CS-BIGS 3(1): 30-36
http://www.bentley.edu/csbigs/olinsky.pdf

Applying Monte Carlo Simulation to Determine the Likelihood of Cheating on a Multiple-Choice Professional Exam

Robert DiSario
Bryant University, USA

Alan Olinsky
Bryant University, USA

John Quinn
Bryant University, USA

Phyllis Schumacher
Bryant University, USA

This paper outlines statistical arguments used in an attempt to determine if cheating occurred on a multiple-choice exam. The arguments include the testimony in a court case involving accusations of cheating on a 100-question professional multiple-choice examination with four choices for each question. In response to the fact that the prosecution employed a witness who was an expert in statistical analysis, one of the authors was engaged by the defense to conduct an independent statistical analysis of the exam scores. The prosecution's witness utilized a simulation to demonstrate, in his opinion, the relative certainty of cheating by the defendant in the case. The authors performed their own analysis, including simulations, to counter the testimony of the prosecution. The results presented in this paper highlight the fact that in the absence of definitive proof, in spite of a statistical analysis of data, there is still a need to make subjective interpretations when trying to decide if cheating has occurred on a multiple-choice test.

Keywords: *Monte Carlo Simulation, Exam Cheating Detection, Crystal Ball*

1. Introduction

Most teachers, at some time in their careers, have been faced with the dilemma of what to do in the case of suspected cheating on in-class exams. Cheating on multiple choice exams can be especially hard to confirm. Attempts to verify cheating typically require computer intensive analyses such as simulations. The following paper outlines an example of the use of statistical analysis to try to decide if cheating has occurred on a multiple choice examination. It is an interesting example as it involves a case which ended up in litigation and statistical arguments are provided both to prove and to disprove the occurrence of cheating. It does, in the final analysis, display the impossibility of any definitive answer to the question in the absence of eye-witness testimony and so highlights the importance of assuring the integrity of testing materials prior to an examination and

appropriate proctoring of examinations.

In the spring of 2006, one of the authors was asked to analyze data and determine the likelihood of two individuals cheating on a professional multiple-choice examination, which consisted of 100 questions with four possible answers each. One of the two individuals was subsequently charged with cheating on the exam. The results of the analysis were presented as testimony for the defense by the author in a courtroom. This testimony was necessitated by the prosecuting attorney bringing in their own statistician as an expert witness.

This case involved two individuals, who both answered 93 out of 100 questions correctly on this examination. They both missed the same 5 questions with the same choice of answers and also missed 2 additional different questions each. The exam was carefully proctored at a local community college and these individuals were not seated near each other. There were no accusations of the use of electronic devices or other types of communication between these 2 individuals. The opportunity for cheating occurring during the exam was practically nonexistent.

A Short Literature Review of Detecting Cheating on Multiple-Choice Exams

Methods for detecting cheating or collusion typically utilize probability distributions and tests of significance to identify student pairs suspected of cheating. Some methods rely on the detection of outliers which generate indices showing unusual patterns of answers between pairs of respondents (Wesolowsky, 2008). For example, Nelson (2006) explored the use of a specific index in trying to detect cheating on multiple choice examinations, while Post (1994) utilized various probability models, as logit and probit, to examine cheating on these examinations. Bellezza and Bellezza (1989), updated in Bellezza and Bellezza (1995), included the use of the binomial distribution to examine the likelihood of pairs of students obtaining the exact same wrong answers on multiple choice tests. Cizek (2001) provides a detailed background of cheating, including describing some statistical techniques used in the detection of cheating and steps that could be used to avoid it. Harpp and Hogan (1996) described the use of software to detect cheating by pairs of students taking multiple choice exams. Palazzo (1996) examined cheating on online homework software but also considered cheating on multiple choice tests, for which he provided a review. McManus, Lissauer, and Williams (2005) examined the results of medical postgraduate multiple choice examinations in the UK to look for patterns of similarly answered questions by use of a computer program, and applying regression. Sundermann (2008) considered cheating on multiple choice tests by looking at the Harpp-Hogan Index which is the ratio of the number of exact errors in common to the number of different responses (EEIC/D). Mogull (2004) examined cheating on multiple choice exams by using joint probabilities, assuming independence, to calculate the likelihood of students answering with the same incorrect alternative for a given question. In addition, Feinberg and Kadane (1983) reviewed and applied Bayesian analyses to various legal proceedings, in general.

