The Expected Value of Cheating

Craig H. Wisen
University of Alaska Fairbanks

Cheating in a traditional test setting is estimated to be of a magnitude similar to the proportion of firms that backdate employee stock options. This paper describes a methodology for the detection of cheating. The simple method for detecting cheating is presented in this study and applied to a large undergraduate class. The tool is easy to implement and provides valuable lessons in estimating the expected value of cheating, statistical size, type II errors, type I errors, and hypothesis testing.

INTRODUCTION

Cheating in academia occurs in a wide variety of contexts. The purpose of this study originated as an exam for an undergraduate finance class composed of 178 students. The study is subject to several constraints and limitations, because the test was not explicitly designed to detect cheating, and the test setting was not structured to be an experimental design on human subjects. Nevertheless, the results are instructive on several dimensions.

Faculty, staff, and students were genuinely surprised more by the ease at which cheating could be detected than by the extent to which cheating occurred. Approximately 25% of students cheated on the exam based on the cheater detection algorithm, which utilized a telltale pattern of improbable responses. Approximately 87% of the cheaters were male. By comparison, male students represented approximately 70% of the sample in which no indications of cheating were detected. The 25% rate of cheating is biased downward for two reasons. First, the algorithm was limited to the detection of a single cheating strategy. Second, the criteria for identifying cheating used a conservative estimate of observing the improbable responses.

Developing methodologies to detect cheating date back over a thousand years. For example, the practice of biting a gold coin was a test of softness, since counterfeit coins were debased with metals that were more resistant to deformation. Stock option backdating is a relatively recent discovery of cheating (Lie, 2005) that is based on a detection method similar to the current study, specifically a telltale pattern of improbable exercise prices. Cheating at the firm level is also similar in magnitude to student cheating in the current study. Heron and Lie (2009) estimate that 29.2% of firms manipulated grants to top executives at some point between 1996 and 2005. Of the 2,000 companies alleged to have manipulated stock option grants there were 12 criminal convictions. Roughly 150 companies were required to restate their financials (Lattman, 2010). It should be no surprise that the moral compass of counterfeiters and corporate executives is not dramatically different from that of students or teachers.

Teachers in the U.S. are often rewarded monetarily and through promotions for improving classroom performance. The common means of assessing improvement is through standardized tests. This creates an incentive to "teach to the test" and to correct wrong answers after the fact. Jacob and Levitt (2003)

developed a method for detecting retroactive cheating by teachers and administrators on standardized tests. One mechanism was a pattern of suspicious answer strings.

Certain methods for detecting and/or preventing student cheating rely on elaborate technology (Bedford, Gregg, and Clinton, 2011). While formulae that detect student cheating after the fact have been developed, they tend to lack general applicability. For example, one formula has been worked out that can prove cheating within fractions of one percent, but it involves comparing strings of identical answers from pairs of random students (Mogull, 2004).

The purpose of the test in the current study was to assess student comprehension of course material rather than to detect, or punish, cheating. The pervasiveness of cheating that occurred in the present study did not trigger a formal investigation or punishment of any student. There were two primary reasons that factored into the decision to forgo a formal investigation of cheating. The first reason was similar to the amount and severity of sanctions for option backdating. It was expected that the investigation of a large group of students accused of cheating would consume an inordinate amount of time and be unlikely to result in meaningful penalties. The second reason was that students who cheated received no competitive advantage or reward. In fact, students who copied answers from a different version of the exam would have fared better by guessing.

Students could determine whether they appeared to have cheated by calculating the probability of observing multiple correct responses to the alternate version of the exam. As a case study, the structure of the test renders it very informative on several levels, and the results create a catalyst for class discussion in the following areas:

- Pervasiveness of cheating in business and in academia
- Expected value of cheating
- Methodologies of cheating and detection
- Statistical size and power of hypothesis testing

> *The measure of a man's real character is what he would do if he knew he would never be found out*. Thomas Babington Macaulay

The power of any test to detect cheating is difficult to estimate, since one cannot easily estimate the probability that the test will reject a false null hypothesis, i.e. the false negative rate. Estimating the probability that a test for cheating will not make a Type II error requires knowledge of how frequently cheaters are not detected.

