American Economic Journal: Economic Policy 2018, 10(1): 298–325
https://doi.org/10.1257/pol.20150066

Cheating and Incentives: Learning from a Policy Experiment†

By César Martinelli, Susan W. Parker, Ana Cristina Pérez-Gea, and Rodimiro Rodrigo*

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). We provide evidence suggesting that students may have learned to cheat, with the number of cheating students per classroom increasing over time under treatments that provide monetary incentives to students. (*JEL* D83, I21, I28, O15)

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.

* Martinelli: Department of Economics, George Mason University, 4400 University Drive, MSN 1B2 Fairfax, VA 22015 (email: cmarti33@gmu.edu); Parker: School of Public Policy, University of Maryland, 2101 Van Munching Hall, College Park, MD 20742, and Centro de Investigación y Docencia Económicas (CIDE), Mexico City (email: swparker@umd.edu); Pérez-Gea: Courant Institute of Mathematical Sciences, New York University, 251 Mercer Street, New York, NY 10012 (email: acperezgea@gmail.com); Rodrigo: Department of Economics, Georgetown University, 37th and O Streets, N.W., Washington, DC 20007 (email: rr1041@georgetown.edu). The authors are grateful to two anonymous referees for very helpful comments. Research for this paper was funded by Asociación Mexicana de Cultura while Martinelli was at ITAM.

† Go to https://doi.org/10.1257/pol.20150066 to visit the article page for additional materials and author disclosure statement(s) or to comment in the online discussion forum.

[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).

In this paper, we explore the effects of incentives to students and teachers on cheating using a database 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 at increasing performance, there was some evidence of cheating in the incentivized exams, as described by Behrman et al. (2015). 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.

We start our analysis by identifying cheating students extending methods borrowed from the education measurement literature. These methods rely on the statistical detection of pairs 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 from 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 detect actual cheating versus statistical anomalies, 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.

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 percent to 7.5 percent in different years. For the student incentives group, cheating increases from about 11 percent the first year to about 27 percent the second year and to 30 percent the third year. For the teacher and student incentives group, cheating increases from about 7 percent the first year to about 23.5 percent the second year and to 32 percent the third year. For the teacher only incentives group, per contra, cheating increases from about 7 percent the first year to nearly 10 percent the third year; statistically different from 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 are also 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.

[2] See, e.g., Wollack (1997, 2003, 2006); Wesolowsky (2000); van der Linden and Sotaridona (2006); Zopluoglu (2013); and Romero, Riascos, and Jara (2015).

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 (NCLB) 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 been a growing interest in incentives programs that include incentives to students and teachers linked to performance in standardized tests. Recent work on teacher incentives programs include Muralidharan and Sundararaman's (2011) carried out in rural India; Glewwe, Ilias, and Kremer's (2010) in rural Kenya; and Springer et al.'s (2010) in Nashville, Tennessee public schools. Recent work on student incentives programs include Angrist and Lavy's (2009) study of high school student incentives in Israel; Kremer, Miguel, and Thornton's (2009) of sixth grade girls in Kenya; and Fryer's (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. (2015), who report on the ALI intervention effects on learning, are careful to isolate the effects of cheating on inflating test scores, but do not elaborate on the determinants of cheating, the characteristics of cheating students, or the evolution of cheating over time.

On other unintended consequences of school accountability programs, Reback, Rockoff, and Schwart (2014) show that NCLB had some small effects on increasing student achievement but at the possible cost of reducing incentives for teachers to accept jobs in schools likely to have difficulty meeting targets and with differential effects on tenured vis-à-vis untenured teachers. 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.

Other recent work on cheating has focused on peer effects and the use of external monitors. Carrell, Malmstrom, and West (2008) use self-reported data from US military academies to show that peer dishonesty (as measured by the presence of high school cheaters in the classroom) results in a substantial increase in the probability that students 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 (2015) use a dataset drawn from a national evaluation standardized test in Italy that included random monitoring by external inspectors to show that grades are inflated in the absence of such inspectors.

Our results illustrate what has been dubbed (e.g., by Charness, Masclet, and Villeval 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 incentivized 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 (Bénabou and Tirole 2011, 2006). In this sense, cheating in the classroom resembles illegal activities in the society at large. Glaeser, Sacerdote, and Scheinkman (1996), in their seminal work on social interactions and crime, interpret the high variance of crime rates across US cities as evidence of the importance of social interactions in crime, leading to interdependent decisions. 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.

The remainder of this paper is organized as follows. Section II describes the policy intervention on which our analysis 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 III explores empirically cheating at the student level, focusing on the impact of incentives and experience on the probability of cheating. Section III also explores empirically cheating at the classroom level. Section IV concludes.

I. Incidence of Cheating in the ALI Experiment

A. *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 2008–2009 academic year and ended with the 2010–2011 academic year (Behrman et al. 2015). A total of 88 high schools (*preparatorias*) participated in the experiment; Figure 1 illustrates the location of the schools. The schools were randomly assigned to 4 different groups; 20 schools were assigned to each of 3 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.

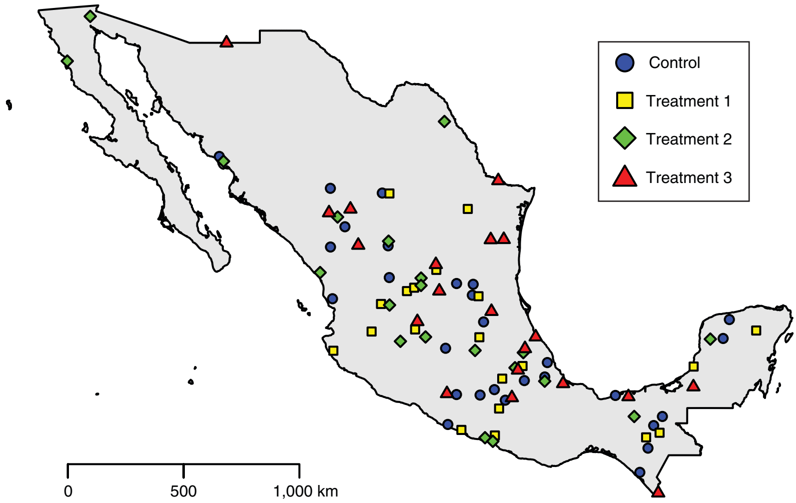

Figure 1. ALI Schools

(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.

As described in detail in Behrman et al. (2015), T3 was designed to align incentives and promote collaboration among all the agents at school including students, teachers, and administrators toward the common goal of improving mathematics achievement. For student payments, T3 differs from T1 in that students in T3 receive a group monetary incentive in addition to the incentive based on their individual-level performance. Similarly for teacher payments, T3 differs from T2 in that teachers in T3 receive a group incentive based on performance in classes taught by other math teachers.

Incentive payments were based on standardized curriculum-based mathematics exams in tenth, eleventh, and twelfth grade given at the end of each academic year. Payments were calculated according to the improvement in mathematics learning over the school year for tenth and eleventh graders and on the final level of learning for twelfth graders. The score on the ninth 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 ninth 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 385 to 1,440 US dollars at the exchange rate at the beginning 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.