Wesolowsky (2000) has done much research in this area and has developed software which is intended to identify cheating. Wesolowsky's method looks for pairs of students with unusually high numbers of uncommon answers. He models non-cheating behavior since this requires the fewest assumptions. His software is suitable for screening large classes. His method also uses the Bonferroni inequality to prevent false accusation due to "dredging" the data. In practice at McMaster University, all multiple choice exams given in large lecture hall settings are run through his software. Pairs of students are flagged by the program. This list is then compared to the actual location of the students at the exam location. This information is then available for further consideration of possible cheating. A major application of Wesolowsky's software was to investigate accusations of cheating in Texas with a statewide assessment program called the Texas Assessment of Knowledge and Skills (TAKS) which is administered to students in grades 3 through 11. The analysis from the program indicated that there was "substantial" cheating (Benton and Hacker, 2009).

We had initially explored the use of Wesolowky's program to analyze the test results. However, this technique is mainly recommended for use in confirming cheating in situations where students are sitting near each other at a testing facility which was not the case here. We chose to use a Monte Carlo simulation since it was thought that it would be easier to explain a simulation in court as Wesolowsky's program is predicated on probability theory that would be difficult to explain to a layman. Finally, when the expert witness for the prosecution utilized a simulation, we felt it necessary to respond in court in a similar fashion.

2. Prosecution Testimony

The prosecution claimed that the two individuals suspected of cheating had been given source materials. The source materials in question did not include questions or answers but rather the sections of the professional code from which the questions were selected.

In addition to their other witnesses, the prosecution invited an expert statistical witness who testified that the probability of two test takers out of 64, with no advance information, receiving a score of 93 with 5, 6 or 7 wrong answers of the 7 answered incorrectly by one of the two individuals with the same choices was only 14 out of 100,000 or .00014. He claimed to have determined this probability by simulating the results of a class of 64 students taking this exam 100,000 times. The prosecution witness further stated that this exam would have to be administered 7,000 times to see this kind of match, since .014% of 7000 is approximately 1. Since the two individuals in question had this kind of match, he concluded that there was practically certain evidence of cheating.

He also used a comparison of percentiles ranks of the defendant on the current sitting and the previous sitting of the exam, in order to support the prosecution claim. He claimed that the defendant's rank increased by over fifty percent in the two years between sittings. We subsequently found that his percentiles were not accurate and we were not able come get the same results when replicating the simulation. Finally, the prosecution mentioned the fact that these two individuals in question are now married to each other to support the charge of their having had advance information.

3. Defense Testimony

3.1 Percentiles of Exam Ranks

We will address the question of percentiles first. The correct calculation of percentiles are provided in Table 1 for the top scoring test takers in the 1994 and 1996 test sittings. The defendant's score appears in bold. The percentiles are calculated by a method used by David Stockburger (2008). The scores are first ranked from highest to lowest and then for each score, the percentile is calculated by adding the percentage of values below the score and one half of the percentage of values equal to the score. Therefore, in the 1996 sitting, the defendant's scored better than 61 of 64 candidates and two candidates shared the same rank so the percentile is 61/64 plus 1/2(2/64) = 96.8% It can be seen that there was a change from the 82.76th percentile to the 96.88th percentile from an earlier exam for the defendant. This increase of 14.12 percentage points is much lower that the change claimed by the prosecution. It should also be noted that the defendant did rank in the top 10 (out of 58 candidates taking the exam) on the previous exam and that although the ranking changed from 10th to 2nd for this individual, this could easily be explained by the fact that she had two additional years to prepare for the exam. In fact, the individuals with the top 10 scores are promoted. The defendant therefore should have been promoted in 1994. However two questions were challenged by other test takers and the correct answers for these two questions were subsequently changed. The tests were then re-graded and consequently, this individual dropped out of the top 10 and was not promoted.

