Checking for Collaboration in Online Multiple Choice Testing

Harsha Perera, Rajitha M. Silva, and Tim B. Swartz *

Abstract

In this paper, we propose a novel two-step procedure to detect potential collaboration among students during multiple-choice tests, with the aim of maintaining academic integrity in online education and testing settings. Cheating in online environments can involve much larger groups of students, and therefore, traditional pairwise detection methods may not be effective. In the first step, we identify suspicious common responses using probabilistic reasoning and apply the UPGMA algorithm to cluster students who may have collaborated. In the second step, we calculate the probability of timing-related patterns among clustered students to provide additional evidence of collaboration. We provide an example of how to implement these two steps and demonstrate their effectiveness in identifying potential cheating incidents. Our proposed method offers a practical solution for maintaining academic integrity in online education and testing settings.

Keywords : cheating, clustering, geometric mean, probabilistic reasoning.

*H. Perera is Lecturer, and T. Swartz is Professor, Department of Statistics and Actuarial Science, Simon Fraser University, 8888 University Drive, Burnaby BC, Canada V5A1S6. Rajitha M. Silva is Senior Lecturer, Department of Statistics, University of Sri Jayewardenepura, Nugegoda, Sri Lanka. Swartz has been partially supported by the Natural Sciences and Engineering Research Council of Canada. Conflict of interest: The authors declare that they have no competing interests, and all authors confirm accuracy. Availability of data and materials: The data used to support the findings of this study are available from the corresponding author upon reasonable request.

1 INTRODUCTION

The COVID-19 pandemic promoted a surge of online instruction in the education sector where it has been argued that much of the infrastructure is here to stay (Li and Lalani 2020). Associated with online instruction is online testing which poses obvious challenges related to course integrity. Online invigiliation of students using cameras is seen by some as intrusive and a violation of privacy (Coghlan, Miller and Paterson 2021). Moreover, it is clear that invigilation with cameras does not fully prevent online cheating. For example, off-camera activities which constitute cheating would be unobserved.

With the complications associated with online environments, many instructors have chosen multiple choice testing for its relative simplicity. However, multiple choice testing also faces integrity issues. For example, there is evidence of increased contract cheating in STEM subjects (Lancaster and Cotarian 2021).

In this paper, we propose integrity checks for multiple choice testing that follows a specific protocol. First, multiple choice tests are given fixed time durations which are set to pose a challenge to students. Specifically, tight time constraints would not allow students the time required to seek extensive help from alternative sources such as tutoring agencies. Second, in this framework, multiple choice questions are required to be the same for each student (which manages workload for the instructor), and the questions arrive in random order for each student. The random order is an important component of the protocol as it does not allow students (sitting in the same room or in a chat room) to work on the same question at the same time. The delivery of questions in a random order is easily facilitated by educational platforms such as Canvas. Third, our protocol insists that students must answer a question before moving on to a subsequent question, and may not return to modify a previously answered question. Although this requirement is unpopular with students, it is an important component in retaining integrity. Again, this type of requirement is easily implemented in educational platforms such as Canvas.

Our integrity check is a two-step process. In the first step, students are clustered accorded to the similarity of their selection responses in the test. Once suspicious groups

of students have been identified, then additional probability calculations are obtained for their timing patterns. The time taken to respond to questions is also provided by some educational platforms such as Canvas. We suggest that confirmation of collaboration in both of the two steps provides a strong indication of cheating.

There exists a considerable literature concerning the detection of cheating in multiple choice testing. However, because of the pervasiveness of multiple choice testing across disciplines, the research tends to be dispersed across seemingly unrelated journals. For example, Belezza and Belleza (1989) developed a technique called error similarity analysis is based on a probability calculation for the number of incorrect answers between pairs of examinees using an underlying Binomial distribution. A common Binomial parameter (i.e. for all questions) was estimated from the dataset of incorrect answers. Nath and Lovaglia (2009) provided a probabilistic assessment of cheating which is based on the pairwise comparison of students. A cutoff probability of detection 0.001 was utilized which does not appear to take into account the number of students (i.e. the issue of multiple comparisons) nor the number of test questions. Wesolowsky (2000) also provided pairwise detection techniques. The approach was based on a compound Binomial probability calculation which is then approximated by a normal distribution. A feature of the approach is that it accounts for student ability and differentiates between test questions. Richmond and Roehner (2015) developed methods for pairwise detection that were not inspired directly via probability. Instead, they introduced a diagnostic based on the number of common correct questions between two subjects (referred to as "overlap") as a function of the geometric mean of the correct number of responses by the two subjects. As with our approach, all of the aforementioned papers recognize that similarity of incorrect responses is more indicative of cheating than similarity of correct responses.

