Economica LSE

Economica (2020) **87**, 885–900
doi:10.1111/ecca.12331

Catching Cheating Students

By MING-JEN LIN† and STEVEN D. LEVITT‡

†*National Taiwan University* ‡*University of Chicago*

Final version received 12 November 2019.

We develop a simple algorithm for detecting exam cheating between students who copy off one another's exams. When this algorithm is applied to exams in a general science course at a top university, we find strong evidence of cheating by at least 10% of the students. Students studying together cannot explain our findings. Matching incorrect answers proves to be a stronger indicator of cheating than matching correct answers. When seating locations are randomly assigned, and monitoring is increased, cheating virtually disappears.

INTRODUCTION

Student cheating is a perennial issue. In recent years, 70 students in a New York City top public high school were caught using smart phones to cheat on state exams (Baker 2012), and cheating scandals have rocked Harvard, Stanford, Dartmouth and the Air Force Academy, to name just a few.[1]

These well-publicized scandals are only the tip of the iceberg. McCabe (2005) surveyed 8000 college students in the USA and Canada, finding that 11% of them admit to 'copying from another student on an exam without their knowledge', 10% say they have 'helped someone else cheat on test', and 9% acknowledge copying from another student 'with their knowledge'.

This paper develops an algorithm for identifying cases of students copying off one another's exam answers. We test this algorithm using data from a course taught at a top university in which the professor suspected cheating may have occurred. We find compelling evidence of cheating involving at least 10% of the class's 242 students on a midterm exam. When seating positions were randomly assigned and monitoring was increased for the final exam, almost all evidence of cheating disappeared. We are able to rule out that the observed correlations in answers across students who voluntarily sit next to each other is due to studying together, as opposed to cheating on the exam, because of an unusual experiment carried out in advance of the final exam. Students seated themselves voluntarily, with the expectation that the seats that they chose would be the ones in which they would take the exam. These seating choices were recorded. Prior to the actual test, however, students were reassigned to different seats. Thus we are able to observe the patterns in correlations among students who wanted to sit together, but then were not allowed to.

In spite of this apparent widespread cheating, there has been little academic attention devoted to the detection of cheating.[2] Zitzewitz (2012) surveys more than 100 pages in the emerging field of forensic economics, not one of which addresses student cheating.[3] In the statistics discipline, this issue has received more attention. Wesolowsky (2000) develops a method similar to ours, but there are a number of important differences. Wesolowsky's method ignores a key aspect of our approach (differentiating between matching answers that are correct and matching answers that are incorrect—empirically we find that matching incorrect answers are much more informative), estimates a much more highly parametrized functional form, and does not provide an easy metric for

judging how unusual an individual pair of students' answers is. Wesolowsky also does not intervene on seating choices, and so does not observe the randomized seating counterfactual that we have in our analysis.

The remainder of the paper is structured as follows. Section I describes the background of the cheating incident that we analyse. Section II presents a simple reduced-form regression analysis of the cheating patterns. Section III develops and implements a more systematic algorithmic approach to the problem, and also considers alternative explanations for correlations like studying together. Section IV concludes.

I. The Cheating Incident and the Data

In spring 2012, 242 students registered in an introductory natural sciences course at a top US university. The course had three midterms and a final exam.[4] All of the exams were multiple choice, with four possible answers per question. Students recorded their answers on Scantron sheets. There was no punishment for incorrect guesses; that is, a wrong answer yielded no points, as did leaving the question blank. The average percentage correct on the exams fell in the range 75–85.

The first three midterms were held in a classroom with nearly every seat occupied. A single teaching assistant (TA) proctored the exam. During the third midterm, a student came to the TA reporting suspicions that another student had been cheating. The proctor did not take any action regarding the cheating during the midterm, but did report this information to the professor after the exam. In response, the professor sent out an email, saying that 'cheating is morally wrong', and encouraged students to admit their wrongdoing. No students voluntarily came forward, although a second student said she had also witnessed cheating. This prompted the professor to once again call for student confessions, bolstered by the threat that he was going to contact us—the authors of this paper—and have us catch the cheaters.[5] Again, no one came forward, and the professor did indeed contact us two weeks before the final.

The data available to us include students' answers to each question on all four exams, as well as seating information for the third midterm and the final exam. In addition, we were able to carry out an unusual experiment involving the final exam. Students entered the exam room and selected their own seats, as was usual practice. These seating choices were recorded. Before the exam actually began, however, students were shuffled into randomly assigned seats for the test taking. This provided us with the opportunity to observe correlations in answer patterns among students who would have liked to sit together (and perhaps studied together), but were then separated.

Further steps were taken to prevent cheating on the final exam. Unlike the first three exams, where only one TA served as a proctor, four proctors were present during the final. Finally, the professor created two different versions of the final exam; the questions in both versions were the same, but the order in which they were asked was different. Students randomly received one of the two different versions of the test.

II. Reduced-form Detection of Cheating

We begin our analysis of possible cheating with a simple reduced-form regression approach in which the unit of observation is a *pair* of students on a particular exam. For each possible pair of students, we calculate the number of questions for which those students gave the same *correct* answer, and also the number of questions for which those students gave the same incorrect answer. If the number of common answers is high, then

this may be an indication of cheating, although of course there may be other explanations as well.

Copying from a student to one's left or right is the simplest way to cheat. Thus the key variable of interest in the regression is an indicator variable that is equal to 1 if the students sit next to each other, and takes the value 0 otherwise.[6] Given the room setup, it is difficult to cheat from two seats away with an empty seat in the middle, but triplets of students might effectively cheat. Therefore we include an indicator for students whose seating pattern is 'student 1—empty seat—student 2' as well as an indicator if the seating pattern is 'student 1—some other student—student 2'. Cheating from in front or behind another student is not easily done, so we would not expect this type of proximity to lead to elevated numbers of shared answers due to cheating. On the other hand, if there are other factors that lead to correlation in answers for students who sit near each other (e.g. because they study together, or good students congregate near the front of the room), then the back–front indicator will capture these effects. In some specifications we also control for the gender composition of the pair, or whether they are part of the same academic department.[7]

Table 1 shows the results of these regressions, using as the outcome variable the number of shared incorrect answers across the two students. The results in columns (1) and (2) correspond to the third midterm in which cheating is suspected. Columns (3)–(6) reflect the final exam. In columns (3) and (4), the right-hand side variables associated with seat locations are the initial, voluntary seats occupied by the students; columns (5) and (6) reflect the assigned seats given to the students—where they actually sat when