Power is usually dependent on three factors: Sample size, magnitude of effect, and statistical significance. All else being equal, increasing the sample size is one of the easiest ways to increase the statistical power of a test, because sampling errors decrease as the sample size increases. The magnitude of effect in the current setting is difficult to measure. Cheating occurs through several strategies, and copying the answers of a peer or competitor is only one tactic. When a student apparently looks at another student's exam, it is extremely difficult to determine whether cheating has occurred. In the extreme case, should a student briefly glance at another student's answers, the magnitude of effect would be quite subtle and nearly impossible to penalize. One can easily increase the power of a test by using a larger significance criterion. Doing so would unfortunately mean that a greater proportion of innocent students would be falsely flagged as having cheated.

TESTING STRUCTURE

The test consisted of 31 multiple-choice and two short-answer questions. There were two versions of the exam, referred to as *A* and *B*, which differed from one another in subtle ways on 18 questions. Both versions were identical in all respects on the remaining questions. The set of 18 questions can be thought of as cheater-detection questions in addition to simultaneously testing comprehension of course material.

The two exam versions were collated into an alternating sequence within a commingled stack. A smaller stack of commingled exams was then distributed to each row of students. The first page of each version of the exam consisted of an identical sheet of formulas. Students were instructed to wait until all exams were distributed prior to turning the page and starting the test. Students were probably unaware that different versions of the exam were present, because they looked nearly identical. Those who thought there was only one version of the exam would also be unaware that the stacks of exams were collated into an alternating sequence of the two versions.

An example of the first cheater-detection question from the midterm exam is shown below:

***Version A* Exam, Question 1:**

Assume your score without cheating would be fifty-eight percent. If you cheat and are not detected your score would increase to seventy-two percent; however, the probability of getting caught is thirty-eight percent. Individuals who are caught cheating receive a score of zero. Calculate the expected value of cheating.

a.	**30.64%**	**f.**	**13.36%**	**k.**	**-17.81%**
b.	**23.44%**	**g.**	**9.60%**	**l.**	**-20.56%**
c.	**22.60%**	**h.**	**-9.04%**	**m.**	**-22.60%**
d.	**20.56%**	**i.**	**-9.60%**	**n.**	**-23.44%**
e.	**17.81%**	**j.**	**-13.36%**	**o.**	**-30.64%**

***Version B* Exam, Question 1:**

Assume your score without cheating would be fifty-eight percent. If you cheat and are not detected your score would increase to seventy-two percent; however, the probability of getting caught is forty-eight percent. Individuals who are caught cheating receive a score of zero. Calculate the expected value of cheating.

a.	**30.64%**	**f.**	**13.36%**	**k.**	**-17.81%**
b.	**23.44%**	**g.**	**9.60%**	**l.**	**-20.56%**
c.	**22.60%**	**h.**	**-9.04%**	**m.**	**-22.60%**
d.	**20.56%**	**i.**	**-9.60%**	**n.**	**-23.44%**
e.	**17.81%**	**j.**	**-13.36%**	**o.**	**-30.64%**

The question, with different inputs, and its solution method were presented in lecture in the weeks prior to the exam. The question, again with a different set of inputs, and its solution method were also presented in the review session prior to the exam. Students were told during lecture that this particular question was highly likely to appear on the midterm. Students were *not* told prior to taking the exam that this question could simultaneously be used as a tool to detect cheating.

The irony that a question tasking 178 students to calculate the expected value of cheating was also detecting cheating during the test itself was not warmly received in the days following the exam. Although the correct answers to the expected value of cheating were negative, many students must have believed that the expected value of their own cheating attempt on this question was positive.