Unlike our methods, none of the above approaches use traditional clustering techniques nor a secondary check based on response times. The introduction of the second step regarding response times appears to be an additional diagnostic tool that has not been previously considered. We also emphasize that cheating behaviours in an online environment may differ from those in a fully proctored environment. For example, group cheating

(as opposed to pairwise cheating) appears more plausible in an online environment.

In Section 2, we set up the problem of interest by listing the ways in which a student may cheat under multiple choice testing subject to specific protocols. In Section 3, the first step of the two-step process is described. Groups of students who may have collaborated are identified using a hierarchical clustering algorithm. The similarity measure used in the clustering algorithm to distinguish students is developed based on probabilistic reasoning. Section 4 considers the groups of students identified by the clustering algorithm and then develops a probabilistic measure to detect whether the timing of their responses is unusual. Unusual responses provide further evidence of collaboration. Section 5 provides a real data example under anonymity. In this dataset, it is unknown whether students have collaborated. The two-step procedure is applied to this data, and the results are explored. In Section 6 we conclude with a short discussion.

2 HOW MIGHT A STUDENT CHEAT?

We repeat the required protocol for multiple choice testing: First, the multiple choice test is given a fixed time duration which is set to pose a challenge to students. Second, the multiple choice questions are the same for each student but arrive in random order to each student. Third, the protocol insists that students must answer a question before moving on to a subsequent question, and may not return to modify a previously answered question.

Under this protocol, we envision the following ways in which a student may cheat in a multiple choice testing environment consisting of m questions:

1. A student may enlist a non-student to write the test.

2. A student may seek assistance from a non-student or a prohibited resource.

3. A group of students may collaborate where one student is designated the "baseline" student. The students work together to answer question 1 for the baseline student. If any of the collaborating students have the same question 1 as the baseline student,

then they also respond. Then the students work together to answer question 2 for the baseline student. If any of the collaborating students have a current question which corresponds to either question 1 or question 2 for the baseline student, then they respond. The process continues until the baseline student has completed all m multiple choice questions, at which time the remaining students will have also completed all of their questions.

Our testing procedure only addresses the third type of behaviour which we attempt to identify. The key insight is that some of the collaborating students (not the baseline student) will exhibit extremely short response times for some of their questions (i.e. questions which had been previously answered by the baseline student). These students will also exhibit some long response times. Probability calculations are developed in Section 4 to see if the timing patterns are unusual.

3 STEP-1: CLUSTERING OF STUDENTS

Consider a class of N students taking m multiple choice questions according the protocol described at the beginning of Section 2. Let $X_{ijk} = 1(0)$ correspond to the selection (non-selection) of response k corresponding to question j by student i where $k = 1, \ldots, n_j$. Often in multiple choice testing, $n_j = 4$ for all questions $j = 1, \ldots, m$, but we retain the flexibility of allowing different numbers of responses for individual questions. Under the assumption that the students consist of a random sample from a population of students, the standard statistical model using this framework is

$$(X_{ij1}, X_{ij2}, \ldots, X_{ijn_j}) \sim \text{multinomial}(1, p_{j1}, p_{j2}, \ldots, p_{jn_j}) \tag{1}$$

where p_{jk} denotes the probability that the i-th student selects response k for the j-th question such that $\sum_{k=1}^{n_j} p_{jk} = 1$. In (1), the parameters p_{jk} are unknown but may be simply estimated by $\hat{p}_{jk} = (1/N) \sum_{i=1}^{N} x_{ijk}$ where x_{ijk} is the observed response corresponding to the random variable X_{ijk}.

A first reaction may be to consider the number of common responses between student i_1 and student i_2 given by the similarity statistic

$$s_{i_1 i_2}^{(1)} = \sum_{j=1}^{m} \prod_{k=1}^{n_j} x_{i_1 jk} x_{i_2 jk}. \tag{2}$$

However, a difficulty with (2) is that more context is required when assessing the possibility of collaboration. For example, consider two students who have identical responses on all test questions, i.e. $s_{i_1 i_2}^{(1)} = m$. There is a difference between the case where the two students both obtained perfect scores (possibly bright students working independently) and the case involving two students whose agreement on some questions involve responses that are incorrect and unlikely to be selected. There is a greater suspicion of collaboration when both students choose a highly unpopular response.