TABLE 1
MATCHING INCORRECT ANSWERS AMONG PAIRS

	3rd midterm		Final pre		Final post	
Test	(1)	(2)	(3)	(4)	(5)	(6)
Left–right pair	1.105***	1.104***	0.096	0.046	0.111	0.077
	(0.253)	(0.254)	(0.186)	(0.198)	(0.195)	(0.193)
Front–back pair	−0.062	−0.062	0.039	0.025	0.039	0.026
	(0.149)	(0.151)	(0.153)	(0.156)	(0.159)	(0.165)
Two apart: middle student	0.614*	0.611*	−0.149	−0.195	−0.078	−0.117
	(0.268)	(0.274)	(0.195)	(0.191)	(0.219)	(0.218)
Two apart: no middle student	0.147	0.119	−0.567	−0.538	−0.397	−0.359
	(0.237)	(0.223)	(0.357)	(0.360)	(0.378)	(0.374)
Constant	2.340***	2.316***	3.981***	3.962***	3.980***	3.961***
	(0.013)	(0.019)	(0.013)	(0.022)	(0.014)	(0.022)
Controls	No	Yes	No	Yes	No	Yes
N	19,110	19,110	22,578	22,578	22,578	22,578
R^2	0.003	0.010	0.000	0.023	0.000	0.023

Notes
Each observation is a pair of students who took the exam. Each regression uses the number of matching incorrect answers on the given exam as the dependent variable. The variables of interest are dummies indicating if the students sat in the specified arrangement while taking the given test. Columns (3) and (4) use each student's chosen seating position; columns (5) and (6) use each student's position after randomly reassigning seats. Odd columns do not include controls for gender and school, while even columns do include those controls. Bootstrapped standard errors are reported.
*, **, ***indicate significance at the 10%, 5%, 1% level, respectively.

taking the final exam. The odd columns include no controls; the even columns include controls.

Students who sit next to one another on the midterm have an additional 1.1 shared incorrect answers. This estimate is highly significant statistically.[8] Note that because the typical student gets most of the questions correct, the mean number of shared incorrect answers across all pairs of students is only 2.4. Thus students who sit next to each other have roughly 50% more shared incorrect answers than would be expected by chance. In contrast, we see no evidence that trying to sit next to one another in the final—but being relocated before the test begins—leads to shared incorrect answers. If, for instance, studying together is the cause of correlated answers among people who choose to sit next to each other, then we would expect more shared incorrect answers even after those students are relocated prior to the test. Students who actually sit next to each other in the final—after having been relocated and in the presence of heavy monitoring—also show no evidence of correlated incorrect answers in columns (5) and (6) of Table 1.

Pairs of students sitting two seats away with another student in between also have elevated levels of matching incorrect answers, on the midterm only. If there is an empty seat in between, however, then the correlation in incorrect answers disappears. This suggests that triplets of students may have worked together to cheat. There is no impact on shared answers among students sitting front to back on any of the tests.

Table 2 is identical to Table 1 except that the dependent variable is the number of shared correct answers. The pattern of coefficients is quite similar to the previous table. Students sitting next to each other during the midterm have an extra 1.2 shared correct answers (but off a baseline of over 30 correct shared answers, so in percentage terms the

TABLE 2
MATCHING CORRECT ANSWERS AMONG PAIRS

	3rd midterm		Final pre		Final post	
Test	(1)	(2)	(3)	(4)	(5)	(6)
Left–right pair	1.203*	1.204*	−0.124	−0.082	−0.388	−0.363
	(0.526)	(0.555)	(0.770)	(0.741)	(0.724)	(0.735)
Front–back pair	0.580	0.578	0.191	0.214	0.190	0.213
	(0.450)	(0.452)	(0.700)	(0.727)	(0.697)	(0.698)
Two apart: middle student	0.762	0.770	0.486	0.568	0.129	0.190
	(0.532)	(0.524)	(0.883)	(0.873)	(0.870)	(0.899)
Two apart: no middle student	−0.636	−0.550	−2.233	−2.211	−1.168	−1.104
	(0.849)	(0.876)	(1.163)	(1.163)	(1.488)	(1.500)
Constant	31.176***	31.154***	50.750***	50.638***	50.751***	50.639***
	(0.036)	(0.053)	(0.057)	(0.084)	(0.057)	(0.080)
Controls	No	Yes	No	Yes	No	Yes
N	19,110	19,110	22,578	22,578	22,578	22,578
R^2	0.001	0.009	0.000	0.003	0.000	0.003

Notes
Each observation is a pair of students who took the exam. Each regression uses the number of matching correct answers on the given exam as the dependent variable. The variables of interest are dummies indicating if the students sat in the specified arrangement while taking the given test. Columns (3) and (4) use each student's chosen seating position; columns (5) and (6) use each student's position after randomly reassigning seats. Odd columns do not include controls for gender and school, while even columns do include those controls. Bootstrapped standard errors are reported.
*, **, ***indicate significance at the 10%, 5%, 1% level, respectively.

impact is small). This coefficient is statistically significant at the 0.05 level. Students who are two seats apart with another student in between on the midterm once again have positive coefficients (but this time statistically insignificant). None of the other seating variables coefficients is particularly predictive; if anything, sitting two seats apart with an empty seat in the middle reduces the number of shared correct answers.

III. A More Systematic Algorithmic Approach to Detecting Cheating

The regressions above show that students who sit next to each other tend to have an increased number of both correct an incorrect answer pairs on average, but for identifying likely cheaters, it is the abnormality of the answers of a particular pair of students that is critical. In order to identify unlikely occurrences of matching answers, we first need to establish a baseline expectation with respect to the expected number of matching answers for any pair of students. To do so, we begin by modelling a student's answer on a particular question on either the third midterm or the final exam using a multinomial logit of the form

$$(1) \qquad p_{sa} = \Pr(y = a) = \begin{cases} \frac{e^{X_s\beta_a}}{1+\sum_{a=1}^{3} e^{X_s\beta_a}} & \text{if } a<4, \\ \frac{1}{1+\sum_{a=1}^{3} e^{X_s\beta_a}} & \text{if } a=4, \end{cases}$$

where s indexes students, and a reflects the answer that the student gave to that question. There are four possible answers to each question (a = 1, 2, 3, 4). In our basic

Table 3
Residual Matches on Third Midterm Less Residual Matches on Final

	Matching incorrect		Matching correct	
	(1)	(2)	(3)	(4)
Left–right pair	0.958***	0.967***	0.671	0.670
	(0.281)	(0.263)	(0.467)	(0.482)
Front–back pair	0.094	0.105	0.458	0.458
	(0.204)	(0.210)	(0.436)	(0.427)
Two apart: middle student	0.489	0.490	0.244	0.248
	(0.324)	(0.319)	(0.610)	(0.628)
Two apart: no middle student	−0.516	−0.471	−0.250	−0.237
	(0.442)	(0.422)	(0.761)	(0.800)
Constant	−0.005	−0.004	−1.063***	−0.979***
	(0.017)	(0.024)	(0.034)	(0.051)
Controls	No	Yes	No	Yes
N	18,915	18,915	18,915	18,915
R^2	0.001	0.006	0.000	0.001

Notes
Each observation is a pair of students who took the exam. Each regression uses residual matches on the third midterm less residual matches on the final as the dependent variable. Residual matches are the number of observed matching answers less the number predicted by the logit regression. The variables of interest are dummies indicating if the students sat in the specified arrangement. Odd columns do not include controls for gender and school, while even columns do include those controls. Bootstrapped standard errors are reported.
*, **, ***indicate significance at the 10%, 5%, 1% level, respectively.

