Journal of Applied Testing Technology, Vol 23(Special Issue), 54-61, 2022

Identifying Statistically Actionable Collusion in Remote Proctored Exams

Kirk Becker[1]* and Huijuan Meng[2]

[1]Kirk Becker, Pearson VUE, 4627 N Leclaire, Chicago, IL 60630, USA;
kirk.becker@pearson.com

[2]Senior Psychometrician, Amazon Web Services (AWS), 9963 Cyrandall Drive, Oakton, VA, 22124, USA;
huijuam@amazon.com

Abstract

The rise of online proctoring potentially provides more opportunities for item harvesting and consequent brain dumping and shared "study guides" based on stolen content. This has increased the need for rapid approaches for evaluating and acting on suspicious test responses in every delivery modality. Both hiring proxy test takers and studying unauthorized test content (e.g., "study guides" or brain dumps) result in characteristic patterns of responses, many of which are detectable through collusion analysis. The ability to identify and rapidly revoke test results are one component of stopping test takers from engaging in these behaviors, both in online proctored and test center testing. Existing collusion analyses have typically taken the approach of evaluating all response pairs sequentially, potentially requiring several days to evaluate a set of test results. This paper demonstrates matrix-based methods for quickly calculating exact overlap counts for large data sets, as well as approaches for determining criteria for flagging suspicious results or invalidating results. We discuss and compare the results for simulations and probability calculations and discuss the operational implications of these decisions.

Keywords: Test Security, Online Proctored Exams, Cheating Detection

1. Introduction

Like the move from paper-based to computer-based tests in the 1990s and 2000s, the use of online proctoring with licensure and certification programs raises questions about the comparability of results with test-center based results. In many discussions on this topic, the issue of test security is raised as a topic of particular concern. While proponents of online proctoring have pointed out that cheating does occur in test centers, access to the physical environment and computer hardware of the testing event provides new opportunities to cheat. Proxy test taking, as well as access to operational test questions through brain dumps or "study guides", are two of these opportunities that may be detected using forensic analysis.

Collusion analysis is the evaluation of excessive levels of response similarity to detect proxy test taking or shared responses. There are numerous methods for identifying answer copying and sharing, typically based on the overlap of incorrect responses, correct and incorrect responses, or all item responses (Zopluoglu, 2017). This general approach was originally introduced in the 1920s to identify answer copying in physical test locations (Bird, 1927; 1929), but has gained prominence as a means of detecting virtual collaboration as well. Groups of test takers who have all studied the same item set are identified by these analyses. Becker and Makransky (2011) demonstrated that response similarity was also relevant to proxy test takers, with repeat test takers showing extremely high similarity between responses

*Author for correspondence

over time. While an individual proxy testing for just one other person would be difficult to identify (e.g., having your twin brother take the test for you), a "professional" proxy test taker testing for numerous individuals would be easier to detect.

Evaluation of test results may be done for a targeted group (e.g., for an individual or test center), or for the overall testing population (Maynes, 2017), and the criteria for flagging test takers as well as the computational resources necessary will vary based on that decision. Evaluating the overall population for collusion requires the comparison of every test taker with every other test taker, resulting in $N * \frac{N\text{-}1}{2}$ calculations. When the calculations themselves are lengthy, the frequency with which these can be run may also be limited. Relying on traditional methods of collusion analysis would prevent the real-time (or near real-time) evaluation of test results, which has particular implications for the security of programs employing online proctoring. Frequent evaluation of test results for collusion, and the invalidation of results (or requirement to retest), would reduce current invalid test results, as well as reducing the likelihood of future cheating using these methods due to the known risk of getting caught. Motivated by this desire to perform frequent collusion analysis, the authors lay out in this paper an application of matrix multiplication which, in combination with pre-calculating flagging criteria, will dramatically speed up these analyses.

2. Response Similarity

Response similarity indices "look at the likelihood of agreement between two response vectors given the assumption of independent responding" (Zopluoglu, 2017). In practice this may involve counting correct responses in common, item scores in common, exact incorrect responses in common, or all responses in common. High scoring test takers will have many more correct responses in common with all test takers, and low scoring test takers will have more incorrect or exact incorrect responses in common with other low scoring test takers. For this reason, response similarity is insensitive to collusion between test takers with very high scores, and the measures of response similarity must condition on total scores.