Randomization was carried out using a school-based block randomization design where schools were grouped into nine blocks based on school size and the previous school year's graduation rate. Behrman et al. (2015) provide evidence on the quality of the randomization testing for significant differences between the treatment groups and the control group for a number of school-level characteristics and for differential attrition. The statistical tests support that the randomization was successful and that attrition of students was random with respect to the program. We restrict attention to the sample of students in the sample for the three years of the program in order to study cheating of the same individuals over time and carry out our own tests of the quality of randomization and attrition which are provided in Tables A3 and A4 in the Appendix. In Table A3, we estimate the effect of individual and school-level characteristics on treatment status (including all treatments combined and separately) using logit models. In Table A4, we show levels of attrition for each treatment group and for the control group. The tables are overall highly supportive that the randomization was successful and that attrition did not differ significantly between the treatment and control groups, even though there are occasional individual characteristics that appear to be significant among the many characteristics considered.

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 about the test based on past tests.

B. *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, 2006). We focus the analysis on the

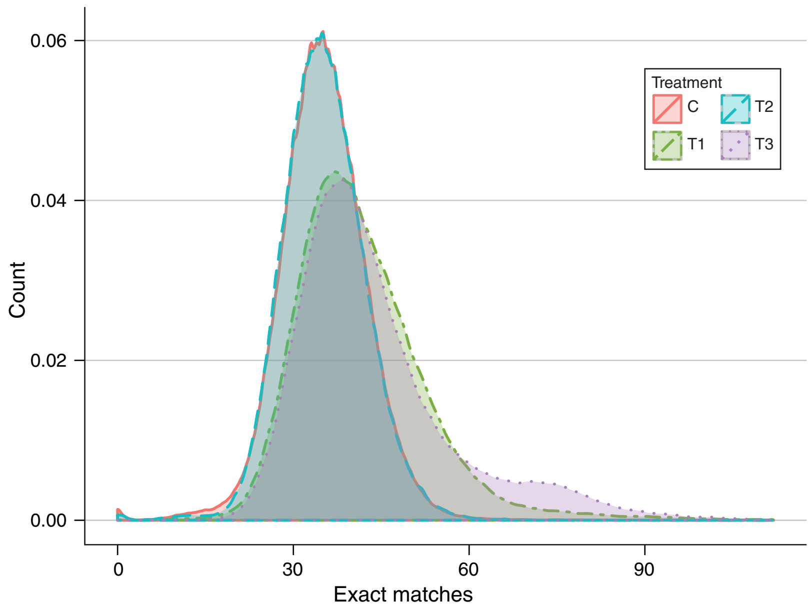

Figure 2. Distribution of Exact Matches for Pairs of Students

ω index and the GBT, which have been judged to perform better in the recent education literature.[3]

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 2 illustrates the distribution of the number of exact matches for sampled pairs of students in the same classroom for each treatment in the twelfth grade exam, the last year of the program (2010–2011). 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.

[3] We explored 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.

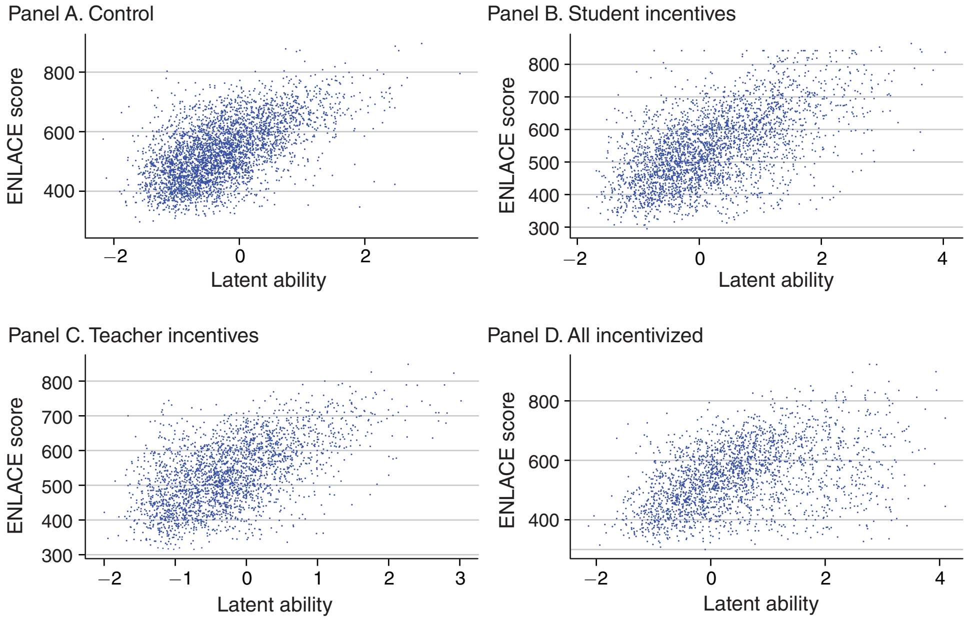

Figure 3. Latent Ability versus Preprogram Score

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), none. 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, 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.

As an illustration, Figure 3 plots latent ability for twelfth grade students in 2010–2011, according to an NRM estimation, against their preprogram, ninth 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.

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

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

Then the index GBT_{ij} is computed as the probability of h_{ij} matches or more given the total number of questions in the exam.

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 devices) likely confined to pairs of students in the same classroom. We tested for cheating with every pair of students in the same school and raised the threshold for accusing a given pair to the point where only 5 percent 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 4 illustrates our procedure for choosing α. In the top figure, we consider twelfth grade students in the teacher and students' incentive group in 2010–2011, that is the cohort that went through the program all three years. 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 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 percent of the percentage of students for relatively high values of α. The two lines are very close for α above 10^{-2}, but they differ sharply for α below 10^{-3}. At $\alpha = 10^{-5}$,

