

When is fast too fast? Setting score and time thresholds for credentialing programs

Casey Johnson, Ph.D.

Diane Talley, Ph.D.

Angelica Rankin, Ph.D.

Jill R. van den Heuvel, Ph.D.



Abstract

Are candidates responding too quickly to your credentialing exam items? When is being quick on an exam, too quick? When should the validity of the results be called into question? Candidate score by time analyses strive to address these questions. Fast response times at both the item and/or exam level can be indicative of item preknowledge. Credentialing programs frequently review candidate score by time for evaluating potential security concerns at the candidate, item, and exam levels. The analyses are appealing as they can be easily performed and interpreted, involve a visual element (score by time plots), and statistical probability (estimating the joint probability of each candidate's score and time combination). Additionally, policy decisions can be reflected with differing flagging thresholds (conservative versus liberal rules). The current exploration crosses multiple credentialing programs, multiple exams within each of those programs, and exams at different lifecycle stages representing different rates and likelihoods of item exposure. This paper explores analyses of candidate score by candidate time in an attempt to define: 1) what is reasonable, 2) what is extreme, 3) what actions are defensible, and 4) what policies should be in place.


Introduction

Imagine spending half a year or more and tens of thousands of dollars developing a certification exam and within just a few months, or weeks, or even days of the exam release, most (if not all) of the exam items are available on the internet. What are some indicators of such an exposure problem, other than candidates reaching out directly to report the availability of items? What should an exam sponsor do when it becomes obvious that exam scores are dramatically increasing while exam time spent is decreasing?

Attention to indications of item exposure and test fraud is essential to the integrity of a testing program and fairness to candidates (AERA, APA, & NCME, 2014). Compromised security of an exam poses a threat to the validity of exam scores, decreasing the accuracy of scores for their intended interpretation and use. Methods of detecting such compromise are needed for exam programs of all sizes and serving a multitude of purposes.

Most research in this area is conducted using large-scale education or licensure programs. Yet there are many small to mid-sized programs in certification that have the same concerns with regards to the validity of exam scores when exam content has been exposed and candidate scores are reflective of item preknowledge as opposed to job-related knowledge, skills, and abilities. These programs are often delivered using multiple fixed forms as opposed to computer adaptive administrations and offer continuous/on-demand exam administration as opposed to exam administration windows.

Fortunately, modern data forensics offers methods for detecting when candidates may have preknowledge of exam content. Early methods of data forensics focused on fit statistics, indicating when a candidate's response patterns significantly varied from expected behavior. These methods proved problematic for criterion-referenced exams given a lack of power due to the homogeneity of the population and because person abilities closely align with item difficulty by design. Newer methods emerged that identify candidates with preknowledge of and/or exposed exam content such as detection of answer copying or similarity of responses between exam candidates (Zopluoglu, 2016), Differential Person Functioning (DPF) comparing content believed to be exposed with unexposed content such as scored and unscored items (O'Leary & Smith, 2016), cluster analysis (Wollack & Maynes, 2016), and response time analyses using IRT and Bayesian models (van der



Linden & van Krimpen-Stoop, 2003; van der Linden & Guo, 2008; van der Linden, 2009; Qian, Staniewska, Reckase, & Woo, 2015; Boughton, Smith, & Ren, 2016).

Analyses using response time is typically used in one of two different ways. One analysis framework is to examine candidate response times for individual items. A second framework, proposed in this study, is to examine overall candidate response time in relation to total exam scores.

Methods focused on candidate response time for individual items assume that candidates with item preknowledge will respond to items more quickly or have response time patterns that are aberrant in comparison to predicted patterns. Analyses within this framework are valuable for detecting both candidates with item preknowledge and candidates engaging in item harvesting (memorizing items to share with future candidates) (Boughton et al, 2016). Response time methods overcome some of the prior weaknesses in data forensics by: 1) focusing on a continuous variable (time) suitable for statistical analyses, and 2) being robust to situations where item difficulty and person ability are closely aligned with stronger statistical power (van der Linden & Guo, 2008).

Response time is defined by the speed with which candidates respond to items as well as information regarding the effort required to respond to an item, or item intensity (van der Linden & Guo, 2008). Response time models are designed to detect various types of anomalous behavior, including item preknowledge. While response time models demonstrate value in identifying specific potentially compromised items, drawing candidate-level conclusions poses a challenge. If the goal is to identify candidates who potentially had preknowledge of exam content, then a defensible threshold specifying the number of anomalous items necessary to constitute a threat to validity would be necessary. Boughton et al (2016) used a percentile established for the purpose of the study (but not recommended as a defensible threshold), flagging candidates whose number of flagged items were at the 95th percentile in comparison with all other candidates' item-level flags. A method to determine a defensible threshold was not found in published literature.