specifications, we use the student's percentage correct on the final to predict the answers that a student gives to a particular question on the third midterm or the final. We use scores from the final exam only because the percentage correct on the midterms may be inflated for cheating students. In computing the student's percentage correct, we exclude the results for that particular question.[9]

Let $\widehat{p_{ia}^{q}}$ denote the estimated probability that student i gives answer a on question q. Further, denote a as the correct answer, and b, c and d as incorrect answers. If two students i and j answer a particular question independently, then the probability that they both choose answer a, conditional on the variables included as controls in the multinomial logit, is given by $\widehat{p_{ia}^{q}} \times \widehat{p_{ja}^{q}}$. For each pair of students ij, the expected numbers of matching right and wrong answers are given by

$$E(\text{matching right answers}) = \sum\nolimits_{q} \widehat{p_{ia}^{q}} \times \widehat{p_{ja}^{q}}, \tag{2}$$

$$E(\text{matching wrong answers}) = \sum\nolimits_{q} \widehat{p_{ib}^{q}} \times \widehat{p_{jb}^{q}} + \widehat{p_{ic}^{q}} \times \widehat{p_{jc}^{q}} + \widehat{p_{id}^{q}} \times \widehat{p_{jd}^{q}}. \tag{3}$$

We then compute two potential indicators of cheating based on unexpected concordance of answer patterns: (i) the residual between the observed and predicted number of matching *correct* answers (Δ_c), and (ii) the residual between the observed and

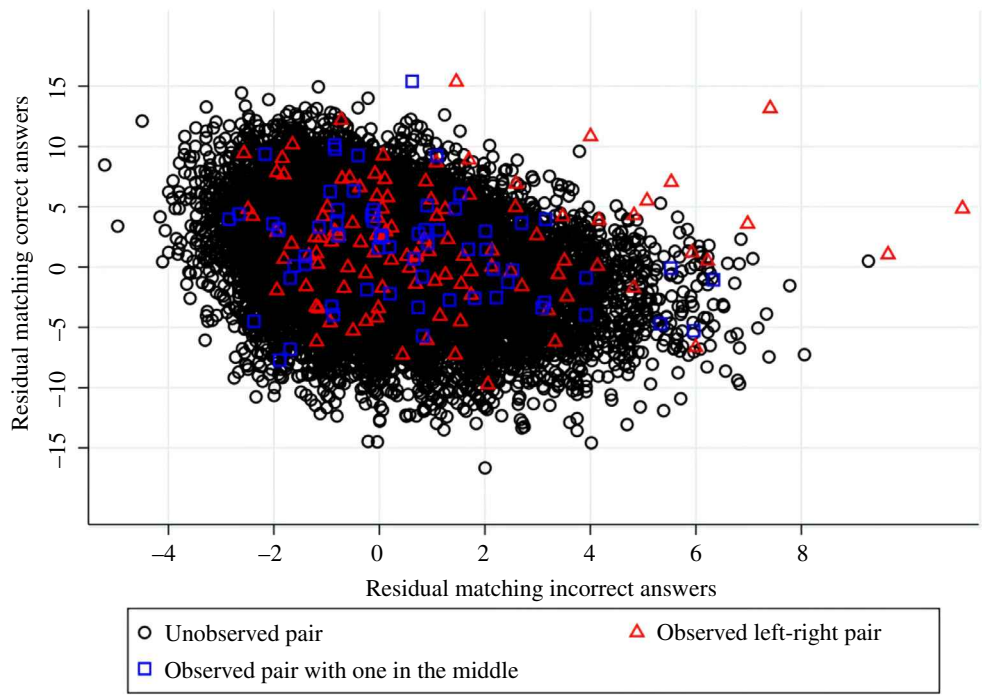

FIGURE 1. Third midterm—residual matching answers for student pairs.
Notes: The figure shows residual matching correct and incorrect answers on the third midterm. A multinomial logit estimated the expected number of matching answers using students' percentage correct on the final exam. Each student pair is plotted based on their seating position on the exam. [Colour figure can be viewed at wileyonlinelibrary.com]

predicted matching *incorrect* answers (Δ_i). *A priori*, it is uncertain which of these two measures will be most predictive empirically. We estimate the probabilities implied in model (1) using data for the 214 students that took both the third midterm as well as the final.[10] We then create a dataset of all possible student pairs (22,791 = 214 × 213/2) for which we compute the number of matching correct and incorrect answers as well as the expected number of matching correct and incorrect answers, using equations (2) and (3), respectively. Finally, we compute Δ_c and Δ_i as the differences between the observed and predicted number of matching correct and incorrect answers, respectively. As a further check of our reduced-form results presented earlier, Table 3 reports identical specifications to those shown in the earlier tables, except that the dependent variable is now the residual number of matching answers on the third midterm minus the residual number of matching answers on the final. We still observe highly significant results for left–right pairs on matching incorrect answers; for matching correct answers, the sign is positive, but the point estimate is only half as large as in Table 2 and no longer statistically significant. This suggests that some of the observed excess in matching correct answers is driven by students of similar abilities sitting next to one another.

Figure 1 shows a scatterplot of Δ_c against Δ_i on the third midterm, where cheating was suspected. Each symbol in the plot represents a pair of students in the class. Triangles correspond to pairs of students who sat next to each other during the third midterm; squares are pairs of students who had one seat in between them, with that seat

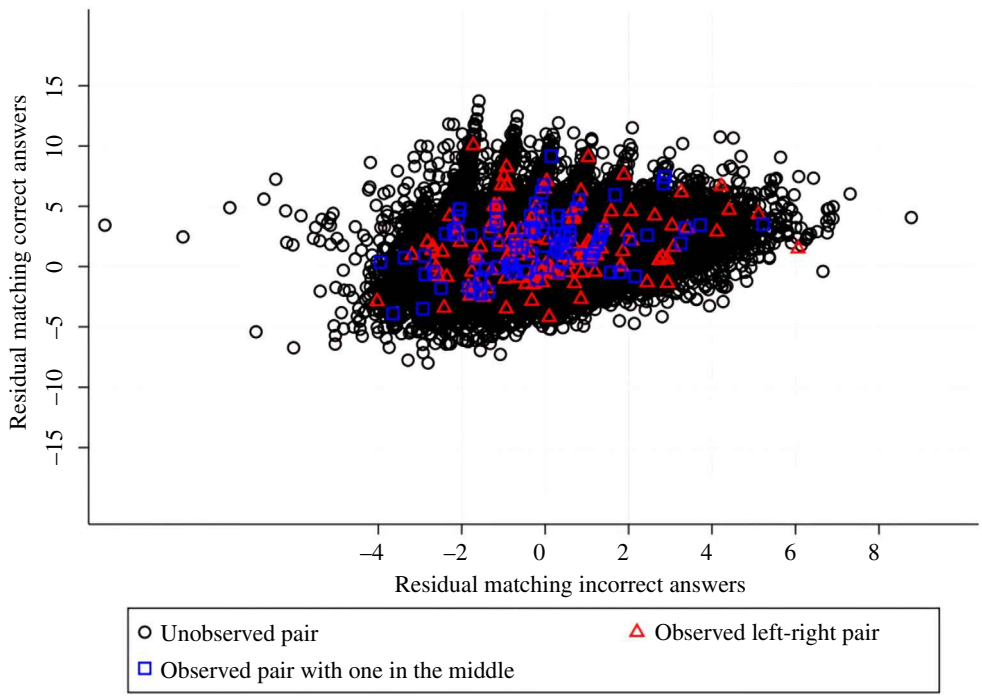

FIGURE 2. Final pre-randomization—residual matching answers for student pairs.
Notes: The figure shows residual matching correct and incorrect answers on the final. A multinomial logit estimated the expected number of matching answers using students' percentage correct on the final exam. Each student pair is plotted based on their chosen seating position pre-randomization. [Colour figure can be viewed at wileyonlinelibrary.com]