Table 1. Top 12 Grades, Ranks and Percentiles for Students Taking Exam in 1994 and in 1996

1994 Exam			1996 Exam		
Exam Grades	Rank	Percen-tile	Exam Grades	Rank	Percen-tile
93	1	99.14	94	1	99.22
91	2	97.41	93	2	96.88
89	3	95.69	93	2	96.88
84	4	93.1	91	4	94.53
84	4	93.1	90	5	92.97
83	6	90.52	89	6	91.41
82	7	88.79	88	7	89.84
81	8	87.07	87	8	87.5
80	9	85.34	87	8	87.5
79	10	82.76	86	10	85.16
79	10	82.76	85	11	82.81
77	12	80.17	85	11	82.81

3.2 Probability of Getting Questions Wrong with a Matching Choice

As an initial response by the defense, prior to hearing the prosecution witness, we had provided a more realistic prospective, then the very narrow simulated probability which he obtained by matching the 7 incorrect answers of one of the individuals.

First of all, we analyzed all 9 questions that the two suspected candidates missed in common and differently. These questions are outlined in Table 2. This table contains 9 questions, each of which was answered incorrectly by one of the two individuals. The 5 questions that they answered incorrectly with the same answer choice are marked with an X. The next column gives the percent of all test takers answering this question incorrectly. The last column in the table contains the percent of all test takers answering this question with the same wrong choice as the defendant. In Table 3, the 5 questions with the same incorrect answer are tabulated for all 64 test takers. It can be seen from Table 2 that for the 5th question that the defendant answered incorrectly, 96.9% of the test takers answered incorrectly and actually chose the same incorrect answer. From Table 3, we see

that this represents 62/62 or 100% agreement among those who answered it incorrectly. This is strong evidence that this is a poor question and should not be considered when calculating subsequent probabilities. The same can be said for the 3rd question since nearly one half of the test takers (43.5%) answered this incorrectly and all choose the same incorrect answer. The 7th incorrect answer was also missed by a large number of test takers (79.7%) with 73.4% choosing the same incorrect answer, which from Table 3 is equivalent to 92% of the incorrect answers being the same. In consideration of the percentages in Table 3, we can see that if one looks only at the questions answered with the same incorrect answers as the accused, one can see that the choice was the most popular with all candidates except for possibly the first question where even there 45% of those who answered incorrectly match the defendant's choice. Nevertheless, we proceeded with the data as provided in accordance with the approach taken by the prosecution.

Table 2. Questions Answered Incorrectly by the Individuals Suspected of Cheating

Question	Defend-ant	Co-def-endant	Same	% with wrong wnswer	% with same wrong choice
1	C	C	X	31.3	14.1
2	B	A			
3	D	D	X	45.3	45.3
4	A	D			
5	A	A	X	96.9	96.9
6	D	D	X	54.7	37.5
7	D	D	X	79.7	73.4
8	D	A			
9	B	A			

3.3 Probability of Any Pair of Test Takers with Matching Wrong Answers

As another argument, we considered all possible pairings of the 64 test takers. Using a counting formula for combinations (${}_{64}C_2$) there are 2,016 possible pairs. Upon analysis, we found 73% of these comparisons with at least 5 matching exact wrong answers. One pair even had as many as 21 matching incorrect answers. In addition, we generated a list of all pairs in which both test takers achieved a score of at least 80. There were 55 such pairs (85.9%) in which both test takers had at least 5 matching exact wrong answers. In fact, 33 pairs (51.6%) with 5, 12 pairs (18.8%) with 6, 8 pairs (12.5%) with 7, and 2 pairs (3.1%) had exactly 8 matching exact wrong answers.

Table 3. Five Questions Answered with Same Wrong Answer by the 64 Candidates

Number with incorrect answer	Number with defendant's answer	% matching defendant
20	9	45
29	29	100
62	62	100
35	24	69
51	47	92

3.4 Probability of a Test Taker Matching a Specific Set of Wrong Answers

We then determined that if a test taker randomly answered any seven questions, the probability of 5 or more exact matches with another test-taker is .01284 a number which, although small, is 100 times greater than their witness's value. This probability is obtained by using a Binomial probability function with n = 7 and the probability of match, p = .25. Thus if X = the number of matches out of 7, P(X = 5, 6, or 7) = .0115 + .00128 +.0006 = 0.01284. Also, it has to be remembered that this last consideration is a worst case scenario, as we are assuming that the individual is purely guessing randomly on each of the 7 questions. In fact, if reasonable people made educated choices, this probability would be much higher. Based on the Bellezza's paper (4), using the probability of p = .4, which is the recommended adjusted probability, since all incorrect alternatives are not equally likely, recalculation of the previous probabilities yields a probability of .0962. This probability of .4 is based on five-alternative answer questions and since the questions on this test were four-alternative, Belleza suggests that the probability of a match may actually be higher than 0.4.