A second methodology proposed uses candidate response time in comparison with exam scores to identify candidates with preknowledge of exam items. To date, there appears to be no published research investigating the use of total exam time with total raw scores to identify candidates who may have preknowledge of content. This method, Bivariate Score by Time (BST) analysis, determines the probability of each candidate's specific total exam time and total raw score given the estimated population total time and score means using a bivariate probability density function. BST, used in conjunction with other data forensics and programmatic information, shows promise in identifying systemic issues related to exam security. It also has the advantage of providing candidate-level evidence of preknowledge in a way that is straightforward and easy for stakeholders to comprehend and use (Foley, 2016). There is no published research to date that describes or suggests this method.

The current research seeks to address this gap by exploring a practical methodology for exam stakeholders managing programs of varying sizes to reveal potential instances of misconduct that would warrant further investigation. There can be valid and appropriate reasons why any given candidate may have a higher than expected score with a lower than expected time. However, information regarding exam score and time, combined with other statistical and non-statistical evidence, can lead to a convergence of evidence to use in making security-related decisions for a given program.

Methods

To conduct this exploratory study, data from three voluntary Information Technology (IT) certification programs was used. The data used was scrubbed of personally identifiable information at the candidate level, and was used with permission from each of the three included certification programs. The three programs varied widely in the number of exam titles within the certification program portfolio as well as in the number of candidates participating in the overall program as well as across the different exam titles. The first program has multiple certification exams, with varying pre-requisites but generally smaller overall volumes. The second program has exams representing multiple certifications with levels that candidates move through as they “progress” through the certification program such that the “lowest” level exam sees the greatest volumes and the “highest” level exams see the smallest volumes. The third program is a large program in which two exams from a single area were included in this study. The three programs were selected to generally represent small (Program 1), medium (Program 2), and large (Program 3) sized programs.

The certification programs involved additionally vary in how they secure their exams and certification programs. The programs included vary from open certification programs (anyone can decide to register for and take an exam) to closed certification programs (a candidate must apply to enter the certification program based on their work/role/experience in the field). There are also differences in the security measures instituted within the exam development process. One program requires multi-factor authentication by Subject Matter Experts in order to access the item bank for the exam being developed. Another program leads “closed door” development activities where “restricted access” laptops are provided and cell phones and other personal devices are not permitted within the development area. Finally, there are differences in the exam administration experience, though all require some level of proctoring (none are “open-administrations”). One program requires palm vein scans to access the test center, while another engages in remote online proctoring.

The data used ranged in date from October 2013 through January 2019. In total, across the three programs, there were 12 distinct certification exams studied with a total of 75,128 administrations.

Procedure

A bivariate normal probability density function was used to estimate the joint probability of each candidate's score and time combination for each exam. The parameters of the bivariate normal probability density (i.e., the mean and standard deviation of both exam time and exam score) were estimated based on exam data for 1) the first 10–30% administrations and 2) all administrations. The exact number of administrations was determined on a case-by-case basis, where noticeable deviations occurred in the moving averages of exam score and Rasch person measure. Thus, data from the first 10–30% of administrations allow us to estimate the score and time distribution from data assumed to be less influenced by the item exposure. In most datasets, the mean time decreased and the mean score increased over time when using the entire sample, which may be an indicator of item exposure.

For each candidate record¹, the probability measure indicates the probability that the candidate achieved their particular score and time combination, or a score and time combination more extreme, under the assumption that scores and times follow a bivariate normal distribution defined by the mean and standard deviation of score and time and with zero correlation between time and score.

Candidates are then flagged separately for 1) probability <0.001 and 2) probability <0.01. Low probabilities reflect unusual combinations of score and time in the direction of high scores and low times. Unusual combinations reflecting low scores and high times do not show low probabilities. Thus, each low probability calculation indicates anomalous behavior consistent with the hypothesis that candidates who have preknowledge of exam content will obtain high scores in a short amount of time. Four flagging areas were defined: High Score—Low Time (potentially indicative of item preknowledge); High Score—High Time (potentially slow, but appropriately able candidates); Low Score—Low Time (potential mis-administrations or item harvesting); Low Score—High Time (potential item harvesting).