occupied; circles are a 1% sample of all other pairs of students. The further to the right a pair is in the figure, the greater the number of unexpected shared incorrect answers. The higher up a pair is in the figure, the greater the number of excess correct answers. The triangle and square symbols are greatly overrepresented in the north-eastern part of the figure, consistent with cheating. Although pairs of students sitting next to each other represent only 0.5% of all possible pairs of students, three of the six rightmost data points are pairs of students who were next to each other. Although the pattern is less clear in the vertical dimension, the single greatest anomaly on shared correct answers is a pair seated next to one another. A number of triangles are outliers in the north-eastern direction, that is, they have high residuals on both dimensions. Some squares are near the far right of the figure, but the pattern is much less obviously extreme than for the triangles.

For purposes of comparison, Figures 2 and 3 mirror Figure 1, but for the seating positions originally selected on the final (Figure 2) and the actual seating positions on the final (Figure 3). In stark contrast to Figure 1, there is no visual evidence that students wishing to sit next to each other or actually sitting next to one another have unusual answer patterns. This supports the interpretation of cheating on the midterm.

Figure 4 provides a more systematic means of capturing the degree to which sitting next to one another produces anomalous patterns on the midterm. The horizontal axis in Figure 4 captures ranges of excess incorrect answers (Δ_i), with the rightmost columns

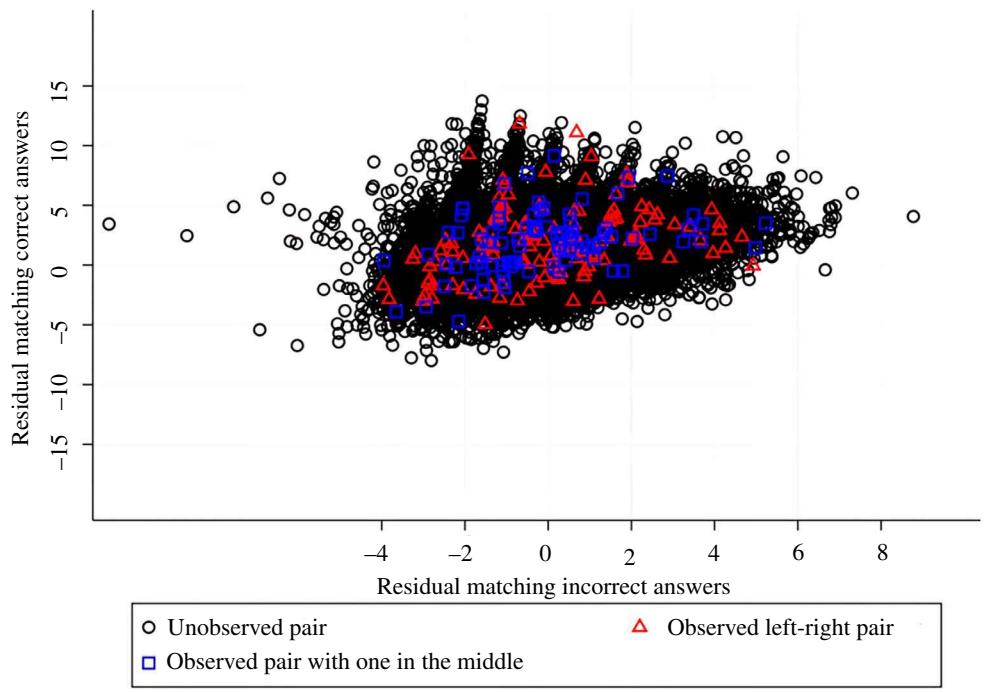

FIGURE 3. Final post- randomization—residual matching answers for student pairs.
Notes: The figure shows residual matching correct and incorrect answers on the final. A multinomial logit estimated the expected number of matching answers using students' percentage correct on the final exam. Each student pair is plotted based on their assigned seating position post-randomization. [Colour figure can be viewed at wileyonlinelibrary.com]

corresponding to the highest (most suspicious) values. The height of a column represents the hazard rate of the frequency with which students who sit next to each other produce outcomes in that range, compared to all pairs of students. Standard error bands are shown as vertical lines. If students sitting side-by-side look similar to randomly chosen pairs of students, then the hazard rates in all columns would be equal to 1. If the hazard rate in a particular column is 2, then that means that students sitting next to each other are twice as likely to generate answers in that range.

As can be seen in Figure 4, pairs of students who sit next to each other in the midterm produce answer patterns that diverge greatly from random pairs of students. (Note that the hazard rates are presented on a log scale because the differences in the tails are so