3.5 Monte Carlo Simulation

Not knowing what approach would be taken by the prosecution's expert witness, we independently had obtained a Monte Carlo simulation of this class retaking the exam ten thousand times. This simulation used a probability distribution which matched the likelihood of responding to each question with a certain alternative with the relative frequency which that alternative was chosen by the class in question. For example since question number 1 was answered by the 64 members of the class with the probability distribution in Table 4, these percentages were used to generate answers for question 1 in the simulation. This procedure was used to simulate answers to all 100 questions for 64 students.

Table 4. Answers to Question 1 Provided by Test Takers

Choice	Number Selecting Choice	Percent Selecting Choice
A	1	1.56
B	0	0.00
C	8	12.50
D	55	85.94
Total	64	100.00

The results of the simulation, which were obtained using Crystal Ball, an EXCEL add-in, are presented in Figure 1. The horizontal axis represents the number of incorrect exact matches and the vertical scale is the corresponding probability. As calculated by Crystal Ball and denoted at the bottom of the table the total probability of getting 5 or more was found to be 90.6%. The interpretation of this is that there was over a 90% chance of five or more matches on exact wrong answers by test takers on this exam. It would seem that this is a very common event and certainly not rare as when simulating a match to a specific test takers incorrect answers. It should be noted that only 9,917 of the 10,000 simulations are displayed as Crystal Ball omits the display of outliers by default.

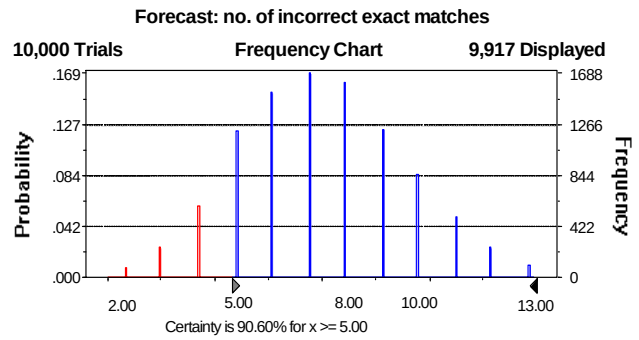

Figure 1. Simulation of Total Number of Incorrect Exact Matches

3.6 Replication of Prosecution Simulation

We then considered a rebuttal of the simulation results provided the prosecution witness. In addition to the fact that we could not reproduce his results, we felt that the attempt to focus on the specific 7 questions answered incorrectly by the one of the individuals was too narrow. After hearing the testimony of the prosecution's expert witness, we replicated the simulation that he apparently ran. He provided no documents in support of his simulation and we were forced to proceed on the basis of his oral testimony.

We ran a simulation, in which we replicated 10,000 exams of these 64 test takers, again utilizing the actual relative frequencies with which each alternative was chosen on each question. Although we believe that it is improper to select the individual accused of cheating for comparison with all of the other simulated students, we wanted to replicate the prosecution results. Therefore, against our better judgment, we kept the defendant's scores as a constant and ran a simulation in which we examined the number of exact wrong answers that became matches with his 7 wrong problems. In this simulation, we do not require that the grade on the test match the score of 93 which the defendant achieved. The results of this simulation are presented in Figure 2. In fact, after 10,000 runs, we found that approximately 10% of the class achieved at least 5 of the same wrong answers as the defendant. It is interesting to note that this probability is very close to the probability of .0962 obtained by using the Binomial formula earlier with the probability of .4 suggested by Belleza.