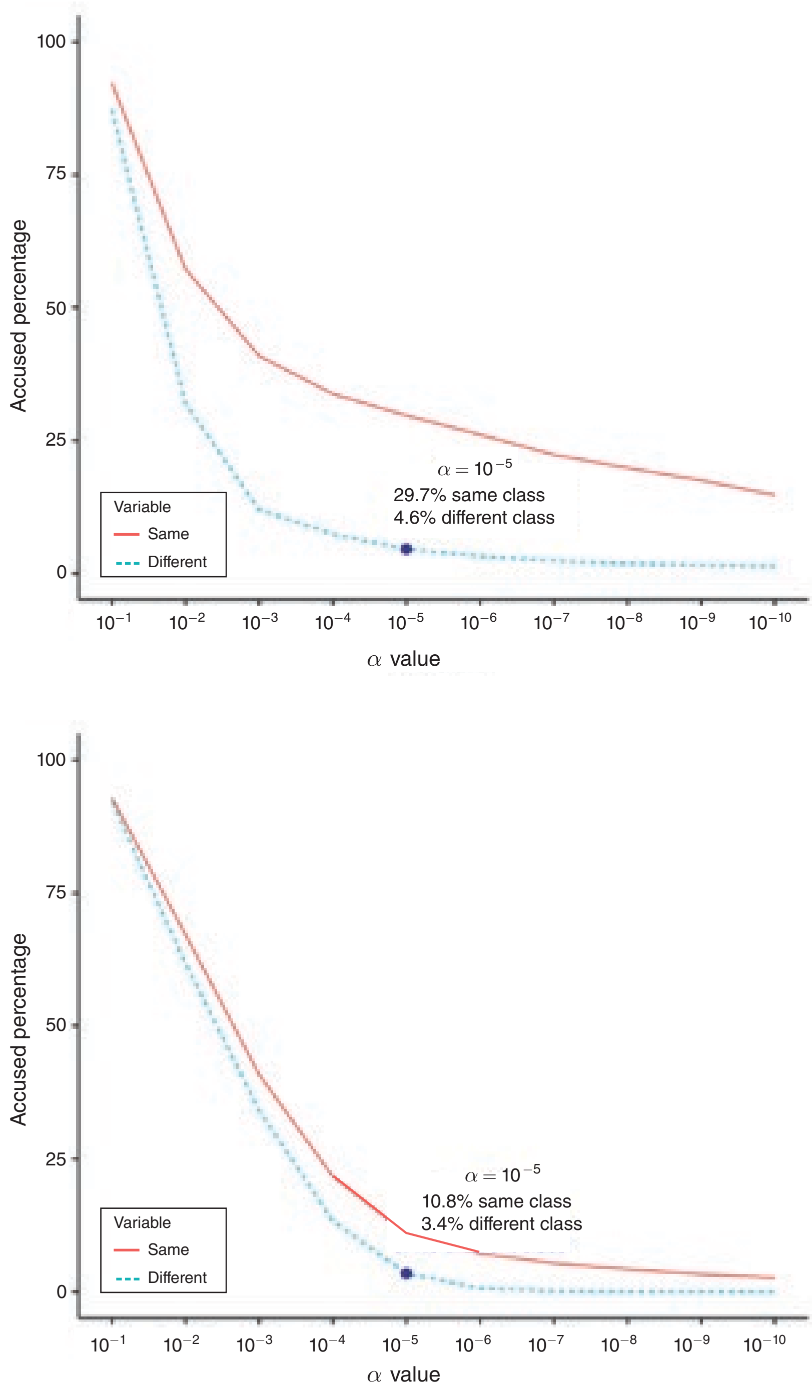

Figure 4. Estimated Probability of Cheating: All Incentivized versus Control Group

the percentage of students accused because an unusual similarity with a student in the same classroom is 29.7 percent, while the percentage of students accused exclusively because an unusual similarity with a student in a different classroom is

4.6 percent. Further lowering α does not help in distinguishing between the two categories. To construct these graphs, we took a sample of students from other classes equal to the size of the own class in order to avoid mechanically finding more false positives in larger schools.

In the bottom figure, we repeat the exercise but consider the control group. As expected, since there are no incentives for cheating, the two lines are much closer. At $\alpha = 10^{-5}$, the percentage of students accused because of an unusual similarity with a student in the same classroom is 10.8 percent, while the percentage of students accused exclusively because of an unusual similarity with a student in a different classroom is 3.4 percent. We take the latter percentage as an approximation to the percentage of students erroneously accused of cheating.

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 reestimated 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 similarly, 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 on using the ω index with a threshold value $\alpha = 10^{-5}$ and calculating item response parameters and individual talent using both the control group and the treatment groups.[4]

The question might be raised of whether students who study together could have similar answer sheets and thus be more likely to be accused of cheating with respect to students who do not study with others (a frequent student defense when accused of copying). The length of the multiple-choice exam that we used and the high threshold for detecting cheating makes the statistical likelihood of false accusations of cheating to students who worked together very unlikely. Furthermore, the difficulty of the exams increases the implausibility due to the high proportion on average of incorrect answers on the exam. Finally, if students working together to improve their probability of getting correct answers were able to explain the majority of the cheating observed, we would also have observed substantial cheating across different classrooms (where cheating is not feasible).

C. *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 through the program for three years. This cohort began in tenth grade in the ALI program in the 2008–2009 year and completed twelfth grade

[4] In Behrman et al. (2015), the method used to identify cheaters was Wesolowsky's (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

	1st year	2nd year	3rd year	
	10th grade	11th grade	12th grade	All years
Panel A. Percentage of students involved in cheating by treatment group				
No incentives	7.223	5.135	7.506	6.622
Student incentives	11.323	26.687	29.693	22.567
Difference in means versus no incentives	−4.099	−21.551	−22.187	−15.946
	(0.227)	(0.318)	(0.362)	(15.430)
Teacher incentives	7.356	8.969	9.599	8.641
Difference in means versus no incentives	−0.133	−3.834	−2.093	−2.020
	(0.180)	(0.167)	(0.232)	(0.089)
All incentivized	7.347	23.510	31.878	20.912
Difference in means versus no incentives	−0.123	−18.375	−24.372	−14.290
	(0.177)	(0.378)	(0.439)	(0.171)
Observations	11,530	11,530	11,530	34,590
Panel B. Test for differences between treatment groups				
Student versus teacher incentives	3.97	17.72	20.09	13.926
	(0.256)	(0.388)	(0.453)	(0.183)
All versus teacher incentives	0.009	14.54	22.28	12.270
	(0.191)	(0.458)	(0.544)	(0.202)
Student versus all incentivized	3.98	3.18	−2.18	1.656
	(0.255)	(0.539)	(0.612)	(0.244)

Panel C. Test for differences between years	1st versus 2nd year	2nd versus 3rd year
Student incentives	15.364	3.006
	(0.389)	(0.497)
Teacher incentives	1.612	0.629
	(0.209)	(0.274)
All incentivized	16.163	8.367
	(0.457)	(0.666)

Note: Standard errors are in parentheses.

in the ALI program in the 2010–2011 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 percent and 7.5 percent, depending on the year. In the student incentives group, the estimated percentage of cheaters increased from about 11.3 percent the first year to nearly 30 percent the third. Similarly, in the group that received incentives for teachers and students, the estimated percentage of cheaters increased from around 7.4 percent to nearly 32 percent. The estimated percentage of cheaters in the teacher incentive group, per contra, barely increased from around 7.4 percent the first year to about 9.6 percent 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 5 percent. 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 percentage of accused individuals for a given treatment and year is z, a conservative estimate can be derived from $z' + 0.05(100 - z') = z$ for $z > 5$ and $z' = 0$ for $z \leq 5$, yielding $z' = \max\{0, (z-5)/0.95\}$. From this conservative viewpoint, cheating in the control group and in the teacher incentives group may have been below 5 percent throughout the program. Cheating in the student incentive group, per contra, went from around 6.6 percent the first year to nearly 23 percent the second year and to nearly 26 percent the third year. Similarly, cheating in the teacher and student incentives group went from close to 2.5 percent to about 19 percent the second year and to about 28 percent the third year. That is, even from this viewpoint, incentives had a very large effect on cheating, and the effect compounded with time in the program. The high level of cheating in spite of monitoring of the exams is perhaps surprising. Unfortunately, we are not able to distinguish between student-level efforts to hide copying and ineffective monitoring due to lax effort of the monitors or altruism toward the students.