The following counts are relevant to the calculation of response similarity indices:

- Correct overlap, the number of items two test takers have both answered correctly
- Incorrect overlap, the number of items two test takers have both answered incorrectly
- Exact incorrect overlap, the number of items for which two test takers have chosen the same incorrect response
- Items in common, the total number of items in common between two test takers
- The number of items that are correct for each test taker

There are numerous options for the interpretation and analysis of these values. Multiple-choice items will have greater similarity when a single distractor is much more likely than the others. Alternative item types (e.g., multi-select items, list ordering, etc.) will frequently have a larger number of possible responses. This is not an issue when looking at score similarity but should be considered when including exact incorrect matches in an analysis. Matching scores (0 or 1) are more likely than exact incorrect responses (A,B,C,D), so a higher level of score overlap is necessary to identify collusion from scores. Conversely, because exact incorrect responses are less likely than matching incorrect responses, their inclusion may increase the sensitivity of analyses. Because pretest items may vary between test takers, and because pretest items are generally less exposed than scored items, many operational collusion analyses include only scored items. This paper presents approaches for improving the efficiency of collusion analysis, determining criteria for flagging test taker pairs, as well as results of the application of these approaches.

3. Matrix Multiplication

Matrix multiplication provides a highly efficient method for calculating all the counts necessary for collusion analyses. Table 1 provides a sample of the minimum data necessary to produce several matrices used for collusion analysis. In this data format there is one row per individual per item, so a 50-item test would have 50 rows of data for each test taker. A Registration ID is a unique identifier for a test taker and a specific sitting of a test, while a Test Taker ID is a unique identifier for a test taker that is used

Table 1. Minimum data fields for collusion analysis matrices

Registration ID	Test Taker ID	Item ID	Item Response	Item Score	Item ID Response
R1	T1	49	A	1	49_A
R1	T1	50	B	0	50_B
R2	T2	1	A	1	1_A
R2	T2	2	D	0	2_D
R3	T2	51	C	1	51_C
R3	T2	52	C	1	52_C

across multiple tests. Because repeat test takers will show high levels of similarity when answering items they have previously seen, response similarity results should not flag high overlap between a test taker and their previous test across registrations (e.g., between R2 and R3 in the table, as these are both results from test taker "T2"). Highly dissimilar results for the same test taker could indicate a proxy test taker (Becker & Makransky, 2011). The Item ID is a unique identifier for a test item, while the Item Response is the option or response a test taker provided for the item. When working with alternative item types there can be a large number of different responses. Item Score is the point value given to the test taker response, although additional considerations not covered here are necessary for tests including polytomous items. Finally, the Item ID Response is a concatenation of the Item ID field and the Item Response field.

Data in Table 1 can be used to create the following four matrices:

1. Registration ID by Item ID, all data
2. Registration ID by Item ID, Item Score=1
3. Registration ID by Item ID, Item Score=0
4. Registration ID by Item ID Response, Item Score=0

Table 2 shows matrix 4 for a sample of data. This table is created by first selecting all rows where item score is 0. The table has one row for each unique registration ID, and one column for each unique item ID response. There are 3 columns for each item – one for each incorrect response. A test taker would have a 0 in all three columns if they answered correctly, and a 1 in one of the columns if they answered incorrectly. Test taker R1 for example answered items 1, 2, and 50 incorrectly, with responses of C, A, and A. Test taker R2 answered items 1 and 50 correctly, and item 2 incorrectly with a response of B.

Numerous statistical analysis platforms including SAS, R, and S-Plus include matrix multiplication. For demonstration purposes, we use syntax from R (R Core

Table 2. Registration ID by item ID response sample table

Registration ID	1_B	1_C	1_D	2_A	2_B	2_D	50_A	50_C	50_D
R1	0	1	0	1	0	0	1	0	0
R2	0	0	0	0	1	0	0	0	0
R100	0	0	0	0	1	0	0	0	1

Team, 2020) to demonstrate the creation of matrixes and matrix multiplication. A data file including the fields above can be used to create matrices (in R this would use the syntax “table (Registration ID, Item ID)”), which transforms the data into a table with one row per test taker, one column per item or item response, and 0/1 in each cell. Using matrix multiplication to multiply the table by its transpose (in R syntax, it is “matrix%*%t (matrix)”) will quickly produce a new matrix with the desired counts for calculating different response similarity indices.