In the weeks prior to the exam, lecture topics included the option backdating scandal, corruption, and rogue trading. The discussion included how business, society, and complex natural systems enforce penalties for misrepresentation and included a variety of views on the morality of theft. The pedagogical style emphasized conceptual understanding and did not require the memorization of formulas. The question shown above required using a formula from the exam cover sheet:

$$E[Value|cheat] = p(penalty) + (1-p)(loot)$$

Where *p* is the probability of detection, *penalty* is the punishment for cheating, and *loot* is the increment that the test score would increase if cheating were not detected. The inputs for *Version A* are *p*=38%, *penalty*=58%, and *loot*=14%, and for *Version B, p*=48%, *penalty*=58%, and *loot*=14%. The correct response for *Version A* is -13.36% (j.) and the correct response for *Version B* is -20.56% (l.). Incorrect responses are not equally likely in a random setting, but empirically the most frequent incorrect response from *Version A* exams was -20.56% (l.), and the most frequent incorrect response from *Version B* exams was -13.36% (j.)

Each version of this question had the same set of 15 responses denoted "a." through "o." The probability of guessing the correct response needed to be the same for each version, which is equivalent to specifying that the probability of guessing the incorrect response is the same for each version of the question. The most likely incorrect ways of solving the problem were used to determine the incorrect responses. This is a good quality for most types of multiple-choice questions. Equal proportions of incorrect and correct responses were combined to form the common set of responses that were identical for each version of the question.

The probability of selecting the incorrect response, assuming no cheating and a purely random response strategy, would be 14 out 15. The probability of selecting the correct response to the alternate version of the exam, again assuming no cheating and a purely random response strategy, is 1 out of 14. The estimate of the proportion of the class that appeared to have cheated on the exam was defined by observing 3 or more responses on cheater-detection questions that were the correct response to the alternate version of the question. Assuming a random selection strategy, the probability of observing 3 or more incorrect responses that were also the correct answers to the alternate version of the exam would be less than (1/14)^3, or approximately 0.04%. This assumes that each of the 18 cheater-detection questions had 14 incorrect responses.

Issues that were addressed in post-exam discussions included the following:

A) Should different exam versions be clearly demarcated?

B) The expected value of cheating is inversely related to mastery of course material, severity of the penalty, and probability cheating will be detected.

C) Cheating may have increased over the last few decades because sanctions for cheating have decreased, there exists a greater amount of activities that distract students from studying, and improved technologies may increase the loot more than it increases the probability of detecting cheating.

RESULTS

Results were evaluated based upon the number of incorrect responses that *would have been considered correct* had they been entered on the alternate version of the exam. Students who had three or more correct responses from the alternate version of the exam were classified as "Definitely Cheated" (i.e. 0.04% probability of chance occurrence). Students who had two correct responses from the alternate version of the exam were classified as "Probably Cheated" (i.e. 0.50% probability of chance occurrence) and students with one or fewer correct responses from the alternate version of the exam were classified as "Did Not Cheat." In the tables below, the "Random Response Probability" is the percent of exams one would expect to observe in each of the three categories, assuming all students used a test-taking strategy of selecting responses randomly*, as opposed to copying "correct" incorrect answers*.

TABLE 1
PROBABILITY OF STUDENT CHEATING

	Did Not Cheat	Probably Cheated	Definitely Cheated	Class Total
Random Response Probability	100% to 93%	0.5%	0.04% or less	
Exam Score	51.2	41.6	36.2	46.6
Student Count	118	14	45	177
Proportion of Class	66.7%	7.9%	25.4%	100.0%

Males were more likely to have been categorized as "Definitely Cheated" than females. Gender assignment was based upon the first name of the student and could be subject to error in cases where the first name was ambiguous and the student did not attend class aside from the midterm, thus remaining unidentified. There was only one such case, which was eliminated from the study, resulting in a class total of 177.