Therefore, we propose the more complex similarity statistic

$$s_{i_1 i_2}^{(2)} = \sum_{j=1}^{Q_{i_1 i_2}} \Phi \left(\frac{1 - \hat{p}_{k_j}^2}{\sqrt{\hat{p}_{k_j}^2 (1 - \hat{p}_{k_j}^2)}} \right) \tag{3}$$

where $Q_{i_1 i_2}$ is the subset of questions from $(1, \ldots, m)$ where students i_1 and i_2 have common responses, $k_j \in (1, \ldots, n_j)$ is the common response for question j, $\hat{p}_{k_j}$ is the estimated probability of response k_j, and Φ is the cumulative distribution function for the standard normal distribution. The idea is that the two students are more similar (i.e. $s_{i_1 i_2}^{(2)}$ increases) when the suspicion increases that they have collaborated.

The motivation for (3) begins with the assertion that there is nothing suspicious (in terms of collaboration between two students) when they provide different responses to a question. This is why the summation in (3) is restricted to the set $Q_{i_1 i_2}$. We then note that when two students provide the same response k_j to question j, this corresponds to an event which is Bernoulli($p_{k_j}^2$) under the assumption that there is no collaboration. The suspicion of collaboration is greater for smaller values of p_{k_j} (i.e. less likely responses). Therefore, the occurrence of the event may be standardized and approxi-

mated by $v = (1 - \hat{p}_{k_j}^2)/\sqrt{\hat{p}_{k_j}^2(1 - \hat{p}_{k_j}^2)}$. The quantity $v \in (0, \infty)$ may be thought as an index of suspicion involving collaboration where larger values of v are more suspicious. And given the standardization, the quantity $\Phi(v)$ has the analogous appearance of a probability that increases as $\hat{p}_{k_j}$ decreases. Finally, as probabilities of independent events (i.e. test questions) ought to be multiplied, the logarithm is introduced.

3.1 Clustering

Cluster analysis has an extensive history dating back to at least Driver and Kroeber (1932). In cluster analysis, the objective is to group "similar" objects into "clusters". Many approaches and algorithms have been proposed and a survey of the clustering literature is given by Xu and Tian (2015).

Our approach uses one of the most elementary and popular clustering algorithms known as UPGMA (unweighted pair group method with arithmetic mean) clustering. UPGMA has been attributed to Sokal and Michener (1958) and falls under the general framework of hierarchical clustering. An appealing feature of UPGMA is that users can easily write their own clustering code specific to their application. Every hierarchical clustering algorithm is reliant on a similarity (or dissimilarity) measure which enables the comparison of observations. In our implementation of UPGMA clustering, we use the measure given by (3). According to development of (3), $s^{(2)}_{i_1,i_2}$ is a probabilistic measure of suspicion of collaboration between students i_1 and i_2 derived under the assumption that they were acting independently. Larger values of $s^{(2)}_{i_1,i_2}$ are more suspicious and therefore the metric serves a similarity measure between the two students.

The UPGMA algorithm is simply described. It is an agglomerative algorithm which begins with N clusters (i.e. each student belongs to its own cluster). However, in a given iteration of the algorithm, let C_i denote the set of students in cluster i and let n_i be the number of students in cluster i. The following iterative steps are followed:

1. calculate the average cluster distance between each pair of clusters i and j, $D_{ij} = \frac{1}{n_i+n_j} \sum_{p \in C_i} \sum_{q \in C_j} s^{(2)}_{pq}$

2. merge the two clusters i and j for which D_{ij} is minimum

Note that the UPGMA algorithm (as described above) continues until there is a single cluster remaining which contains all students. Therefore, for practical purposes, a stopping criterion is required. We terminate the algorithm when a threshold number M of clusters is attained. This identifies M groups of students where each group may be further investigated. To simplify the investigation, it seems unlikely that the largest cluster would be deemed suspicious.

4 STEP-2: PROBABILITIES WITH RESPECT TO TIMING PATTERNS

Having identified a cluster of students with similar selection responses using the Step-1 methodology of Section 3, we now investigate the timing response patterns associated with these students.

We let T_{ij} denote the response time taken by the i-th student on question j of the multiple choice test, $i = 1, \ldots, N$ and $j = 1, \ldots, m$. Corresponding to the random variable T_{ij}, we let t_{ij} denote the observed response. Although times are strictly positive, it may be reasonable to assume

$$T_{ij} \sim \text{Normal}(\mu_j, \sigma_j^2)$$

where estimates $\hat{\mu}_j$ (the sample mean) and $\hat{\sigma}_j$ (the standard deviation) are obtained from the data $\{t_{ij}\}$.