Matrix 1 × Matrix 1^T will produce a table containing the number of items in common between each pair of test takers (CommonItems), with the test length for each test taker in the diagonal. Matrix 2 × Matrix 2^T will produce a table containing the number of correct items in common between each pair of test takers (CommonCorrect), with the diagonal (or row maximum) containing the score for the test taker in that row. Matrix 3 × Matrix 3^T will produce a table analogous to Matrix 2 but for incorrect items in common between each pair of test takers (CommonIncorrect). Finally, Matrix 4 × Matrix 4^T will produce the exact incorrect overlap between each pair of test takers (CommonExIncorrect). If desired, the table created from Matrix 3 minus the table created by Matrix 4 will provide the incorrect but non-identical response count (CommonNExIncorrect). Similarly, summing the tables created from Matrix 2 and Matrix 4 will provide the total responses in common between each pair of test takers (CommonResp). Note that the dimensions of the resulting tables will not be identical if there are test takers with zero correct or all correct items in the data. For subsequent analyses we will be using the sum of CommonCorrect and CommonResp, which is the total responses in common or Response Similarity Index (RSI).

Using an Intel i7 quad core CPU (2901 Mhz), multiplying a 10,000 by 100 matrix took 6.4 seconds in R, while a 40,000 by 300 matrix took 510 seconds in R. The larger matrix in this case produced 1.6 billion pairs (or 799 million unique pairs). This is the matrix size required to count exact incorrect overlap for 40,000 test takers on a 100-item test with 3 incorrect options per item. Anecdotally, performing these calculations through loops can take days to complete.

A Method for Interpreting Response Similarity

There are numerous methods for interpreting the significance of response overlap, including indices introduced by Angoff (1974), Holland (1996), Belov and Armstrong (2009), Maynes (2017), and others. It is beyond the scope of this paper to review and compare all of these, however, the approach we present for pre-calculating flagging values and simulating response distributions will be relevant to many different indices.

Early research on answer copying (Bird, 1927; 1929) made use of test data to estimate the null distribution of response overlap. This approach is limited in terms of the availability of large samples of operational test data and may be affected by collusion/proxy test taking in the data. An alternative to empirical distributions is the use of simulated test data. To establish statistically defensible cut-off values for an exam, a simulation process is presented that can be easily conducted by testing organizations to determine the desired flagging criteria for their exams.

The basic idea is to specify the number of comparison pairs (N) for each unique raw score pair to support the desired RSI flagging results. For a 50-item exam, in total, there are 1,225 unique raw score pairs (exclude 0 and 50), e.g., {1,1}, {1,2}, {1,3}, etc. Simulated data show for each score pair how frequently different levels of response overlap occur when there is no collusion. Operationally, the use of highly improbable levels of overlap (e.g., 1/10,000), combined with the requirement that flagged test takers overlap with multiple other test takers, will result in a conservative security analysis standard that identifies colluding groups.

In this example, item parameters are from a 50-item exam and calibrated under the Rasch model. All items are multiple-choice items with 4 options. This general process can be easily modified, but we expect flagging criteria to be robust to normal variation in item statistics.

1. Use exam item parameters to compute raw-to-theta score conversion table to determine theta scores for each raw score except for 0 and perfect score.
2. For each theta value, use IRT-b parameters to generate item scores (0/1) and compute the observed raw score.
3. If the observed raw score matches the true raw score, keep the simulated exam record, otherwise discard the data.
4. Repeat steps 2 and 3 until the total number of comparison pairs meets the specified sample size N.

5. For all cases with simulated correct answers, code the correct response as "1" by convention; for incorrectly answered items, randomly code the incorrect response as either "2" or "3" or "4".

This processed results in uniformly distributed raw/theta scores, with their associated simulated item score and item response vectors. These simulated data can then be used to calculate response similarity counts as detailed above. Probabilities for raw counts, or for indices computed from raw counts, can then be calculated from these data. For example, for raw score pair 30 and 31, if the number of comparison pairs is 10,000, and only one pair yields the highest total overlap value 34, the probability of observing this value would be 0.0001 (1/10,000). If an achievable overlap value (e.g. 35) is not seen across 10,000 pairs, the probability of observing this value would be less than 0.0001. Table 3 shows an example for score pair {31, 40}.