Table 1 also provides tests of significant differences between years of cheating levels and by treatment group. These tests confirm that cheating increased significantly between years in the student incentives group and in the all incentivized group, and that significant differences in cheating exist by treatment group in the three years of analysis, particularly in years two and three. The increase in cheating between the first and the second year is particularly stark. There are at least two possible explanations for this increase. First, students in the first year may not have been fully cognizant about the operation of the program nor the likelihood of obtaining the incentives. Second, and related, the monetary incentives promised by the program may not have been completely credible until the first year payments were carried out. There is also a minor, but statistically significant, increase in cheating between the second and the third year in the program.

Note that since T3 students receive more total money, one might expect cheating to be higher than in T1. Cheating might also be higher in T3 through additional collaboration in cheating due to the group incentives present in T3 but not in T1. In fact however, Table 1 shows that levels of cheating are similar in both groups in all three years and so neither the additional payment nor the group incentive payment seem to have much effect over the individual-level incentive payment structure.

II. Empirical Analysis of Individual Cheating Behavior

A. *Descriptive Analysis*

In Table 2, we compare the characteristics of cheaters with those of other students in the cohort that were three years in the program, from tenth to twelfth grade, where we define as a cheater any student who was a member of at least one cheating pair in any year through 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 on average a higher baseline score (i.e., the ENLACE ninth grade math score). Overall, this initial picture suggests that those engaging in cheating have a somewhat more privileged family background and greater social networks.

Table 2—Cheaters versus Non-cheaters: Descriptive Analysis

	Cheaters	Non-cheaters	Difference	All students
Age	15.265	15.365	−0.101	15.335
	(0.867)	(0.891)	(0.018)	(0.885)
Gender (female = 1)	0.483	0.474	0.009	0.477
	(0.500)	(0.499)	(0.010)	(0.499)
Family members	4.660	4.900	−0.242	4.820
	(2.200)	(2.357)	(0.052)	(2.310)
Monthly household income: 2,000 to 4,000 pesos	0.322	0.306	0.016	0.311
	(0.467)	(0.461)	(0.010)	(0.463)
Monthly household income: 4,000 to 8,000 pesos	0.169	0.137	0.032	0.147
	(0.374)	(0.343)	(0.008)	(0.354)
Monthly household income: More than 8,000 pesos	0.110	0.071	0.039	0.083
	(0.313)	(0.256)	(0.006)	(0.276)
Have internet	0.204	0.128	0.076	0.152
	(0.403)	(0.334)	(0.008)	(0.359)
Three or four books at home	0.283	0.248	0.035	0.259
	(0.451)	(0.432)	(0.010)	(0.438)
Five or more books at home	0.119	0.084	0.035	0.095
	(0.324)	(0.278)	(0.007)	(0.294)
Mother with secondary education	0.313	0.318	−0.005	0.316
	(0.464)	(0.466)	(0.010)	(0.465)
Mother with high school or technical education	0.208	0.155	0.053	0.171
	(0.406)	(0.362)	(0.008)	(0.377)
Mother with college or more	0.120	0.071	0.049	0.086
	(0.324)	(0.256)	(0.006)	(0.280)
Father with secondary education	0.284	0.291	−0.007	0.289
	(0.451)	(0.454)	(0.010)	(0.453)
Father with high school or technical education	0.203	0.172	0.030	0.182
	(0.402)	(0.378)	(0.009)	(0.386)
Father with college or more	0.185	0.103	0.082	0.129
	(0.279)	(0.304)	(0.007)	(0.335)
I have a positive attitude[a]	0.914	0.898	0.0167	0.903
	(0.279)	(0.303)	(0.007)	(0.296)
I believe I am a failure[a]	0.081	0.086	−0.005	0.084
	(0.273)	(0.280)	(0.006)	(0.278)
I am as capable as most people[a]	0.930	0.924	0.005	0.926
	(0.256)	(0.264)	(0.006)	(0.262)
How often hang out with friends[b]	0.090	0.073	0.016	0.078
	(0.286)	(0.260)	(0.006)	(0.269)
Have scholarship	0.284	0.347	−0.063	0.328
	(0.451)	(0.476)	(0.010)	(0.469)
Preprogram math score	0.190	−0.081	0.271	0.000
	(1.068)	(0.958)	(0.021)	(1.000)
ALI math score	0.305	−0.132	0.437	0.000
	(1.065)	(0.941)	(0.020)	(1.000)
Observations	3,471	8,059		11,530

Notes: Cheaters are defined as those students detected cheating in any of the three years of the program. Standard errors are in parentheses.

[a] Dummy variables where 1 means agree and 0 disagree with the statement.

[b] Dummy variables where 1 means few or never and 0 frequently or always.

TABLE 3—EFFECTS OF INCENTIVES AND BASELINE ACHIEVEMENT ON THE PROBABILITY OF CHEATING

	Marginal effects from logit of cheating									OLS on $1 - \Phi(\omega)$
	10th graders			11th graders			12th graders			12th graders
Student incentives	0.038	0.031	0.025	0.232	0.221	0.213	0.232	0.223	0.216	0.015
	(0.027)	(0.025)	(0.025)	(0.051)	(0.041)	(0.038)	(0.057)	(0.047)	(0.043)	(0.004)
Teacher incentives	0.001	−0.001	−0.002	0.073	0.073	0.073	0.038	0.037	0.034	0.005
	(0.028)	(0.026)	(0.026)	(0.036)	(0.034)	(0.035)	(0.047)	(0.045)	(0.045)	(0.004)
All incentivized	0.001	−0.006	−0.007	0.212	0.201	0.204	0.247	0.238	0.240	0.017
	(0.026)	(0.024)	(0.0240)	(0.050)	(0.047)	(0.047)	(0.052)	(0.050)	(0.049)	(0.003)
Baseline score		0.018	0.004		0.033	0.029		0.024	0.014	0
		(0.004)	(0.006)		(0.009)	(0.009)		(0.012)	(0.013)	(0.002)
Score × Student incentives			0.028			0.025			0.036	0.005
			(0.008)			(0.016)			(0.025)	(0.002)
Score × Teacher incentives			0.007			−0.008			0.027	0.005
			(0.012)			(0.014)			(0.018)	(0.002)
Score × All incentivized			0.014			−0.016			−0.023	0.003
			(0.009)			(0.017)			(0.019)	(0.002)
Controls	No	Yes	Yes	No	Yes	Yes	No	Yes	Yes	Yes
Observations	11,530	11,530	11,530	11,530	11,530	11,530	11,530	11,530	11,530	11,530

Notes: Standard errors are in parentheses. Controls include: age; age squared; gender; family members; monthly household income; have internet; books at home; mother with secondary education, upper secondary or technical education, college, or more; father with secondary education, upper secondary or technical education, college, or more; I have a positive attitude; I believe I am a failure; I am at least as capable as most people; how often hang out with friends; have scholarship; and preprogram math score.

B. *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 and on the probability of the student being detected as a cheater, obtained from logit regressions.[5] The marginal effects are taken with respect to the control group (no incentives). We restrict our analysis to students who remain in the cohort of reference the three years of the program providing results for this cohort as it went through tenth, eleventh, and twelfth grade. We provide three specifications for each year, first a specification with no control variables, second a specification with our standard set of control variables, and finally, an additional specification, which includes interaction terms between the baseline score and the different incentive treatments. Finally, we add a column which rather than using a threshold α and assigning individuals to be cheaters or non-cheaters, uses the (continuous) level of the minimum $1 - \Phi(\omega)$ index (that is the probability that the similarity with other students happened without communication) for each individual. This last specification is provided to test that the relationship of the explanatory variables studied here is not contingent on the threshold chosen.