TABLE 2
PROBABILITY OF STUDENT CHEATING BY GENDER

	Did Not Cheat	Probably Cheated	Definitely Cheated	Class Total
Random Response Probability	100% to 93%	0.5%	0.04% or less	
Male	82	12	39	133
Female	36	2	6	44
Male Proportion	69.5%	85.7%	86.7%	75.1%
Female Proportion	30.5%	14.3%	13.3%	24.9%

A total of 59 students were classified as either "Probably Cheated" or "Definitely Cheated." Approximately 76.3% of the 59 students were classified as "Definitely Cheated." The following quote is helpful in interpreting this result:

> *To keep your character intact you cannot stoop to filthy acts. It makes it easier to stoop the next time.* Katherine Hepburn

Theoretically this result is not surprising. A student who perceives the expected value of cheating to be beneficial on one question is likely to have a similar view on the remaining midterm questions. This belief relies on the assumption that the expected value of cheating increases as the expected exam score without cheating decreases. At the risk of being viewed as tautological, a major finding of this study is as follows: Students who do not study for the exam are more likely to cheat, students who cheat a little are likely to cheat a lot, and students who cheat a lot are more likely to be male.

An unusual characteristic of this multiple-choice exam was that the responses were often not limited to a., b., c., d., and e. The exam therefore did not lend itself to the use of bubble coding sheets, or a separate page dedicated to answers. While grading the exam manually and placing them into "Did Not Cheat," "Probably Cheated," and "Definitely Cheated" categories, there were frustrating cases of students erasing or crossing out the correct answer and replacing it with the correct answer to the alternate version of the exam.

Student reactions to this study were mixed and appeared in stages. The first stage was primarily negative and started with the revelation that the exam included cheater detection questions and ended prior to individual exams being returned to students. Students who did not cheat started applauding when it was revealed that cheating had been detected. The alternate group of students looked at me with eyes like daggers. Replicating this methodology to detect cheating is ill advised for instructors seeking to maximize student ratings of the quality of instruction, although ratings for the course in the present study were excellent.

Graded exams were not handed out in the classroom but were available when the student stopped by the office to pick up the exam. The graded exams were sorted into two stacks. The first stack commingled the "Did Not Cheat" and "Probably Cheated" categories; the second included the "Definitely Cheated" category. No comment was made to the student; nothing was "wrong." At that point, each student knew only that the graded exam was in one stack or the other.

The second stage of student reactions began once they had picked up their exams and calculated the probability that their telltale responses indicated cheating had occurred. Several students requested that extra points be added to their exam scores because they had credible evidence that they had in fact snatched defeat from the jaws victory on multiple questions. It is frustrating to select a wrong answer after initially getting it correct, but the refusal to add extra points was justified by the manner in which they had selected the correct response to the alternate version of the exam. The second stage of reactions included disbelief about the math underlying the indication that cheating had occurred. The students demonstrating disbelief felt that probabilities were an opinion that resulted in a number that ranged from 0 to 1. This particular group of students generally performed poorly on quantitative questions.

The third stage of reactions began a month or two after the exams were returned to students. This stage allowed sufficient time for self-reflection and discussion outside of class. A moderation of anger was observed due in part to the passage of time and because they were assured that no formal accusations of cheating would be brought to bear on any student. The long-term effect of this exam was positive on a departmental level and among students. There were no formal complaints by students, and a consensus developed that favored an environment where cheating is less likely to occur.

Replicating the study would necessitate the following steps:

1) Create multiple-choice questions; determine the four most likely incorrect response techniques and the correct response strategy.
2) Assign Version A question inputs and assumptions and Version B question inputs and assumptions; then sort the ten responses. Create five additional incorrect answers such that the overall set of 15 responses appear reasonable.
3) Print an equal number of Version A and Version B exams and manually collate the two stacks of exams into one stack: an alternating sequence of Version A and Version B exams.
4) Distribute the exams in a normal fashion while actively looking for suspicious activities such as texting, using cheat sheets, whispering, and collusive copying.
5) Collect the completed exams and re-sort them into Version A and Version B stacks.