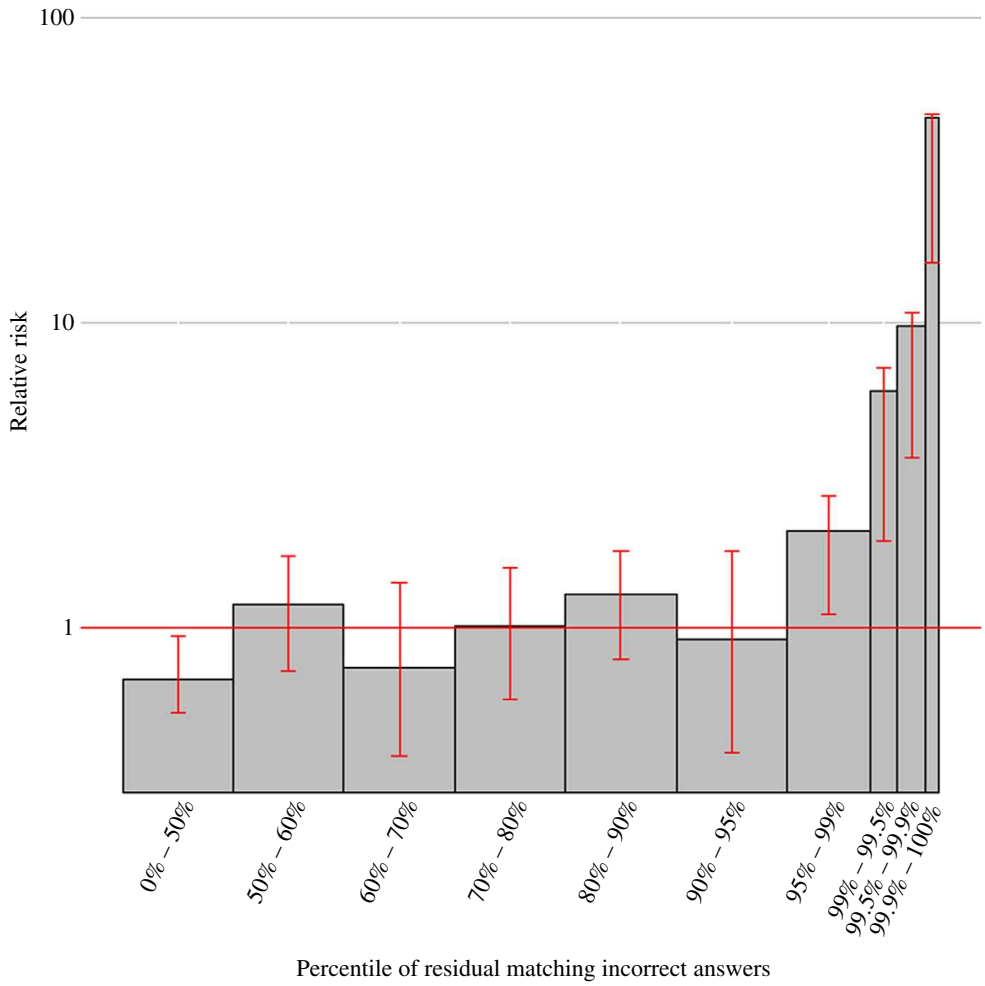

FIGURE 4. Left–right pairs—matching incorrect answers on the third midterm.
Notes: The figure reports relative risk ratios for the percentile of residual matching incorrect answers on the third midterm. The ratios are calculated as the ratio of the proportion of students sitting in left–right pairs who are in the given percentile range to the proportion of students not sitting in left–right pairs who are in the given percentile range. Standard error bars showing a 95% confidence interval are shown as vertical lines. [Colour figure can be viewed at wileyonlinelibrary.com]

large.) Focusing first on the rightmost column of the table, which represents outcomes in the top 0.1% in terms of unexpected matching incorrect answers, pairs of students who sit next to each other are roughly 47 times as likely to fall into this category as a random pair. The null hypothesis that students sitting next to one another are no more likely than any random pair to be in this category is easily rejected. In the next column, reflecting outcomes in the 95.5th to 95.9th percentiles, students who sit next to each other are approximately 10 times more frequently represented than would be expected by chance. This result is also highly significant statistically. Chance would have us expect 0.5% of pairs of students sitting next to each other to appear in one of the two rightmost columns, or less than one pair. In actuality, about 7% of all left–right pairs (14 individual students because some are in multiple pairs) show up in the extreme tail. For the third and fourth

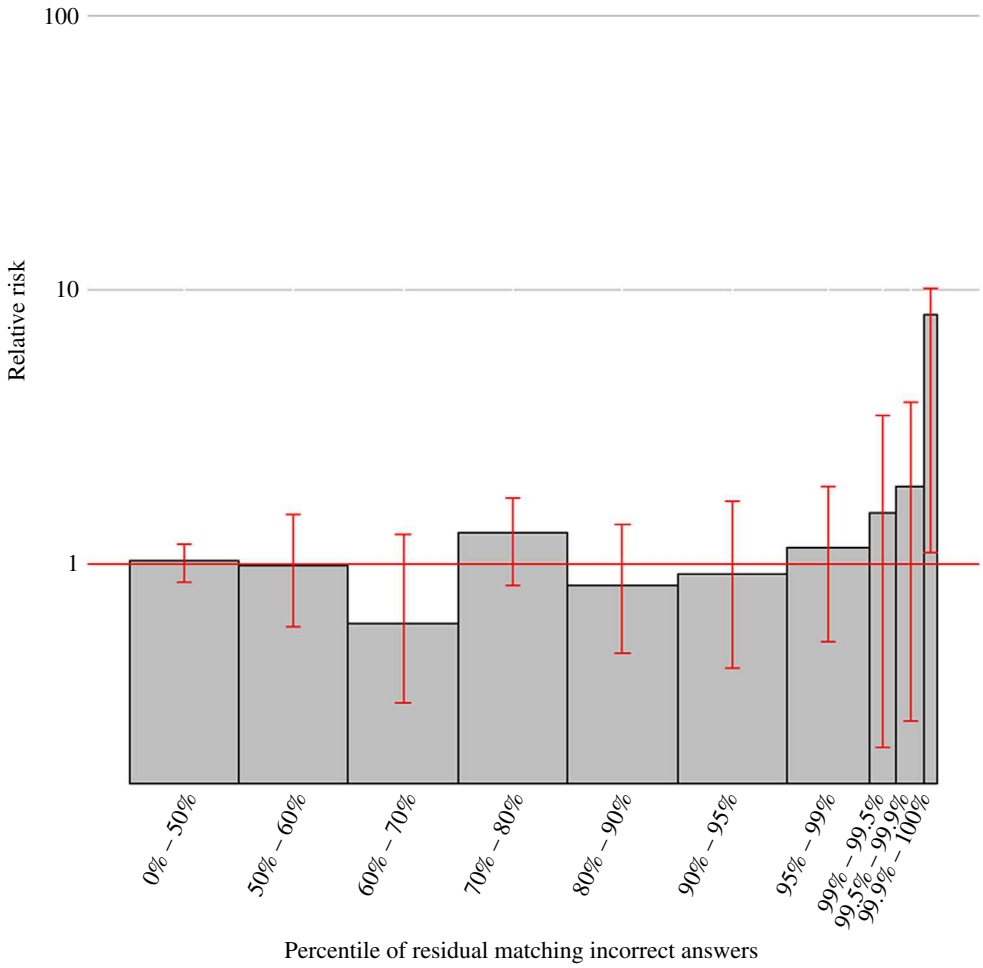

FIGURE 5. Left–right pairs—matching incorrect answers on the final exam pre-randomization.
Notes: The figure reports relative risk ratios for the percentile of residual matching incorrect answers on the final using seating choices pre-randomization. The ratios are calculated as the ratio of the proportion of students sitting in left–right pairs who are in the given percentile range to the proportion of students not sitting in left–right pairs who are in the given percentile range. Standard error bars showing a 95% confidence interval are shown as vertical lines. [Colour figure can be viewed at wileyonlinelibrary.com]

rightmost columns, we also reject the null hypothesis of a hazard rate of 1 for students who sit next to each other: 11% of all left–right pairs show up in those two columns (compared to 4.5% which would be expected by chance). Thus almost 20% of all left–right pairs are above the 95th percentile in residual incorrect answers. Eliminating double-counting of students who appear multiple times, we estimate based on the excess weight in the right tail that more than 10% of students likely cheated.