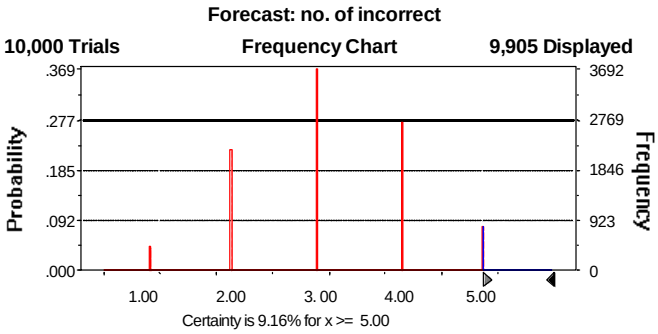

Figure 2. Simulation of Number of Incorrect Exact Matches with One of the Defendants

We then adjusted the simulation to only include the choices of the better students on the 100 questions. Therefore, we chose the top 10 students, with the rationale that these students, subject to other factors, would be promoted. These students had grades of 85 and above. In rerunning the simulation, *using the responses of these 10 individuals to obtain the relative frequency of responding with a particular choice to any question* and again comparing 10,000 individual test takers to the defendant, we found that approximately 7% of the simulated test takers matched exactly on at least 5 of the 7 wrong answers in comparison with the defendant. These are not small probabilities and dramatically higher than those achieved by the prosecution. Again, we are not sure exactly how he framed the problem or assumptions that he made.

However, according to their expert witness's testimony, he ran a simulation that determined the probability of

another test taker not only matching on at least 5 of the same wrong answers, but also achieving a 93 or higher. It should be noted that only 1 student actually had a grade higher than 93. It was in running this simulation that he determined a probability of 14/100,000 of another student matching one of the defendants. Therefore, this unlikely event, and the fact that the other individual in question achieved this, led him to his conclusion that they must have had advance information.

To see for ourselves, we added the stipulation of matching on at least 5 of the same exact wrong answers as one of the defendants and also achieving at least a 93. The use of 100,000 is probably overkill, but we wanted to match our results with that of the prosecution. In fact, when we ran the simulation, we found no one achieving such a match whereas he found 14 out of 100,000 tests.

Although this seemed discouraging from our point of view, we noted that Student 102 received a 91 on the exam and had exactly 3 incorrect matches with one of the defendants. Using the same logic as their expert witness used, we ran a simulation to determine the probability of another test taker receiving at least a 91 and having at least 3 exact matching wrong answers with the defendant. The resulting probability is also close to zero as in the simulation run by their expert. However, there was such an individual and we could use their expert's logic that these two individuals must have also had advance information and cheated. Yet there was no reason to believe that this was true and no such claim by the prosecution!

4. Conclusion

Our initial reaction, which hasn't changed, is that, in this case, with these circumstances, there is not proof of cheating based on a statistical model or simulation. The only way in which the statistical evidence suggested by the prosecution's expert witness might carry any weight is if the two accused individuals sitting side by side in the examination room and physically copied from each other. Then perhaps the statistical simulation might make sense. However it is a fact that this was not the case. In addition, we also have some issues with how the expert witness developed his model and the assumptions which he used which led to such a low probability of occurrence, since we could never replicate his results.

Even if the individuals in question were given the answer key, which contained all of the correct choices to all of the questions ahead of time, and they memorized all of these solutions, and they certainly haven't been accused of having the answer key, there would be no reasonable way in which they would miss the same questions with the same answers. It would be more likely that they might forget some of the answers and get some wrong, but there is no reason to believe that they would match on the exact same wrong answers.

In addition, the accusation that is being made is that they had source material. Even if they studied the referenced material, they would not have known the answers or the order in which they appeared. Therefore, there is no reason to believe that a statistical model would be able to demonstrate that cheating occurred. Indeed, if the two individuals studied together, which they denied, and had formed the same incorrect interpretations of the material, they may have gotten the same wrong answers. However, the questions which they missed with the same answers were answered with the same choice by a majority of all test takers who answered incorrectly.

We also would like to point out that the only way to test for cheating on multiple choice tests is to examine matching wrong answers. What if, in fact, two students cheated and both achieved 100 percent? We are not sure how any of the arguments presented in this analysis would apply. Also, the statistical method would change dramatically if a question were grieved, which commonly happens on this type of exam, and an incorrect response was changed to a correct response. This would result in one less match on exact wrong answers and would have a fairly large effect on the probability calculation.