6) Note the number of incorrect responses on each exam that were the correct response to the other version of exam.
7) Sort the exams into three stacks, "Did Not Cheat" (1 or 0 correct responses to the other version of the question), "Probably Cheated" (2 correct responses to the other version of the question), and "Definitely Cheated" (3 or more correct responses to the other version of the question).
8) Combine "Did Not Cheat" and "Probably Cheated" categories from both exam versions in alphabetical order based on the student's last name (one stack).
9) Combine the "Definitely Cheated" category from both exam versions in alphabetical order based on the student's last name (one stack).
10) Distribute the graded exam on an individual basis and do not provide any indication that the two stacks differ. Let the students determine which stack indicated that cheating occurred after a self-review of their responses.

While this study focused on student cheating, it should be noted that instructors often face the same temptation with potentially much greater penalties if they are detected. For example, an investigation is currently underway in Georgia that involves 44 schools and at least 178 teachers and principals (Severson, 2011). The case called into doubt the validity of the improvement in student learning outcomes, forced dismissals, and led to the resignation of the school superintendent.

SUMMARY

Cheating is often systemic in a university setting, but it is not typically assessed in multiple-choice exams due to a lack of awareness of detection methodologies and the awkwardness of having to deal with the results. The ironic feature of this study is that one of the cheater-detection questions required students to calculate the expected value of cheating.

It is important to note that individuals in this study were not accused of cheating either privately or among their peers, nor was there a threat of academic penalty. The value of this study lies primarily in its ability to assess the prevalence of cheating and incorporating concepts related to statistical size and power in post-exam learning outcomes.

REFERENCES

Bedford, D. Wayne, Gregg, Janie R., & Clinton Suzanne M. (2011). Preventing Online Cheating with Technology: A Pilot Study of Remote Proctor and an Update of Its Use. *Journal of Higher Education Theory and Practice 11(2), 41-58.*

Heron, Randall A., & Lie, Erik. (2009). What Fraction of Stock Option Grants to Top Executives Have Been Backdated or Manipulated? *Management Science* 55(4), 513-525.

Jacob, Brian A., & Levitt, Steven D. (2003). Catching Cheating Teachers: The Results of an Unusual Experiment in Implementing Theory. In William G. Gale & Janet Rothenberg Pack (Eds.), *Brookings-Wharton Papers on Urban Affairs 2003*. Washington, D.C.: Brookings Institution Press. 185-209.

Lie, Erik. (2005). On the timing of CEO stock option awards, *Management Science* 51(5), 802–812.

Lattman, Peter. (2010). Backdating Scandal Ends With a Whimper. *New York Times Dealbook*, 2010, November 11. Retrieved from http://dealbook.nytimes.com/2010/11/11/backdating-scandal-ends-with-a-whimper/

Mogull, Robert G. (2004). A Device to Detect Student Cheating. *The Clute Institute Journal of College Teaching and Learning*. 1(9), 17-22. Retrieved from http://journals.cluteonline.com/index.php/TLC/article/view/1984/1963

Severson, Kim (2011). Systematic Cheating Is Found in Atlanta's School System. New York Times, 2011, July 5, Education Section. Retrieved from www.nytimes.com/2011/07/06/education/06atlanta.html

I would like to thank Kevin C. Chiang, the Finance 202 class at Otago University, and faculty from the School of Commerce, Department of Finance and Quantitative Analysis for their contribution, effort, and invaluable feedback. Craig H. Wisen is an Associate Professor of Finance at the University of Alaska Fairbanks, School of Management, 303 Tanana Dr. Fairbanks, Alaska 99775, chwisen@alaska.edu, 907-474-5531 (voice), 907-474-5219 (fax).