Table 3. Total exact overlap distribution example for score pair: 31, 40

Similarity Values	N	Proportion
21	4	0.0004
22	16	0.0016
23	74	0.0074
24	294	0.0294
25	701	0.0701
26	1228	0.1228
27	1699	0.1699
28	1823	0.1823
29	1681	0.1681
30	1155	0.1155
31	722	0.0722
32	359	0.0359
33	163	0.0163
34	57	0.0057
35	18	0.0018
36	5	0.0005
37	1	0.0001
38	0	0
39	0	0
40	0	0
41	0	0
Total	10,000	1

Relevant to the interpretation of this information is the maximum achievable overlap between responses given two scores. In the case of scores 31 and 40 there are 9 responses that must be different (40 minuses 31) because these are items one test taker answered correctly and the other answered incorrectly. The maximum achievable overlap is therefore 41. The following logic is applied when comparing the maximum achievable overlap to the maximum observed value:

1. If maximum achievable overlap is equal to the maximum observed value and the proportion of pairs with this overlap value is less than or equal to the pre-determined flag probability threshold (e.g., 0.001), then the raw score pair is eligible to flag, and the total overlap flag is the highest observed value.
2. Maximum achievable overlap and maximum observed value are the same but the percentage for the highest observed value is higher than pre-determined flag probability threshold (e.g., 0.001), this raw score pair is NOT eligible to flag.
3. The highest observed value is lower than the highest achievable value, this raw score pair is eligible to flag.

For score pair {31, 40} in Table 3, the highest observed value is 37 in the simulation. That is, among 10,000 simulated pairs of comparison, one yielded this value (p=0.0001), so a value of 37 is eligible to flag based on a probability threshold <=.0001. For the score pair {49, 49}, the highest achievable overlap is 50, which is the same as the highest observed value. However, 144 out of 10,000 pairs of comparison yield this value (p=0.014), so this score pair would not be flagged. We are still evaluating the best methods for setting conservative cuts when the highest observed values are higher than the probability threshold but lower than the highest achievable values, as well as the resilience of this method to reasonable variability in the distribution of item statistics.

4. Applying the Method to Simulated Collusion

Data with collusion were simulated using the approach described in Maynes (2017). An initial set of 4,000 theta values were calculated from live test data response strings, and these were used to generate 1 million pairs of theta values. Nine sets of 1 million response pairs were then generated, ranging from completely independent (0% copying) to 80% similar (40 out of 50 responses taken from the same base response string) responses. The raw scores and overlap for each response string were then calculated for each pair, and the pairs were then flagged based on various probability values ranging from .01 to .000001 using the M4 similarity index (Zopluoglu, 2019) and the simulated null distribution. The M4 similarity statistic calculates the probability of a given number of correct and incorrect matching responses based on a generating function and estimated probabilities of matching correct,

Table 4. Type 1 error for M4 and simulated null distribution

Probability Threshold	M4		Simulated Null Distribution	
	Number of Detections	Observed Rate	Number of Detections	Observed Rate
0.01	2715	0.002715	491	0.000491
0.001	180	0.00018	64	0.000064
0.0001	13	0.000013	5	0.000005
0.00001	2	0.000002	1	0.000001
0.000001	0	0	1	0.000001

Table 5. Power for M4 and simulated null distribution by copying rate

Copying %	M4					Simulated Null Distribution				
	0.01	0.001	0.0001	0.00001	0.000001	0.01	0.001	0.0001	0.00001	0.000001
10	0.014	0.001	0	0	0	0.008	0.002	0	0	0
20	0.056	0.009	0.001	0	0	0.054	0.018	0.005	0.001	0
30	0.159	0.036	0.007	0.001	0	0.206	0.096	0.034	0.014	0.005
40	0.352	0.114	0.031	0.008	0.001	0.481	0.297	0.146	0.075	0.036
50	0.601	0.277	0.1	0.035	0.007	0.738	0.576	0.375	0.247	0.147
60	0.831	0.538	0.266	0.114	0.034	0.9	0.807	0.65	0.513	0.381
70	0.951	0.801	0.547	0.309	0.127	0.969	0.928	0.852	0.753	0.664
80	0.989	0.947	0.83	0.627	0.379	0.992	0.978	0.957	0.912	0.872

matching incorrect, and unmatched incorrect responses. Table 4 shows the Type 1 error rates, which are the number of independent pairs flagged (out of 1 million pairs) for 0% copying. Table 5 shows the power across probability and percent of copied responses. It should be noted that not all response pairs are eligible for flagging because the scores are too high. Both M4 and simulation-based criteria have low false positive rates when an appropriate probability threshold is used, although power is limited with low levels of copying.

