. National Center on Performance Incentives at Vanderbilt University, Nashville, TN, 2010.

Wim J. van der Linden and Leonardo S. Sotaridona. Detecting answer copying when the regular response process follows a known response model. *Journal of Educational and Behavioral Statistics*, 31:283–304, 2006.

George O. Wesolowsky. Detecting excessive similarity in answers on multiple choice exams. *Journal of Applied Statistics*, 27:909–921, 2000.

James Wollack. A nominal response model approach for detecting answer copying. *Applied Psychological Measurement*, 21:307–320, 1997.

James Wollack. Comparison of answer copying indices with real data. *Journal of Educational Measurement*, 40:189–205, 2003.

James Wollack. Simultaneous use of multiple answer copying indexes to improve detection rates. *Applied Measurement in Education*, 19:265–288, 2006.

Cengiz Zopluoglu. [software note] copydetect: An R package to compute statistical indices for detecting answer copying on multiple-choice tests. *Applied Psychological Measurement*, 37(1):

75–77, 2013.

Cengiz Zopluoglu and Ernest Davenport. The empirical power and type I error rates of the GBT and ω indices in detecting answer copying on multiple-choice tests. *Educational and Psychological Measurement*, 72(6):975–1000, 2012.

APPENDIX: MONETARY INCENTIVES IN THE ALI EXPERIMENT

In this Appendix we elaborate on the structure of incentive payments associated to ALI.

6.1. **Treatment 1 (Student incentives group).** Table 8 shows the incentive payment schedule for students at each grade for the student incentives treatment. The amount in each cell represents the payment in Mexican pesos for a student with a given level of performance at the start of the grade (the baseline exam score) and at the end of the grade. As a reference, the exchange rate fluctuated between 12 pesos per US dollar and 13 pesos for US dollar during the yeas of the program. Payment levels were intended to be large enough to be expected to induce behavioral changes. The payments are similar in magnitude to the attendance incentives given by the Oportunidades program and to a scholarship program pioneered by the Secretariat of Public Education. As seen in the table, in the 10th and 11th grades, payments are larger when more learning occurs between the beginning and the end of the year. This feature was designed to avoid rewarding only the highest achieving students. In 12th grade, however, payments are provided only to students in the top two categories, reflecting the goal that students reach at least the proficient level by the time they graduate.

6.2. **Treatment 2 (Teacher incentives group).** In the teacher incentives group, mathematics teachers were rewarded for the performance of the students they taught during the year. The per-student bonus was 5 percent of the bonus payments in the student schedules, except for the modification that teachers were penalized for students who dropped back in levels between the beginning and end year scores in 10th and 11th grade. The reward attached to the performance of each student is described in Table 9.

6.3. **Treatment 3 (Student and teacher incentives group).**

6.3.1. *Students.* In 10th and 11th grade, students received a reward based on their individual performance and also on the performance of the other students in their mathematics class. The first component was calculated in exactly the same way as in the student incentives group. The second component was calculated as a fixed proportion, one percent, of the total payments earned by classmates. In 12th grade, students received a reward based only on individual performance calculated in exactly the same way as in the student incentives group.

6.3.2. *Mathematics Teachers.* The reward to full-time mathematics teachers was the sum of the total performance payments earned by the students in their classes calculated as in the teacher incentives group and a fixed proportion, 25 percent, of the average full-time equivalent adjusted performance payments earned by the other mathematics teachers (across all grade levels).

6.3.3. *Non-Mathematics Teachers.* Non-mathematics teachers received a payment equal to 25 percent of the school-wide average (full-time equivalent) mathematics teacher performance payment. Payments for part-time teachers were adjusted for their own full-time equivalence status.

6.3.4. *Principals and Associate Principals.* Principals received a cash payment equal to 50 percent of the average full-time equivalent mathematics teacher performance payment. Associate Principals received a cash payment equal to 25 percent of the school-wide average full-time equivalent mathematics teacher performance payment, adjusted for their own full-time equivalence status.

TABLE 8. Schedule of incentive payments to students for their own achievement, in pesos

	End of grade			
	Pre-Basic	Basic	Proficient	Advanced
Start of 10th Grade				
Pre-Basic	0	4,000	9,000	15,000
Basic	0	2,500	7,500	13,500
Proficient	0	0	6,000	12,000
Advanced	0	0	4,500	10,500
Start of 11th Grade				
Pre-Basic	0	4,000	9,000	15,000
Basic	0	0	7,500	13,500
Proficient	0	0	6,000	12,000
Advanced	0	0	4,500	10,500
Start of 12th Grade				
Pre-Basic	0	0	5,000	10,000
Basic	0	0	5,000	10,000
Proficient	0	0	5,000	10,000
Advanced	0	0	5,000	10,000

TABLE 9. Schedule of incentive payments to teachers for student achievement, in pesos

	End of grade			
	Pre-Basic	Basic	Proficient	Advanced
Start of 10th Grade				
Pre-Basic	0	200	450	750
Basic	-125	125	375	675
Proficient	-125	-125	300	600
Advanced	-125	-125	225	525
Start of 11th Grade				
Pre-Basic	0	200	450	750
Basic	-125	0	375	675
Proficient	-125	-125	300	600
Advanced	-125	-125	225	525
Start of 12th Grade				
Pre-Basic	0	0	250	500
Basic	0	0	250	500
Proficient	0	0	250	500
Advanced	0	0	250	500