Results

Using the procedure outlined above, joint probabilities were calculated for each candidate, within each exam, for each certification program. Results of these calculations are reported here at the program-level for each of the three Programs described above. Probabilities were calculated using the mean and standard deviations from 1) the first 10%–30% of administrations, and 2) the entire sample of administrations. In many instances, the subset of administrations had a lower raw score mean and higher mean time than the entire sample. Generally, this shows a trend of increased score and decreased time spent on exams and may be indicative of exposure. For example, Exam 2 for Program 1 has a mean time seven minutes lower and mean score almost one point higher when examining the mean of all scores compared with the first n administrations that are believed to be unexposed. Summary statistics for all exams included in the analysis are reported in Table 1.

¹ It is possible that a candidate appears in the data with more than a single record due to varying retake policies and the elimination of personally identifiable information from the data analysis.

Table 1. Raw Data Analyses

Program	Exam	Sample Size	Sample Used to Establish Baseline	Total Test Time (Mins)		Raw Scores	
				Mean	SD	Mean	SD
1	1	1,592	First 10-30%	151.74	30.62	43.40	11.89
			All	148.80	33.91	43.34	11.57
	2	645	First 10-30%	117.68	30.56	38.70	9.38
			All	110.75	36.84	39.87	8.53
2	1	7,898	First 10-30%	67.03	21.52	47.99	10.00
			All	66.97	21.84	48.47	9.64
	2	4,837	First 10-30%	83.17	16.80	52.12	6.51
			All	82.79	17.27	52.87	6.61
	3	2,854	First 10-30%	84.70	16.34	47.60	7.19
			All	84.88	15.99	47.56	7.14
	4	164	First 10-30%	134.13	28.60	81.98	13.89
			All	137.85	25.91	80.33	14.41
	5	577	First 10-30%	94.39	12.42	48.19	7.47
			All	94.28	12.75	48.34	7.43
	6	581	First 10-30%	86.77	16.82	48.63	6.41
			All	87.77	15.05	48.39	6.79
	7	1,350	First 10-30%	72.52	19.47	46.65	8.88
			All	74.68	19.15	46.18	8.57
	8	645	First 10-30%	97.08	20.79	41.97	5.13
			All	96.10	20.83	42.64	5.37
3	1	21,890	First 10-30%	69.00	15.97	50.98	11.07
			All	67.73	16.50	52.14	11.57
	2	16,206	First 10-30%	68.53	16.00	56.73	10.53
			All	70.82	15.66	53.31	11.10

Candidates were flagged for the four possible extreme quadrants (High Score—Low Time; High Score—High Time; Low Score—Low Time; Low Score—High Time). Each probability calculation was performed at two separate significance levels: 1) $p < 0.001$, and 2) $p < 0.01$. Figure 1 illustrates results from one program, flagging candidates with probabilities smaller than 0.001. The maximum time on this examination was 90 minutes except for candidates given accommodations, primarily due to English as a second language. Candidates flagged for High Score—Low Time are the upper left quadrant; candidates flagged for Low Score—Low Time are in the lower left quadrant; candidates flagged for High Score—High Time are the upper right quadrant (primarily candidates with accommodations); and candidates flagged for Low Score—High Time are in the lower right quadrant.