[5] Similar results are obtained using linear probability and probit models, as reported in Tables A5 and A6 in the Appendix. As another robustness check, we vary the level of α used to classify individuals in Tables A7 and A8, with similar results.

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 (statistically insignificant) 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 monetary incentives to copy (see Table A1 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 some evidence for the interaction of the score with student incentives. This last result is consistent with the hypothesis that better students are drafted as sources when incentives are high. Using the actual level of the continuous index (last column) instead of a classification based on the omega index into cheaters versus non-cheaters shows results consistent with those based on using a dichotomous measure of cheating. The coefficient on 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 cheating, e.g., if monitors were to report a student to the school administration.

The regression of $1 - \Phi(\omega)$ on dummies corresponding to the different treatments carries a similar message; student incentives drive a larger suspicious similarity between students than teacher incentives. Interaction terms of the treatment and the baseline score are very small, even if some appear to be significant. In the first year of the program, when there is less experience cheating, some personal characteristics among the controls in Table 3 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.

An interesting question is how much money was actually gained by cheating. To make such a calculation, we must make some assumptions about how much of the program impacts on achievement are due to cheating. (Behrman et al. 2015 carry out such an analysis to identify real program effects on learning.) Assuming that all those we identify as cheating in fact had no real program impacts on achievement, we calculate the difference between what cheaters earned versus what they would have earned if cheating had been impeded. We find that for T3 and T1, respectively, 72 percent and 65 percent of incentive payments actually paid to cheaters would not have been paid.

Table 4—Persistence of Cheating: Probability of Cheating as a Function of Cheating in the Previous Year

	Marginal effects from logit of cheating		OLS on $1 - \Phi(\omega)$
	11th graders	12th graders	12th graders
Student incentives	0.223	0.169	0.013
	(0.038)	(0.036)	(0.004)
Teacher incentives	0.078	0.035	0.004
	(0.032)	(0.041)	(0.004)
Student and teacher incentives	0.210	0.184	0.014
	(0.045)	(0.042)	(0.003)
Cheating prior school year	0.209	0.225	0.018
	(0.026)	(0.033)	(0.006)
Cheating prior school year × Student incentives	−0.059	−0.040	−0.007
	(0.041)	(0.041)	(0.006)
Cheating prior school year × Teacher incentives	−0.021	−0.071	−0.006
	(0.040)	−0.042	(0.007)
Cheating prior school year × All incentivized	−0.047	−0.014	−0.002
	(0.038)	(0.054)	(0.006)
Controls	Yes	Yes	Yes
Observations	11,530	11,530	11,530

Notes: Standard errors are in parentheses. The controls are the same as in Table 3.

C. *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 cheating the previous year for the second and third year in the program of the cohort that began in tenth grade in the ALI program in the 2008–2009 year. We include as controls the same personal and family characteristics from Table 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. (Similar results are obtained using linear probability and probit models, as reported in Table A9 in the Appendix.) We include in the estimation interaction terms between cheating the previous year and incentive treatments; these generally come up as negative, reflecting a much larger increase in cheating among students who did not previously cheat in incentivized treatments than in the control group. In the third column, we again use the continuous index value rather than a dichotomous measure of cheating with similar effects of program participation and correlations with previous participation on cheating.

D. *Cheating by School and Classroom*

We change our perspective now from a focus on individual students to a focus on the classroom. Table 5 provides information about the number of classrooms and (in parentheses) 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.

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)

Figures 5 illustrates the percentage of cheaters by classroom for the control group and the different treatment groups for the cohort of interest in the last year in the program. Each individual bar represents a classroom. Classrooms are grouped by school, and schools are separated by dotted lines. Schools are ordered from left to right in decreasing order with respect to the prevalence of cheating; classrooms within each school are ordered with the same criterion. 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-level data, there is more cheating in classrooms under the student incentive group and the teacher and student incentive group.

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 varies from being more than 50 percent 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 varies from more than 75 percent in one classroom to zero in another. Overall, the data suggest the existence of a "cheating culture" in a few schools and in some classrooms.

We also look at the network defined by pairs of students classified as cheaters by the ω index within each classroom.[6] 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 seems to have given rise to extensive collaboration in illicit communication in these classrooms.

III. 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

[6] Details are available from the authors.

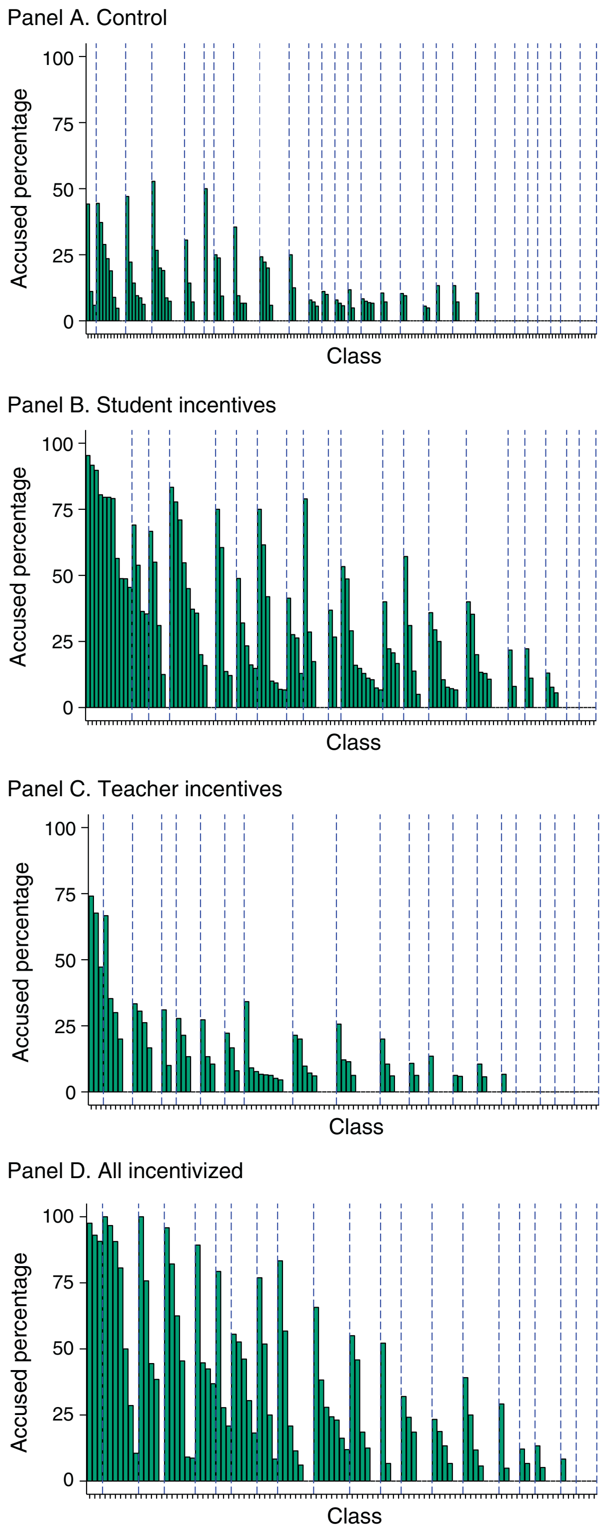

Figure 5. Percentage of Cheaters by Classroom and School in 12th Grade

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 the manipulation of test scores via "teaching to the test" (Kremer, Miguel, and Thornton 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 has 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 5 percent, 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. By the third year in the program, between 20 percent and 30 percent of students may have engaged in cheating in treatments that provided financial incentives to students, depending on assumptions regarding classification errors, 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 little detected cheating in every treatment.