Figures 5 and 6 present parallel results for the initial and actual seatings of the final exam. In contrast to Figure 4, there is little evidence that sitting next to another student is associated with large jumps in shared incorrect answers. In Figure 5 the rightmost column is significant because a single pair of students who initially chose to sit next to one another appear in the top 0.1% of residual matching incorrect answers. We cannot

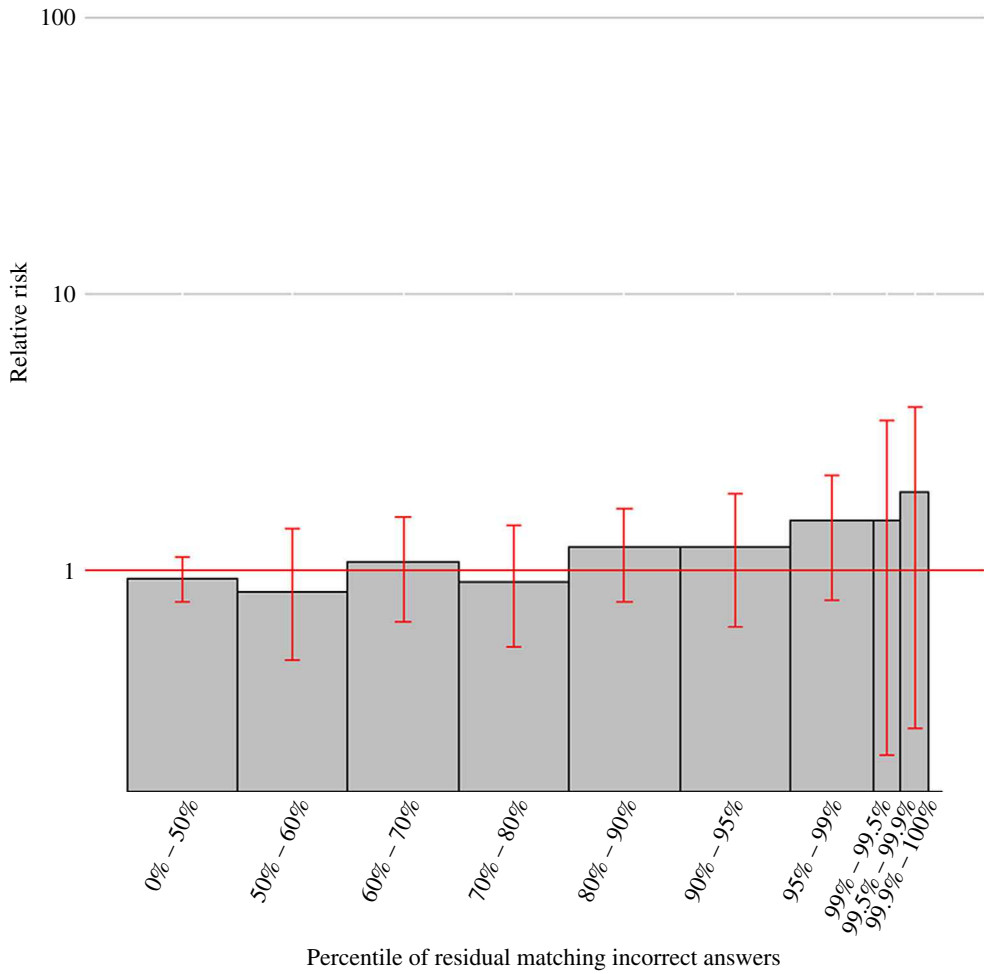

FIGURE 6. Left–right pairs—matching incorrect answers on the final exam post-randomization.
Notes: The figure reports relative risk ratios for the percentile of residual matching incorrect answers on the final using seating assignments post-randomization. The ratios are calculated as the ratio of the proportion of students sitting in left–right pairs who are in the given percentile range to the proportion of students not sitting in left–right pairs who are in the given percentile range. Standard error bars showing a 95% confidence interval are shown as vertical lines. [Colour figure can be viewed at wileyonlinelibrary.com]

reject the null hypothesis of a hazard rate of one for any of the columns of interest in Figure 6.

It is always easier to identify the *number* of cheaters than to identify individual students as cheaters. In this case, however, because the hazard rates are so enormous (almost 50) in the far right tail, the likelihood of false positives for that group is unusually low (about 2%). For the second rightmost column in Figure 4, false positive rates would be about 10%. We elected not to pursue punishment for that group, although others might have made a different choice at that rate of false positives.

Figures 7–9 present the same set of results, but for matching correct answers rather than for matching incorrect answers. Three pairs—six individual students—appear in the two rightmost columns. Of these, one of the pairs also would have been labelled cheaters

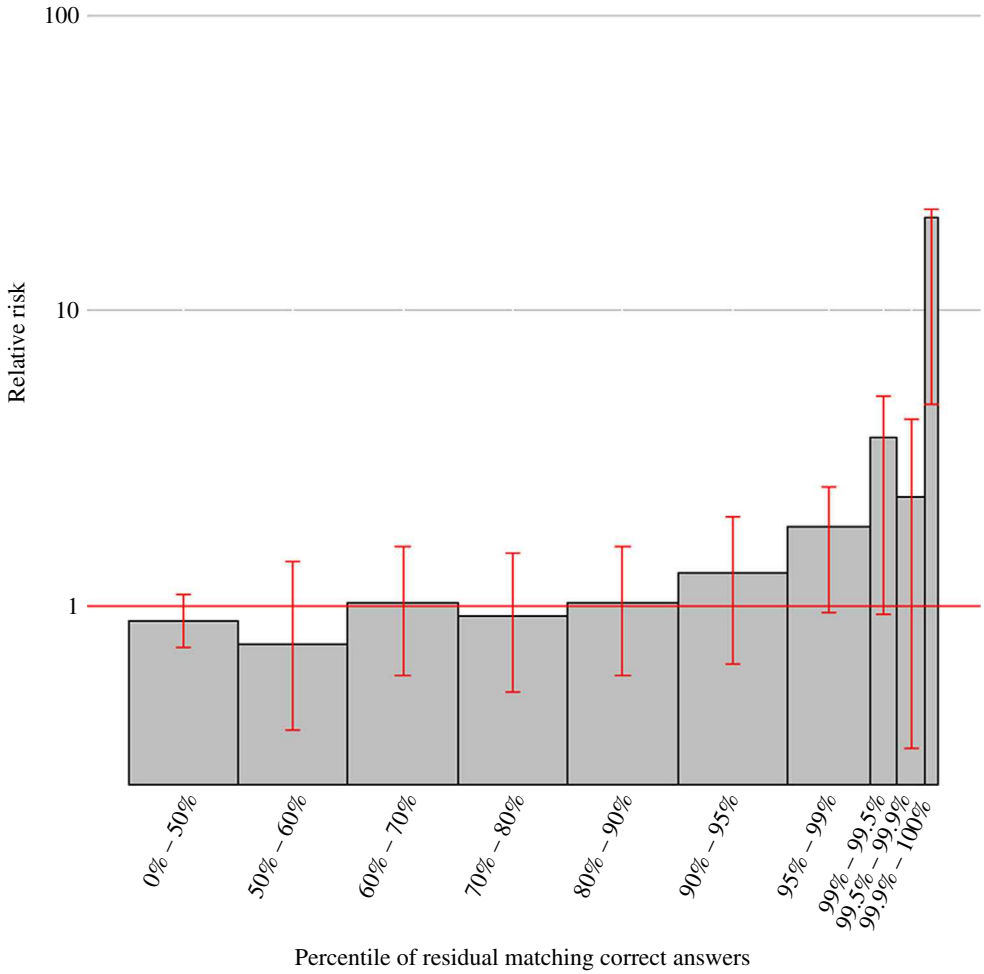

FIGURE 7. Left–right pairs—matching correct answers on the third midterm.
Notes: The figure reports relative risk ratios for the percentile of residual matching correct answers on the third midterm. The ratios are calculated as the ratio of the proportion of students sitting in left–right pairs who are in the given percentile range to the proportion of students not sitting in left–right pairs who are in the given percentile range. Standard error bars showing a 95% confidence interval are shown as vertical lines. [Colour figure can be viewed at wileyonlinelibrary.com]