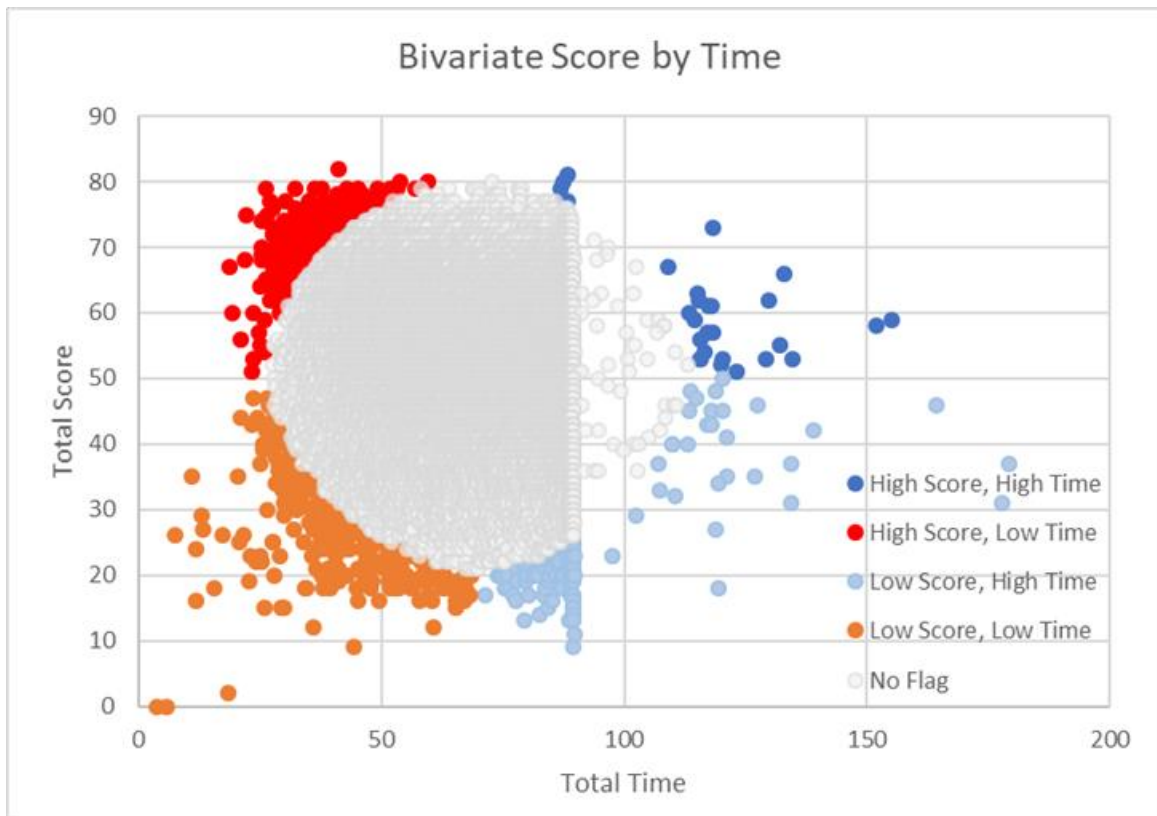


Figure 1. Bivariate Score by Time Scatterplot by Flags for One Analysis

Given that probabilities are calculated based upon a joint distribution of score and time, it is possible for an individual to receive multiple flags (e.g., administrations with extreme scores but moderate times could be flagged as both High Score—Low Time and High Score—High Time, due to the probabilities of the extreme score swaying the joint probability). Due to this artifact, administrations with multiple flags are reported in Table 2. Note that due to the differing overall sample sizes the flags are reported as a percentage as opposed to an n-count. The difference between a p-value of less than 0.001 and less than 0.01 results in an order magnitude increase in those flagged for improbable scores in time. Table 2 reports administrations flagged against means and standard deviations established using the subset of scores.

Table 2. Mean Percent of Sample Flagged for Different Probabilities.

Program	P-values	Average Percent Flagged					
		High Score Low Time	High Score High Time	Low Score Low Time	Low Score High Time	Flagged Pass	Flagged Fail
1	0.001	2.8%	0.3%	2.4%	0.6%	2.6%	0.2%
	0.01	11.3%	1.1%	11.1%	4.4%	9.6%	1.7%
2	0.001	0.7%	0.1%	1.2%	0.7%	0.7%	0.0%
	0.01	5.5%	2.3%	4.4%	4.4%	5.1%	0.4%
3	0.001	0.7%	0.2%	1.7%	1.6%	0.6%	0.1%
	0.01	5.9%	2.0%	7.6%	7.5%	5.7%	0.2%

To establish a baseline that reflects the expected mean score and time in the population a stable average for time and score must be determined. The intention is to identify a proper baseline before the effects of preknowledge impact scores and times. There are multiple ways in which stable means may be established for comparison. For programs with established security policies, a set number or percentage of candidates may be defined (e.g., n=1,000) as a baseline unless evidence of exposure exists prior to that threshold. A second method is to create a moving average graph and identify the point at which scores are noticeably increasing and/or times decreasing (Eckerly, 2016). Figure 2 illustrates the use of a visual inspection of the data. There are more sophisticated ways of determining significant changes (e.g., estimating slope), but these were not applied in the current study. Figure 2 demonstrates that mean score is relatively stable until approximately 100 administrations where a substantial increase is observed. After approximately 100 administrations it is seen that the moving average for score inflates and become inconsistent.