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 crowd 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 (Gneezy and Rustichini 2000; Gneezy, Meier, and Rey-Biel 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.

Finally, while this paper has emphasized that groups where student incentives were paid displayed much higher levels of student copying than the group where only teacher incentives were paid, it is also the case that student incentive groups had both much higher levels of impacts on student achievement (net of cheating) and higher realized incentive payments per student (Behrman et al. 2015). Because the teachers only incentives group costs very little compared with the student incentive payments (and has very little effect on learning), the average cheating per peso of incentive payments paid is actually higher in the teacher only incentives. Our results should not be misinterpreted to imply that the existence of student-level copying is a reason to favor incentives to teachers over incentives to students.

Appendix

In this Appendix, we elaborate on the structure of incentive payments associated to ALI and present additional tables.

A. *Treatment 1* (*Student Incentives Group*)

Table A1 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 years 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 tenth and eleventh 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 twelfth 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.

B. *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 tenth and eleventh grade. The reward attached to the performance of each student is described in Table A2.

Table A1—Schedule of Incentive Payments to Students for Their Own Achievement, in Pesos

	End of grade			
	Pre-basic	Basic	Proficient	Advanced
Start of tenth 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 eleventh 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 twelfth 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 A2—Schedule of Incentive Payments to Teachers for Student Achievement, in Pesos

	End of grade			
	Pre-basic	Basic	Proficient	Advanced
Start of tenth grade				
Pre-basic	0	200	450	750
Basic	−125	125	375	675
Proficient	−125	−125	300	600
Advanced	−125	−125	225	525
Start of eleventh grade				
Pre-basic	0	200	450	750
Basic	−125	0	375	675
Proficient	−125	−125	300	600
Advanced	−125	−125	225	525
Start of twelfth grade				
Pre-basic	0	0	250	500
Basic	0	0	250	500
Proficient	0	0	250	500
Advanced	0	0	250	500

C. *Treatment 3* (*Student and Teacher Incentives Group*)

Students.—In tenth and eleventh 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, 1 percent, of the total payments earned by classmates. In twelfth grade, students received a reward based only on individual performance calculated in exactly the same way as in the student incentives group.

Table A3—Marginal Effects from Logit of Treatment Status versus Control

	All treatments	Student incentives	Teacher incentives	All incentivized
Age	0.009	−0.032	0.070	0.029
	(0.034)	(0.048)	(0.045)	(0.048)
Age squared	−0.004	−0.001	−0.010	−0.005
	(0.004)	(0.006)	(0.005)	(0.006)
Gender (female = 1)	0.000	0.002	−0.007	0.001
	(0.009)	(0.013)	(0.011)	(0.012)
Family members	0.008	0.009	0.009	0.003
	(0.004)	(0.006)	(0.004)	(0.006)
Monthly household income: 2,000 to 4,000 pesos	0.009	0.020	0.034	−0.008
	(0.020)	(0.028)	(0.025)	(0.027)
Monthly household income: 4,000 to 8,000 pesos	−0.010	0.007	−0.001	−0.013
	(0.028)	(0.038)	(0.038)	(0.040)
Monthly household income: More than 8,000 pesos	0.046	0.053	0.073	0.047
	(0.028)	(0.033)	(0.040)	(0.039)
Have internet	0.040	0.052	0.036	0.076
	(0.030)	(0.033)	(0.045)	(0.055)
Books at home (3–4)	0.019	0.063	−0.024	0.009
	(0.019)	(0.023)	(0.019)	(0.024)
Books at home (5 or more)	0.002	0.054	−0.055	−0.010
	(0.025)	(0.032)	(0.034)	(0.034)
Mother with secondary education	0.030	0.022	0.051	0.030
	(0.029)	(0.038)	(0.036)	(0.029)
Mother with high school or technical education	0.060	0.059	0.071	0.057
	(0.029)	(0.038)	(0.036)	(0.029)
Mother with college or more	−0.034	−0.038	−0.013	−0.068
	(0.040)	(0.054)	(0.051)	(0.042)
Father with secondary education	0.051	0.048	0.043	0.067
	(0.017)	(0.023)	(0.027)	(0.021)
Father with high school or technical education	−0.006	−0.015	−0.013	0.010
	(0.021)	(0.030)	(0.035)	(0.028)
Father with college or more	−0.019	0.008	−0.060	−0.012
	(0.031)	(0.038)	(0.033)	(0.049)
I have a positive attitude	−0.010	0.013	−0.024	−0.029
	(0.018)	(0.024)	(0.023)	(0.018)
I believe I am a failure	0.000	0.000	−0.011	−0.001
	(0.018)	(0.028)	(0.025)	(0.020)
I am as capable as most people	−0.021	0.010	−0.037	−0.046
	(0.02)	(0.025)	(0.027)	(0.022)
How often hang out with friends	−0.018	0.002	−0.025	−0.037
	(0.025)	(0.030)	(0.024)	(0.030)
Have scholarship	−0.036	−0.020	−0.057	−0.055
	(0.021)	(0.029)	(0.027)	(0.029)
Preprogram math score	0.002	−0.009	−0.014	0.019
	(0.015)	(0.022)	(0.014)	(0.024)
Total school enrollment	0.017	0.037	0.029	−0.021
	(0.032)	(0.041)	(0.048)	(0.045)
Number of teachers	0.002	0.008	0.007	−0.008
	(0.007)	(0.010)	(0.012)	(0.011)
Number of laboratories	0.012	−0.049	0.077	0.045
	(0.044)	(0.063)	(0.054)	(0.065)
Number of students failing between 1 and 5 classes	0.000	−0.001	−0.001	−0.001
	(0.001)	(0.001)	(0.001)	(0.001)
Observations	11,530	6,538	6,086	5,994

Note: Robust standard errors are in parentheses.

Table A4—Probability of Attriting during the Three Year ALI Experiment

Student incentives	−0.038	−0.036	−0.035	−0.030
	(0.033)	(0.029)	(0.028)	(0.026)
Teacher incentives	−0.000	0.001	−0.003	−0.001
	(0.031)	(0.024)	(0.024)	(0.023)
All incentivized	−0.032	−0.032	−0.036	−0.040
	(0.038)	(0.031)	(0.030)	(0.028)
Individual controls	No	Yes	Yes	Yes
Household-level controls	No	No	Yes	Yes
School controls	No	No	No	Yes
Observations	22,738	22,738	22,738	22,738

Note: Robust standard errors are in parentheses.