In conclusion, we believe that the statistical approach used by the prosecution would be valid only in the circumstance where students are accused of physically cheating during the examination and, therefore, there is no proof of cheating based on the statistical evidence presented by the prosecution's expert witness. Indeed, although we rebutted the statistical argument by analyzing the data and illustrating that statistics can be used to argue that the two candidates did not cheat, it is important to keep in mind the caveat, posed by Cizek, concerning the use of statistical analysis when trying to either prove or disprove cheating. "It is important to note, however, that statistical methods do not obviate the need for human judgment. Even once test results are shown to be highly unlikely, human rationality must be invoked to come to any conclusions about whether alternative causes represent more plausible explanations for the results; that is, there still exists a need to make subjective interpretations about whether the unlikely events represent cheating." (Cizek, 2001, pg. 11).

Unfortunately, the original decision went against the defendant. The case was later appealed to the Superior Court, where the original decision was vacated due to procedural issues.

REFERENCES

Bellezza, Francis S. and Bellezza, Suzanne F. 1989. "Detection of Cheating on Multiple-Choice Tests by Using Error-Similarity Analysis," *Testing of Psychology*, 16(3): 151-155.

Bellezza, Francis S. and Bellezza, Suzanne F. 1995. "Detection of Cheating on Multiple-Choice Tests: An Update," *Testing of Psychology*, 22(3): 180-182.

Benton, Joshua and Hacker, Holly. 2009. "Analysis shows TAKS cheating rampant," The Dallas Morning News, retrieved 3/23/2009 from website: http://www.dallasnews.com/sharedcontent/dws/dn/education/stories/060307dnmetcheating.433e87c.html .

Cizek, Gregory J. 2001. "An overview of issues concerning cheating on large-scale tests," paper presented at annual meeting of the *National Council on Measurement in Education*, April 2001, Seattle, WA.

Feinberg, Stephen E. and Kadane, Joseph B. 1983. "The Presentation of Bayesian Statistical Analyses in Legal Proceedings," *The Statistician*, 32: 88- 98.

Harpp, David N. and Hogan, James J. 1996. "Crime in the classroom: Part II. An Update," *Journal of Chemical Education*, 73(4): 349-351.

McManus, I.C., Lissauer, Tom, and Williams, S.E. 2005. "Detecting cheating in written medical examinations by statistical analysis of answers: pilot study, *British Medical Journal*, 330(7499): 1064-1066.

Mogull, Robert G. 2004. "A Device to Detect Student Cheating," *Journal of College Teaching & Learning*, 1(9): 17-21.

Nelson, Larry R. 2006. "Using selected indices to monitor cheating on multiple-choice exams," *Thai Journal of Educational Research and Measurement, 4(1):* 1-18.

Palazzo, David, J. 1996. "Detection, Patterns, Consequences, and Remediation of Electronic Homework Copying," Submitted to the Department of Physics in Partial Fulfillment of the Requirements for the Degree of Master of Science in Physics, June 1996, Massachusetts Institute of Technology. Retrieved July 10, 2008 from http://dspace.mit.edu/bitstream/1721.1/36817/1/82369073.pdf

Post, Gerald V. 1994. "A Quantal Choice Model for the Detection of Copying on Multiple Choice Examinations," *Decision Sciences*, 25(1): 123-142.

Sundermann, Michael J. 2008. "A Statistical Analysis of Infrequent Events on Multiple-Choice Tests that Indicate Problem Cheating," *Journal of Chemical Education*, 85(4): 568-571.

Stockburger, David W. 2008. *Introductory Statistics: Concepts, Models, and Application*, Retrieved on July 7, 2008 from http://www.psychstat.missouristate.edu/introbook/sbk14m.htm

Wesolowsky, George O. 2000. "Detecting excessive similarity in answers on multiple choice exams," *Journal of Applied Statistics*, 27(7): 909-921.

Wesolowsky, George O. 2008. website: Retrieved July 7, 2008 from http://www.business.mcmaster.ca/msis/profs/wesolo/wesolo.htm .

Acknowledgments: An earlier version of this paper was presented at the Northeast Decision Sciences annual meeting in April, 2006 in San Juan, PR. Robert DiSario is now deceased.

Correspondence: aolinsky@bryant.edu