Recall that an observed timing response t_{ij} can be extreme in either the small or large sense according to the third type of behaviour as outlined in Section 2. Therefore, an

approximate probability corresponding to the occurrence of t_{ij} is given by

$$\begin{aligned} q_{ij} &= \begin{cases} \text{Prob}(T_{ij} > t_{ij}) & t_{ij} > \hat{\mu}_j \\ \text{Prob}(T_{ij} < t_{ij}) & t_{ij} < \hat{\mu}_j \end{cases} \\ &= \begin{cases} 1 - \Phi((t_{ij} - \hat{\mu}_j)/\hat{\sigma}_j) & t_{ij} > \hat{\mu}_j \\ \Phi((t_{ij} - \hat{\mu}_j)/\hat{\sigma}_j) & t_{ij} < \hat{\mu}_j \ . \end{cases} \end{aligned} \tag{4}$$

Therefore, small values of q_{ij} are interpreted as unusual and provide evidence of collaboration. We then provide a summary measure for the i-th student which is given by the geometric mean of the probabilities in (4) taken over all questions

$$\begin{aligned} G_i &= \left(\prod_{j=1}^{m} q_{ij} \right)^{1/m} \\ &= \exp\left\{ \frac{1}{m} \sum_{j=1}^{m} \log q_{ij} \right\} \ . \end{aligned} \tag{5}$$

The quantity G_i can be thought of as an average probability over the set of test questions where small values of G_i provide evidence of collaboration. The second expression in (5) is more amenable to calculation and we emphasize that the diagnostic G_i is readily interpretable.

However, according to the third behaviour described in Section 2, we note that not all students involved in a collaboration will exhibit small values of G_i. Specifically, the baseline student will answer questions with ordinary timing patterns. Therefore, within a cluster, we are seeking small values of G_i for some students as evidence of collaboration. If some students in a cluster are discovered, it does not absolve those students with moderate values of G_i from collaborating.

5 EXAMPLE

We used the midterm Exam 3 data keeping everything anonymous for STAT 201- Statistics for the Life Sciences. We had 181 students and 15 questions in the exam which delivered via canvas. Here, we are going to describe the results of both steps (section 3 and 4), and identify some suspicious students. In Step 1: using the UPGMA algorithm we identified five clusters shown in Table 1 and the dendrogram of the clusters shown in Figure 1.

Cluster Number	**Number of Students**
1	71
2	57
3	42
4	50
5	51

Table 1: Cluster and Student Data

UPGMA algorithm continues until there is a single cluster remaining which contains all students. Therefore, for practical purposes, a stopping criterion is required. We terminated the algorithm when a threshold number M=5 of clusters is attained. This identified 5 groups of students and to simplify the investigation we ignore the three largest clusters, since it seems unlikely that the largest clusters would be deemed suspicious.

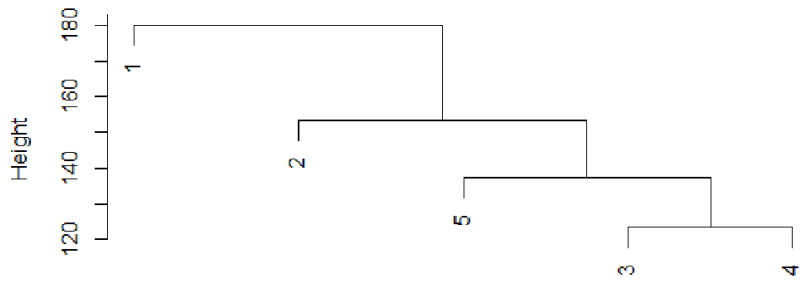

Figure 1: Dendrogram of student clusters for midterm exam 3

After identifying the two smallest clusters (suspicious clusters), we then investigated

the timing-response patterns associated with these students using the quantity G_i explained in section 4. Within a cluster, we seek small values of G_i as evidence of collaboration. However, the presence of some students with low G_i values does not necessarily mean that those with moderate values are not collaborating. We, therefore, obtained histograms of the G_i values within suspicious clusters and used the 25th percentile as the threshold. In Figure 2, the distributions of the G_i values are displayed for each suspicious cluster, with a vertical line at the 25th percentile.