based on anomalies in their incorrect answers. Thus matching correct answers adds relatively little to the potential cheating detection. Moreover, unlike for incorrect answers, there is no overrepresentation in the rightmost column on the final exam for those who wanted to sit together but weren't allowed to, and there is overrepresentation for the students who were randomly assigned to sit next to each other. The fact that randomly assigned students who sat next to each other also have correlated correct answers may point to some cheating on the final. If the extra weight in the tails on the final is indeed due to cheating, then that suggests that four students cheated on the final, still a much lower rate than on the midterm.

We have carried out a similar analysis for students who sat with a chair between them occupied by another student.[11] On the midterm, relative risk ratios

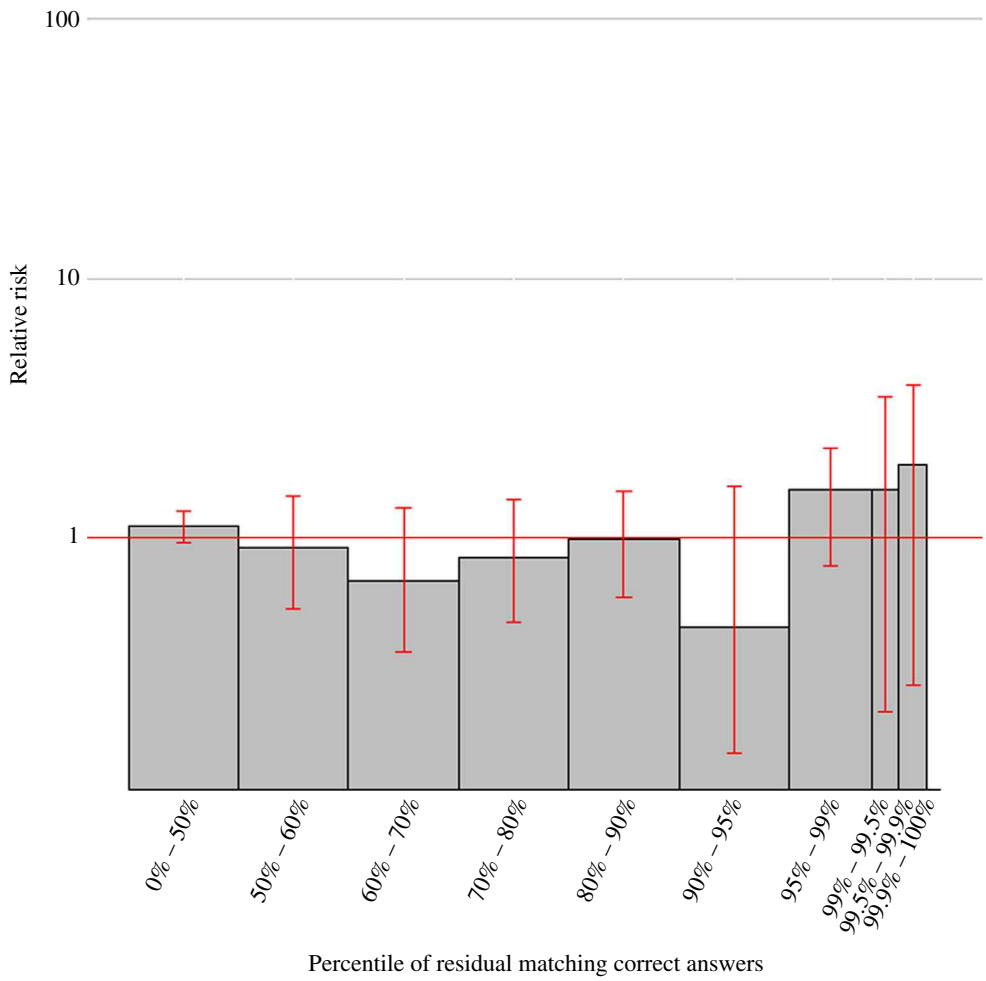

FIGURE 8. Left–right pairs—matching correct answers on the final exam pre-randomization.
Notes: The figure reports relative risk ratios for the percentile of residual matching correct answers on the final using seating choices pre-randomization. The ratios are calculated as the ratio of the proportion of students sitting in left–right pairs who are in the given percentile range to the proportion of students not sitting in left–right pairs who are in the given percentile range. Standard error bars showing a 95% confidence interval are shown as vertical lines. [Colour figure can be viewed at wileyonlinelibrary.com]