Table A5—Probability of Cheating: Linear Probability Models

	10th graders			11th graders			12th graders		
Student incentives	0.041	0.035	0.035	0.216	0.207	0.201	0.222	0.215	0.212
	(0.030)	(0.028)	(0.027)	(0.054)	(0.045)	(0.039)	(0.062)	(0.053)	(0.047)
Teacher incentives	0.001	−0.002	−0.002	0.039	0.038	0.038	0.021	0.020	0.021
	(0.025)	(0.024)	(0.024)	(0.019)	(0.019)	(0.019)	(0.027)	(0.025)	(0.025)
All incentivized	0.001	−0.006	−0.005	0.184	0.175	0.177	0.244	0.237	0.241
	(0.023)	(0.021)	(0.022)	(0.051)	(0.049)	(0.049)	(0.059)	(0.059)	(0.058)
Baseline score			0.003			0.009			0.005
			(0.005)			(0.003)			(0.007)
Score × Student incentives			0.042			0.082			0.073
			(0.011)			(0.028)			(0.037)
Score × Teacher incentives			0.007			0.005			0.021
			(0.012)			(0.008)			(0.011)
Score × All incentivized			0.012			0.013			−0.016
			(0.010)			(0.023)			(0.023)
Controls	No	Yes	Yes	No	Yes	Yes	No	Yes	Yes
Observations	11,530	11,530	11,530	11,530	11,530	11,530	11,530	11,530	11,530

Notes: Robust standard errors are in parentheses. The controls are the same as in Table 3.

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).

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.

Principals and Associate Principals.—Principals received a cash payment equal to 50 percent of the average full-time equivalent mathematics teacher performance

Table A6—Probability of Cheating: Probit Marginal Effects

	10th graders			11th graders			12th graders		
Student incentives	0.038	0.031	0.026	0.221	0.210	0.202	0.224	0.215	0.209
	(0.027)	(0.025)	(0.025)	(0.048)	(0.039)	(0.035)	(0.055)	(0.045)	(0.041)
Teacher incentives	0.001	−0.002	−0.003	0.063	0.064	0.063	0.033	0.032	0.030
	(0.027)	(0.025)	(0.025)	(0.031)	(0.030)	(0.030)	(0.041)	(0.039)	(0.039)
All incentivized	0.001	−0.006	−0.007	0.199	0.190	0.191	0.239	0.231	0.233
	(0.025)	(0.023)	(0.023)	(0.047)	(0.044)	(0.044)	(0.050)	(0.049)	(0.048)
Baseline score			0.003			0.024			0.011
			(0.005)			(0.007)			(0.011)
Score × Student incentives			0.031			0.033			0.041
			(0.008)			(0.016)			(0.026)
Score × Teacher incentives			0.007			−0.005			0.025
			(0.011)			(0.012)			(0.015)
Score × All incentivized			0.013			−0.012			−0.021
			(0.009)			(0.017)			(0.019)
Controls	No	Yes	Yes	No	Yes	Yes	No	Yes	Yes
Observations	11,530	11,530	11,530	11,530	11,530	11,530	11,530	11,530	11,530

Notes: Standard errors are in parentheses. The controls are the same as in Table 3.

Table A7—Probability of Cheating: Marginal Logit Effects Using $\alpha = 0.0001$

	10th graders			11th graders			12th graders		
Student incentives	0.046	0.038	0.033	0.237	0.225	0.213	0.243	0.233	0.226
	(0.033)	(0.030)	(0.030)	(0.051)	(0.040)	(0.037)	(0.057)	(0.047)	(0.043)
Teacher incentives	−0.000	−0.003	−0.004	0.068	0.068	0.066	0.036	0.036	0.033
	(0.035)	(0.033)	(0.032)	(0.037)	(0.036)	(0.036)	(0.048)	(0.046)	(0.046)
All incentivized	0.012	0.004	0.002	0.226	0.214	0.214	0.257	0.248	0.250
	(0.033)	(0.031)	(0.030)	(0.050)	(0.047)	(0.046)	(0.055)	(0.053)	(0.052)
Standardized math score			0.008			0.021			0.020
			(0.008)			(0.008)			(0.011)
Score × Student incentives			0.031			0.046			0.037
			(0.010)			(0.016)			(0.023)
Score × Teacher incentives			0.009			0.003			0.025
			(0.012)			(0.014)			(0.017)
Score × All incentivized			0.015			−0.001			−0.029
			(0.012)			(0.017)			(0.018)
Controls	No	Yes	Yes	No	Yes	Yes	No	Yes	Yes
Observations	11,530	11,530	11,530	11,530	11,530	11,530	11,530	11,530	11,530

Notes: Standard errors are in parentheses. The controls are the same as in Table 3.

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 A8—Probability of Cheating: Marginal Logit Effects Using $\alpha = 0.001$

	10th graders			11th graders			12th graders		
Student incentives	0.051	0.042	0.036	0.243	0.234	0.225	0.255	0.246	0.239
	(0.038)	(0.035)	(0.035)	(0.049)	(0.039)	(0.036)	(0.057)	(0.049)	(0.045)
Teacher incentives	0.007	0.004	0.003	0.064	0.062	0.062	0.062	0.062	0.058
	(0.040)	(0.038)	(0.037)	(0.035)	(0.034)	(0.034)	(0.046)	(0.044)	(0.043)
All incentivized	0.009	0.000	0.000	0.227	0.215	0.215	0.280	0.271	0.272
	(0.038)	(0.035)	(0.035)	(0.050)	(0.046)	(0.046)	(0.054)	(0.053)	(0.051)
Baseline score			0.010			0.031			0.022
			(0.008)			(0.008)			(0.009)
Score × Student incentives			0.041			0.049			0.048
			(0.012)			(0.018)			(0.024)
Score × Teacher incentives			0.012			0.003			0.039
			(0.016)			(0.016)			(0.018)
Score × All incentivized			0.014			−0.004			−0.025
			(0.017)			(0.019)			(0.020)
Controls	No	Yes	Yes	No	Yes	Yes	No	Yes	Yes
Observations	11,530	11,530	11,530	11,530	11,530	11,530	11,530	11,530	11,530

Notes: Standard errors are in parentheses. The controls are the same as in Table 3.

Table A9—Persistence of Cheating Behavior: Linear Probability and Probit Models

	Linear probability model		Probit model	
	11th graders	12th graders	11th graders	12th graders
Student incentives	0.185	0.140	0.207	0.159
	(0.039)	(0.034)	(0.035)	(0.034)
Teacher incentives	0.035	0.018	0.067	0.030
	(0.015)	(0.022)	(0.027)	(0.035)
All incentivized	0.165	0.161	0.194	0.174
	(0.046)	(0.045)	(0.042)	(0.041)
Cheating behavior in the prior school year	0.166	0.224	0.193	0.218
	(0.028)	(0.052)	(0.023)	(0.034)
Cheating prior school year × Student incentives	0.147	0.107	−0.026	−0.016
	(0.074)	(0.078)	(0.043)	(0.044)
Cheating prior school year × Teacher incentives	0.054	−0.074	−0.009	−0.069
	(0.057)	(0.069)	(0.039)	(0.043)
Cheating prior school year × All incentivized	0.157	0.159	−0.017	0.012
	(0.059)	(0.090)	(0.038)	(0.056)
Controls	Yes	Yes	Yes	Yes
Observations	11,530	11,530	11,530	11,530

Notes: Standard errors are in parentheses. The controls are the same as in Table 3.