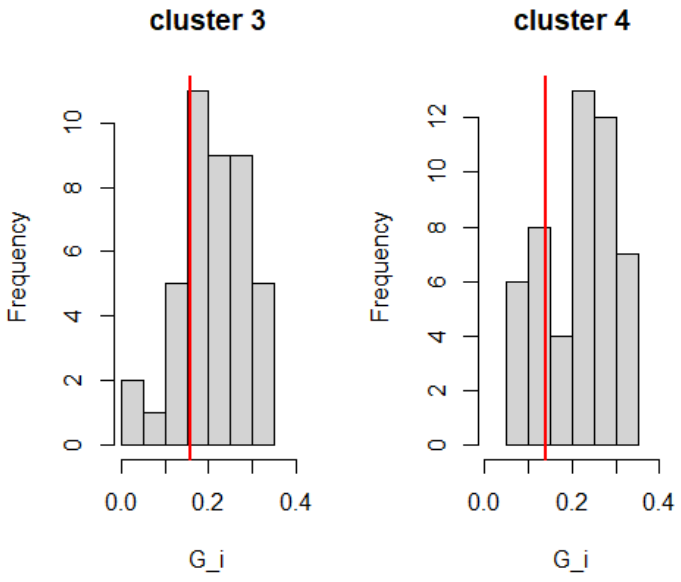

Figure 2: distributions of the G_i values of midterm exam 3 with a vertical line at the 25th percentile.

We repeated the analysis on the data from the final exam for STAT 201, considering the possibility that those who cheated together in midterm 3 exam may have collaborated on the final exam as well. We considered the two smallest clusters in each exam. Interestingly, percentage of students identified as suspects in both the Midterm and Final exams based on the same group of suspected students in the Midterm exam is 30.43%. After setting the threshold to the 25th percentile of G_i values, the suspected percentage reduced to

16.67%.

6 DISCUSSION

When confronted with only the results from multiple choice testing, it is impossible to determine with certainty that there are groups of students who have collaborated. The methods proposed here are based on a two-step procedure where both steps use probability theory to identify groups and to quantify the degree of suspicion regarding collaboration. Therefore, the methods developed here need to be supported with additional evidence to ascertain that cheating has taken place.

A potential benefit of the methods is that if groups are identified and deemed suspicious early in a course, an instructor may initiate conversations that could curtail future collaborative behaviours.

Of course, the methods presented here are subject to multiple choice tests subject to a particular testing protocol which itself discourages cheating. Future research may be considered under alternative testing formats. For example, with short answer questions, it may be possible to develop probabilistic methods involving the commonality of phrases. It may also be possible to develop similarity measures other than $s^{(2)}_{i_1 i_2}$ in (3) which are useful at identifying groups of students. A challenge in this line of work is the validation of methods. It would be necessary to procure real datasets where it is known that particular groups of students have cheated.

7 REFERENCES

Belleza, F.S. and Belleza, S.F. (1989). Detection of cheating on multiple-choice tests by using error-similarity analysis. *Teaching of Psychology*, 16, 151-155.

Coghlan, S., Miller, T. and Paterson, J. (2021). Good proctor or "big brother"? Ethics of online exam supervision technologies. *Philosophy and Technology*, To appear.

Driver, H.E. and Kroeber, A.L. (1932). Quantitative expression of cultural relationships. *University of California Publications in American Archaeology and Ethnology. Quantitative Expression of Cultural Relationships*, 211-256.

Nath, L. and Lovaglia, M. (2009). Cheating on multiplechoice exams: monitoring assessment, and an optional assignment. *College Teaching*, 57, 3-8.

Lancaster, T. and Cotarian, C. (2021). Contract cheating by STEM students through a file sharing website: a Covid-19 pandemic perspective. *International Journal for Educational Integrity*, 17, Article 3.

Li, C. and Lalani, F. (2020). The COVID-19 pandemic has changed education forever. This is how. *World Economic Forum*, Accessed online September 20, 2021 at www.weforum.org/agenda/2020/04/coronavirus-education-global-covid19-online-digital-learning/

Richmond, P. and Roehner, B.M. (2015). The detection of cheating in multiple choice examinations. *Physica A: Statistical Mechanics and its Applications*, 436, 418-429.

Sokal, R.R. and Michener, C.D. (1958). A statistical method for evaluating systematic relationships. *University of Kansas Science Bulletin*, 1409-1438.

Wesolowski, G.O. (2000). Detecting excessive similarity in answers on multiple choice exams. *Journal of Applied Statistics*, 27, 909-921.

Xu, D. and Tian, Y. (2015). A comprehensive survey of clustering algorithms. *Annals of Data Science* , 2, 165-193.