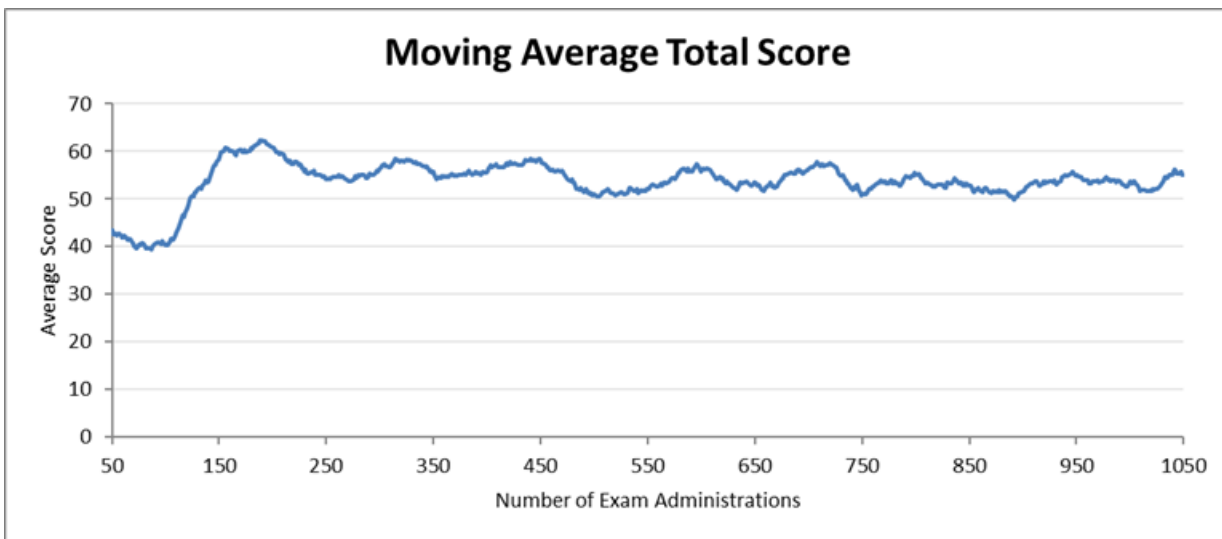


Figure 2. Example Moving Average of Score and Time for a Given Exam

Multiple sample sizes were examined for establishing a stable mean. Table 3 presents the percent of candidates flagged from a sub-sample of approximately 10%–30% of the full exam administration compared with all administrations in the sample. There is not one defined ‘sweet-spot’ that will work with every exam, as in the example above, there are multiple points where stable means can be determined. The analyst must examine the data and use judgement (or analyses) to determine where a stable mean occurs. Again, further research could be done to determine if there are thresholds possible for reviewing slopes within moving averages, but within this study visual inspection was used. The larger the sample, the “truer” the mean is likely to be, however the more prone the sample would be to item exposure issues.

Table 3. Mean Percent of Sample Flagged for Different Starting Number of Cases.

	Sample Size	Average Percent Flagged					
		High Score Low Time	High Score High Time	Low Score Low Time	Low Score High Time	Flagged Pass	Flagged Fail
1	10%-30%	11.3%	1.1%	11.1%	4.4%	9.6%	1.7%
	All	6.7%	1.3%	7.3%	6.2%	6.1%	0.6%
2	10-30%	5.9%	2.6%	4.6%	4.4%	5.5%	0.4%
	All	5.8%	2.3%	5.4%	4.9%	6.0%	0.7%
3	10%-30%	5.9%	2.0%	7.6%	7.5%	5.7%	0.2%
	All	6.1%	1.8%	6.2%	5.1%	5.8%	5.8%

Results in Table 3 show more candidates flagged for High Score Low Time when using a subset of data for Programs 1 and 2, but not in Program 3. Review of the data revealed that Program 3, Exam 1, followed the pattern of increased number of flags when using the subset, but was not the case for Exam 2.

Discussion

This exploratory study described a straightforward approach to identifying candidate preknowledge of exam content using total score and time information. Analysis included examination of two key decisions that need to be made when conducting a BST analysis to determine candidate preknowledge: 1) selecting a p-value to use in determining whether differences are significant, and 2) identifying a stable candidate sample to use as the baseline by which scores and times are compared.

What are reasonable parameters for a BST analysis?