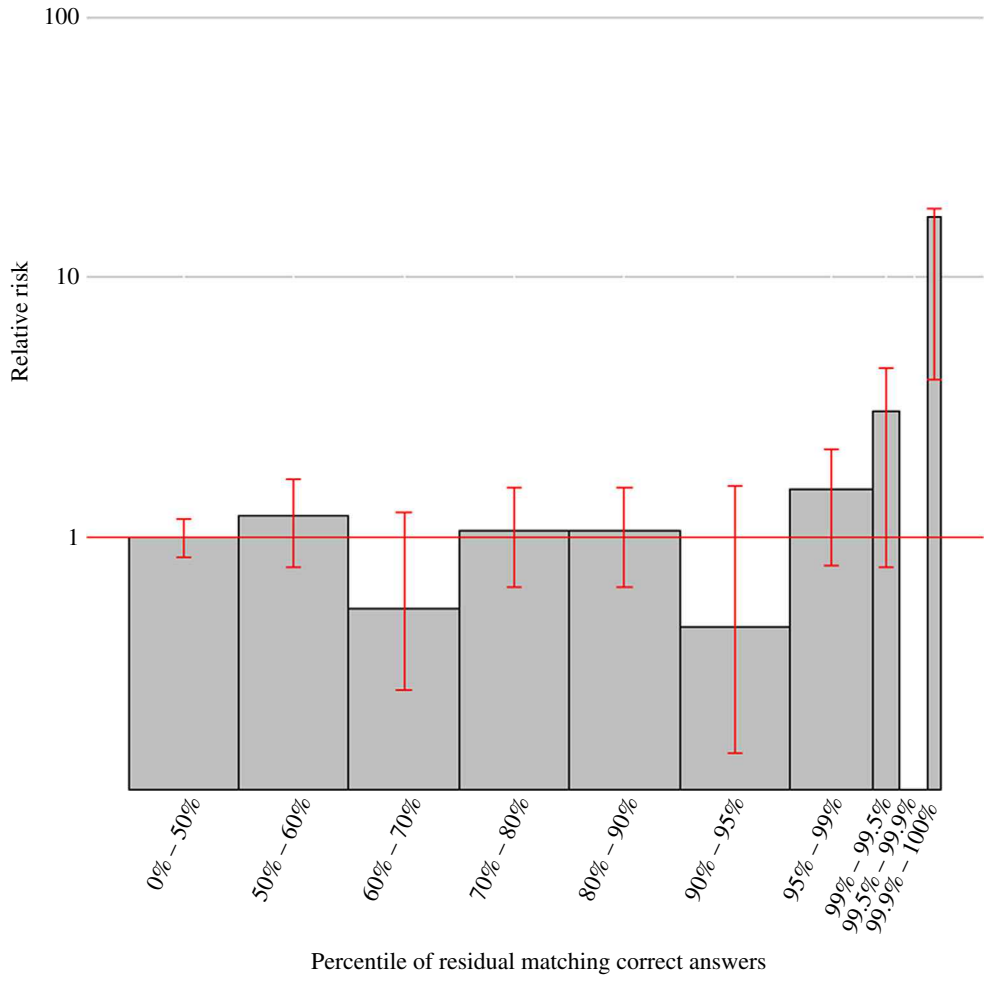

FIGURE 9. Left–right pairs—matching correct answers on the final exam post-randomization.
Notes: The figure reports relative risk ratios for the percentile of residual matching correct answers on the final using seating assignments post-randomization. The ratios are calculated as the ratio of the proportion of students sitting in left–right pairs who are in the given percentile range to the proportion of students not sitting in left–right pairs who are in the given percentile range. Standard error bars showing a 95% confidence interval are shown as vertical lines. [Colour figure can be viewed at wileyonlinelibrary.com]

greater than 10 are present for the rightmost category for both incorrect and correct answers. Doing calculations like those above, eight students sitting with a seat between them are identified as likely cheaters on the midterm. Of those eight, four were identified as cheaters in the analyses above; four of them would have been missed.

IV. CONCLUSION

It is not surprising that students cheat—they have strong incentives to do so, and the likelihood of getting caught is low. What is perhaps more surprising is that so little effort is devoted to catching cheating students. In this paper, we develop a simple algorithm for detecting cheating. In the particular setting in which we apply that algorithm, we

conclude that more than 10% of the students in the class appeared to have cheated in a manner blatant enough to be detected by our approaches.

Perhaps the best supporting evidence for our claims of cheating (and also, perhaps, a powerful explanation as to why so little effort is invested in detecting cheaters) comes from what happened after we carried out our analysis. Based on our initial findings, the professor in the class forwarded the names of the six most suspicious pairs of students to the Dean's office, an investigation was initiated, and a student judiciary court hearing was scheduled.[12] Before the hearing could occur, four of the twelve students confessed. Despite these admissions, the Dean's office nonetheless cancelled the investigation the day before the student court hearing, due to pressure from parents. While this precluded any further admissions of guilt, the professor withheld grades of the presumptive guilty pairs until the first day of the next semester, which resulted in scholarship disqualification. Notwithstanding this punitive action, none of the twelve accused students complained or sought redress.

While our results do not speak directly to effectiveness of honour codes, we are highly sceptical regarding their effectiveness. What we observe in our analysis is that interventions that make cheating more difficult (better proctoring, randomly assigned seats) dramatically reduce cheating. In the presence of honour codes, there is often very little investment in preventing cheating (e.g. students are allowed to take exams in their dorm rooms). One would have to believe that the social/moral costs imposed by the presence of an honour code are extremely powerful, given the strong incentives to cheat.

ACKNOWLEDGMENTS

We thank the professor teaching this course for providing data. We thank Eric Andersen, Dai-Rong Chen, Yue-Shuan Chun, Jason Lai, Dhiren Patki, and Graham Tierney for their excellent research assistance. Financial support from the Ministry of Science and Technology, Taiwan is greatly appreciated.

NOTES

1. Harvard University admitted that 'about 125 students might have worked in groups on a take-home final exam'. Roughly 70 students were forced to withdraw from the university (Perez-Pena and Bidgood 2013). Similar numbers of students were involved at Dartmouth and the Air Force Academy (Associated Press 2015a; Frosch 2007). In March 2015, Stanford University Provost John Etchemendy sent a letter to the faculty expressing concerns over allegations of widespread cheating (Associated Press 2015b).
2. Jacob and Levitt (2003) develop a set of tools for analysing *teacher* cheating, some of which we build on in this paper.
3. Organizations such as the Educational Testing Service (ETS), provider of the SAT and GRE exams, no doubt have developed techniques for detecting cheating, but to the best of our knowledge, these tools have never been made publicly available.
4. Students were required to take only two of the three midterms. The midterms had 50 questions each; the final exam had 80 questions.
5. In his email, the professor warned the students that '[they are] extremely good at catching cheating if you have read *Freakonomics*'. Apparently, none of the cheaters had read *Freakonomics*.
6. Because we have seating charts for only the third midterm and the final, our analysis is restricted to these two tests.
7. For gender we include dummies for 'both female', 'one female, one male' and 'two males'. Each student is assigned to an academic department within the university (e.g. engineering or arts and sciences).
8. Our data have a complex dependency structure because each student appears in multiple student pairs. To account for this, we report the bootstrapped standard errors in the regression tables.
9. An argument could be made for using only the student's performance on the final exam as a control variable, due to cheating concerns on the midterms. Empirically, our results are little changed if we include the midterm scores, or if we add more covariates such as a gender dummy.

10. We drop questions that all students answer correctly, as they provide no information. We also drop a handful of cases where exactly one student gave a particular answer on a question because of non-convergence of the multinomial logit estimation.
11. Full results are available from the authors.
12. Our initial detection algorithms were not as good as those that we eventually developed; that is the reason why only 12 students were identified.

REFERENCES

ASSOCIATED PRESS (2015a). Dartmouth students charged with cheating; available online at https://apnews.com/15c597e9eed843c293f1644ccb210ab6 (accessed 16 November 2019).

—— (2015b). Unusual amount of cheating at Stanford University'; available online at https://www.foxnews.com/us/unusual-amount-of-cheating-at-stanford-university-provost-warns-of-severe-consequences (accessed 16 November 2019).

BAKER, A. (2012). At top school, cheating voids 70 pupils' tests. *New York Times*, 9 July.

FROSCH, D. (2007). 18 Air Force cadets exit over cheating. *New York Times*, 2 May.

MCCABE, D. (2005). Cheating among college and university students: a North American perspective. *International Journal for Academic Integrity*, **1**(1), 1–11.

JACOB, B. and LEVITT, S. (2003). Rotten apples: an investigation of the prevalence and predictors of teacher cheating. *Quarterly Journal of Economics*, **118**(3), 843–77.

PEREZ-PENA, R. and BIDGOOD, J. (2012). Harvard says 125 students may have cheated on a final exam. *New York Times*, 30 August.

WESOLOWSKY, G. (2000). Detecting excessive similarity in answers on multiple choice exams. *Journal of Applied Statistics*, **27**(7), 909–21.

ZITZEWITZ, E. (2012). Forensic economics. *Journal of Economic Literature*, **50**(3), 731–69.