REFERENCES

Angrist, Joshua, and Victor Lavy. 2009. "The Effects of High Stakes High School Achievements Awards: Evidence from a Randomized Trial." *American Economic Review* 99 (4): 1384–1414.

Ballester, Coralio, Antoni Calvó-Armengol, and Yves Zenou. 2010. "Delinquent Networks." *Journal of the European Economic Association* 8 (1): 34–61.

Behrman, Jere R., Susan W. Parker, Petra E. Todd, and Kenneth I. Wolpin. 2015. "Aligning Learning Incentives of Students and Teachers: Results from a Social Experiment in Mexican High Schools." *Journal of Political Economy* 123 (2): 325–64.

Bénabou, Roland, and Jean Tirole. 2006. "Incentives and Prosocial Behavior." *American Economic Review* 96 (5): 1652–78.

Bénabou, Roland, and Jean Tirole. 2011. "Laws and Norms." National Bureau of Economic Research (NBER) Working Paper 17579.

Bock, R. Darrell. 1972. "Estimating item parameters and latent ability when responses are scored in two or more nominal categories." *Psychometrika* 37 (1): 29–51.

Brandes, Barbara. 1986. *Academic Honesty: A Special Study of California Students*. Sacramento: California State Department of Education Bureau of Publications.

Calvó-Armengol, Antoni, and Yves Zenou. 2004. "Social Networks and Crime Decisions: The Role of Social Structure in Facilitating Delinquent Behavior." *International Economic Review* 45 (3): 939–58.

Carrell, Scott E., Frederick V. Malmstrom, and James E.West. 2008. "Peer Effects in Academic Cheating." *Journal of Human Resources* 43 (1): 173–207.

Charness, Gary, David Masclet, and Marie-Claire Villeval. 2013. "The Dark Side of Competition for Status." *Management Science* 60 (1): 38–55.

Cizek, Gregory J. 1999. *Cheating on Tests: How to Do It, Detect It and Prevent It*. New York: Lawrence Erlbaum Associates.

Davis, Stephen F., Cathy A. Grover, Angela H. Becker, and Loretta N. McGregor. 1992. "Academic Dishonesty: Prevalence, Determinants, Techniques, and Punishments." *Teaching of Psychology* 19 (1): 16–20.

Davis, Stephen F., and H. Wayne Ludvigson. 1995. "Additional Data on Academic Dishonesty and a Proposal for Remediation." *Teaching of Psychology* 22 (2): 119–21.

Figlio, David, and Joshua Winicki. 2005. "Food for thought: The effects of school accountability plans on school nutrition." *Journal of Public Economics* 89 (2–3): 381–94.

Fryer, Roland G., Jr. 2011. "Financial Incentives and Student Achievement: Evidence from Randomized Trials." *Quarterly Journal of Economics* 126 (4): 1755–98.

Glaeser, Edward L., Bruce Sacerdote, and José A. Scheinkman. 1996. "Crime and Social Interactions." *Quarterly Journal of Economics* 111 (2): 507–48.

Glewwe, Paul, Nauman Ilias, and Michael Kremer. 2010. "Teacher Incentives." *American Economic Journal: Applied Economics* 2 (3): 205–27.

Gneezy, Uri, Stephan Meier, and Pedro Rey-Biel. 2011. "When and Why Incentives (Don't) Work to Modify Behavior." *Journal of Economic Perspectives* 25 (4): 191–210.

Gneezy, Uri, and Aldo Rustichini. 2000. "Pay Enough or Don't Pay at All." *Quarterly Journal of Economics* 115 (3): 791–810.

Holland, Paul W. 1996. "Assessing Unusual Agreement between the Incorrect Answers of Two Examinees: Using the K-index." Education Testing Service Report PSRTR-96-04.

Jacob, Brian A., and Steven D. Levitt. 2003. "Rotten Apples: An Investigation of the Prevalence and Predictors of Teacher Cheating." *Quarterly Journal of Economics* 118 (3): 843–77.

Kremer, Michael, Edward Miguel, and Rebecca Thornton. 2009. "Incentives to Learn." *Review of Economics and Statistics* 91 (3): 437–56.

Lucifora, Claudio, and Marco Tonello. 2015. "Cheating and social interactions: Evidence from a randomized experiment in a national evaluation program." *Journal of Economic Behavior and Organization* 115: 45–66.

Martinelli, César, Susan W. Parker, Ana Cristina Pérez-Gea, and Rodimiro Rodrigo. 2018. "Cheating and Incentives: Learning from a Policy Experiment: Dataset." *American Economic Journal: Economic Policy*. https://doi.org/10.1257/pol.20150066.

Muralidharan, Karthik, and Venkatesh Sundararaman. 2011. "Teacher Performance Pay: Experimental Evidence from India." *Journal of Political Economy* 119 (1): 39–77.

Reback, Randall, Jonah Rockoff, and Heather L. Schwart. 2014. "Under Pressure: Job Security, Resource Allocation, and Productivity in Schools under No Child Left Behind." *American Economic Journal: Economic Policy* 6 (3): 207–41.

Romero, Mauricio, Álvaro Riascos, and Diego Jara. 2015. "On the Optimality of Answer-Copying Indices." *Journal of Educational and Behavioral Statistics* 40 (5): 435–53.

Schab, Fred. 1991. "Schooling without learning: Thirty years of cheating in high school." *Adolescence* 26 (104): 839–47.

Sotaridona, Leonardo S., and Rob R. Meijer. 2002. "Statistical Properties of the K-index for Detecting Answer Copying." *Journal of Educational Measurement* 39 (2): 115–32.

Sotaridona, Leonardo S., and Rob R. Meijer. 2003. "Two New Statistics to Detect Answer Copying." *Journal of Educational Measurement* 40 (1): 53–69.

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

van der Linden, Wim J., and Leonardo S. Sotaridona. 2006. "Detecting answer copying when the regular response process follows a known response model." *Journal of Educational and Behavioral Statistics* 31 (3): 283–304.

Wesolowsky, George O. 2000. "Detecting excessive similarity in answers on multiple choice exams." *Journal of Applied Statistics* 27 (7): 909–21.

Wollack, James A. 1997. "A Nominal Response Model Approach for Detecting Answer Copying." *Applied Psychological Measurement* 21: 307–20.

Wollack, James A. 2003. "Comparison of Answer Copying Indices with Real Data." *Journal of Educational Measurement* 40 (3): 189–205.

Wollack, James A. 2006. "Simultaneous Use of Multiple Answer Copying Indexes to Improve Detection Rates." *Applied Measurement in Education* 19 (4): 265–88.

Zopluoglu, Cengiz. 2013. "CopyDetect: An R Package to Compute Statistical Indices for Detecting Answer Copying on Multiple-Choice Tests." *Applied Psychological Measurement* 37 (1): 93–95.

Zopluoglu, Cengiz, and Ernest C. Davenport. 2012. "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.