5. Discussion

The validity of test results depends on numerous factors, one of which is the integrity of the testing situation. Online proctoring offers increased access for testing, but also provides new opportunities for collusion and proxy test taking. Incorporating ongoing analysis of response overlap into the testing process can help to improve the integrity of existing test results, as well as discourage future misconduct. The efficiencies provided by matrix-based calculations can drastically reduce the time required to perform these analyses, facilitating their use in both operational and research settings. We have compared two approaches for identifying answer copying, however the approach to calculating response similarity counts can be used with numerous metrics. We encourage researchers to take advantage of the efficiency of calculating overlap to thoroughly compare existing and new collusion indices.

Both the M4 and null distribution simulation presented here make use of test item statistics. Future research will look at the consistency of flagging criteria when different distributions of item statistics are used and when incorrect options are more or less common. The simulation method described here should be largely independent of test difficulty and population distribution. If we find that flagging criteria are consistent across a wide range of item characteristics it will make it easier to implement those criteria widely. A single set of tables showing the overlap required to flag candidate pairs given test length (items in common) and score would allow for consistent and widespread application of collusion analysis.

Test security is rarely about identifying a single individual or pair of test takers, but rather to find and plug holes in test security more broadly. When groups of test takers have access to test content, results of collusion analysis show not just high overlapping pairs but highly overlapping groups. A single pair of results with 1 in-a-million overlap is suspicious, while a group of 73 is clearly actionable (Romo & Bowman, 2020). Whether online testing involves live remote proctors, review of recorded test sessions, or un-proctored testing using an honor code (as with the West Point cheating scandal), collusion analysis can detect and deter cheating. Proxy test takers may show different patterns of overlap than copiers and compiling and publishing results from verified incidents will help the testing community better understand what to look for. In situations where numerous test takers all have the same exceptionally high overlap organizations will be justified in invalidating test results. The inclusion of other forensic flags such as those covered in Cizek and Wollack (2017) and other data forensic work, as well as internal investigations and interviews, are also appropriate.

6. Conflict of Interest Disclosure

There were no external sponsors of this research.

7. References

Angoff, W. H. (1974). The development of statistical indices for detecting cheaters. *Journal of the American Statistical Association,* 69(345), 44-49. https://doi.org/10.1080/01621459.1974.10480126.

Becker, K. A. and Makransky, G. (February, 2011). Verifying Candidate Identity Over Time: Candidate Response Consistency for Repeated Test Items. Paper Presented at the Association of Test Publishers Annual Conference, Phoenix, AZ.

Belov, D. I. and Armstrong, R. D. (2009). Detection of answer copying via Kullback-Liebler divergence and K-Index. *Law School Admission Council Research Report,* 09-01.

Bird, C. (1927). The detection of cheating in objective examinations. *School and Society,* 25(635), 261-262.

Bird, C. (1929). An improved method of detecting cheating in objective examinations. *The Journal of Educational Research,* 19(5), 341-348. https://doi.org/10.1080/00220671.1929.10879954.

Cizek, G. J. and Wollack, J. A. (Eds.). (2017). Handbook of Quantitative Methods for Detecting Cheating on Tests. *Routledge*. https://doi.org/10.4324/9781315743097.

Holland, P. W. (1996). Assessing unusual agreement between the incorrect answers of two examinees using the K-Index:

Statistical theory and empirical support. ETS Program Statistics Research Technical Report No. 96-4. https://doi.org/10.1002/j.2333-8504.1996.tb01685.x

Maynes, D. (2017). Detecting Potential Collusion Among Individual Examinees Using Similarity Analysis. In G. J. Cizek & J. A. Wollack (Eds.) Handbook of Quantitative Methods for Detecting Cheating on Tests. https://doi.org/10.4324/9781315743097-3.

R Core Team (2020). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/.

Romo, V. and Bowman, T. (2020, December 28). More Than 70 West Point Cadets Accused of Cheating in Academic Scandal. NPR. https://www.npr.org/2020/12/21/949025580/more-than-70-west-point-cadets-accused-of-cheating-in-academic-scandal.

Zopluoglu, C. (2017). Similarity, Answer Copying, and Aberrance: Understanding the Status Quo. In: G. J. Cizek & J. A. Wollack (Eds.), Handbook of Quantitative Methods for Detecting Cheating on Tests. https://doi.org/10.4324/9781315743097-2.

Zopluoglu, C. (2019). Computation of the Response Similarity Index M4 in R under the Dichotomous and Nominal Item Response Models. *International Journal of Assessment Tools in Education, Promoting Free/Libre Software Use in Educational Measurement*, 1-19. https://doi.org/10.21449/ijate.527299.