Percent of candidates flagged at two significant levels, $p < 0.001$ and $p < 0.01$, were explored. The first would presume that only 0.1% of administrations should be flagged simply by chance, yet the data flagged 0.9%. The latter would presume 1% of administrations being flagged by chance, yet the data flagged 8%. From a purely statistical standpoint, these results are unlikely. However, this is not proof of fraudulent testing behavior or item exposure. Some portion of those flagged are potentially genuine results with valid reasons for the observed behavior. Further investigation should be taken for flagged candidate records to determine if the validity of the score and time combination.

On average, the difference between a $p < 0.001$ and $p < 0.01$ resulted in nearly ten-times as many administrations being flagged. This is expected, given the significance level increased by the same amount. However, it is important for exam sponsors to consider the consequences of flagging more or fewer candidates. If there were a thousand administrations, and 0.88% were flagged using a $p < 0.001$, that would mean there are nine instances that would require further investigation. If the significance of $p < 0.01$ was used, it might flag 6.5% of administrations; resulting in 65 instances requiring investigation. The level of effort to investigate a larger number of instances may not be feasible for most organizations. An organization may wish to consider how many individuals they are willing to investigate, which is often limited by available resources, and what impact that will have within the organization and the program. It is critically important to

treat all records equally/fairly. Thus, the selection of a flagging criteria should not be based on “who” is being flagged but policies defining why the flagging is occurring and the resulting actions for investigating flagged records. This study did not examine power or type 1 error rates, a recommended topic for future research, potentially using simulations to further investigate.

Two methods were described for determining for establishing stable means and standard deviations; identifying a specific sample size or percent of administrations and using moving averages. Selecting a specific number of administrations early in an administration period in order to establish baseline means that reflect valid candidate scores is typically a policy decision. In this exploration, sample sizes ranged from 5% to nearly 40% of a sample. Ultimately the results indicate that there are only minor differences between the starting means, as long as they accurately reflect a true representation of exam scores. There may be instances where only internal, highly proficient, individuals take the exam before the general public is allowed. This would obviously skew the statistics, and not be representative of expected scores. Alternatively, if the selection of first administrations is too large, such as the entire sample, then any effects from item exposure will be present in the means and standard deviations, and thus skew the analysis.


Various other methods could be used to determine the baseline score and time means and standard deviations. One approach would be to filter candidates from all results who are flagged in other data forensic analyses. Data from a previous administration of the same exam may be used, assuming there is no evidence of exposure during that administration. This may be supportable in instances where exposure is apparent early on in administration and a good baseline cannot be ascertained using the current data.

Ultimately, this step of the process will require a level of reasonable judgement from the organization and those who know the exams the best. In addition, program policy will inform choices such as p-values where a more conservative p-value may be selected for cases in which there are higher consequences (i.e., issues of legal defensibility for type 1 errors). There is no substitute for intimate knowledge of a program. If results are not as anticipated, then further investigation is warranted.

Defensibility Through Evidence and Policies

A common concern with regards to a data forensics analysis such as the BST is whether action based on the data is defensible. Following are some considerations in making that determination. First, if the intended action based on analysis is to identify systemic issues that should be addressed, then acting on data analysis such as BST is easily justifiable. Action in this case would be refreshing content of an exam more regularly and increasing frequency of exam form and item analysis as well as other proactive security methods. This might also include more closely monitoring individual test centers or geographic regions and limiting exam delivery where other methods of prevention fail.

When the goal is to pursue action against individual candidates, the answer is more complex and must include legal consultation. In this case, multiple sources of evidence are recommended to support a claim that a candidate had preknowledge of items on an exam. Triangulation of information from various sources that include both data forensic analyses demonstrating the probability that some form of cheating has occurred (e.g., preknowledge of content) and programmatic information (e.g., number of times a candidate has taken an exam or candidate history that reflects previous aberrant behavior) (Boughton et al, 2016) are needed to provide sufficient evidence in support of the intended actions specified in candidate policies. Even with such evidence, a candidate is never truly proven to have cheated (Cizek, 2016).



Semko and Hunt (2014) suggest that providing a clear policy reserving the right to take action in the event of suspected misconduct lessens the burden of proof, stating “courts, when considering whether to honor test sponsor’s score invalidation decision, will not ask whether the test sponsor ‘proved’ that cheating occurred. Rather, the courts typically treat the inquiry as a matter of contract law between the parties” (p. 238). They cite examples of test fraud on the SAT and ACT exams where the courts focused on evidence of due process. This is not to say evidence of fraud is not necessary, but to point out that thoughtfully (and with legal counsel) developed, communicated, and implemented exam security policies are essential in addition to monitoring the performance of exam items and forms and conducting appropriate data forensic analyses.

Industry testing standards (AERA, APA, & NCME, 2006; ISO, 2012; NCCA, 2014) and best practices (Ferrara, 2017) also specify the need for policies, clearly delineating candidate rights and expectations of ethical behavior in relation to exam content. Both ANSI 17024 and NCCA require specific security policies and procedures be in place for accredited exams and that candidates sign an agreement including information regarding the policies and procedures. This would include any possible actions that might be taken based on evidence of fraudulent behavior. In addition to enhancing fairness to candidates by ensuring policies are clear, publicizing policies and articulating these at the time of testing is thought to be a possible method of prevention, discouraging some candidates from unethical behavior (Ferrara, 2017; Foster, 2013).

Conclusion


The current study applied an unpublished method of using candidate score by time information to detect item exposure using actual exam data across programs of varying sizes. The goal this exploration was to begin a dialogue regarding the method for identifying candidates who have preknowledge of content going into an exam, compromising the validity of test scores and the reputation of the testing programs.

Results of this study are limited to its exploratory purpose, which did not include any controls for variations across programs or exams or investigation of power and type 1 error rates. Future research, both simulation and real data studies, may inform the utility, strengths, and weaknesses of the BST model. With regards to determining a baseline for comparison, thus far, these have been determined using a subset of the candidates administered each exam. However, use of a secondary set of data, similar to what Qian et al (2016) used in their study of response time methods, may also be appropriate. In cases where a test form is exposed very early in the administration, a separate data set (if available) may be needed to establish a reasonable baseline for comparison.

Finally, comparisons with other data forensic analysis techniques such as Differential Person Functioning, Answer Similarity Index, or Response Time analyses would be useful. Comparisons could also extend to non-statistical analyses, utilizing programmatic information to bolster results of BST analyses.

References

- American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (2014). Standards for educational and psychological testing. Washington, DC: American Educational Research Association.
- Boughton, K. A., Smith, J., & Ren, H. (2017). Using response time data to detect compromised items and/or people. In G. J. Cizek & J. A. Wollack (Eds.), Handbook of quantitative methods for detecting cheating on tests (pp. 177–192). New York, NY: Routledge.
- Eckerly, C. A. (2017). Detecting preknowledge and item compromise: Understanding the status quo. In G. J. Cizek & J. A. Wollack (Eds.), Handbook of quantitative methods for detecting cheating on tests (pp. 101–123). New York, NY: Routledge.
- Ferrara, S. (2017). A framework for Policies and Practices to improve test security programs: Prevention, detection, Investigation, and Resolution (PDIR). Educational Measurement: Issues and Practice, 36(3), pp. 5-23.
- Foley, B. P. (2017). Visual displays of test fraud data. In G. J. Cizek & J. A. Wollack (Eds.), Handbook of quantitative methods for detecting cheating on tests (pp. 151–163). New York, NY: Routledge.
- Foster, D. (2013). Security issues in technology-based testing. In J. A. Wollack & J. J. Fremer (Eds.), Handbook of test security (pp. 39-83). New York, NY: Routledge.
- International Standard Organization Committee on Conformity Assessment. (2012). Conformity assessment: General requirements for bodies operating certification of persons ISO/IEC 17024. Geneva, Switzerland: International Standards Organization.
- National Commission for Certifying Agencies. (2004). Standards for the accreditation of certification programs. Washington, DC: Institute for Credentialing Excellence.
- O’Leary, L. S., & Smith, R. W. (2017). Detecting candidate preknowledge and compromised content using differential item functioning. In G. J. Cizek & J. A. Wollack (Eds.), Handbook of quantitative methods for detecting cheating on tests (pp. 323-345). New York, NY: Routledge.
- Qian, H., Staniewska, D., Reckase, M., & Woo, A. (2016). Using response time to detect item preknowledge in computer-based licensure examinations. Educational Measurement: Issues and Practice, 1-10.
- Semko, J.A., & Hunt, R. (2013). Legal matters in test security. In J. A. Wollack & J. J. Fremer (Eds), Handbook of Test Security (pp. 237-255). New York: Routledge.
- van der Linden W. J., & van Krimpen-Stoop, E. M. L. A. (2003). Using response times to detect aberrant response patterns in computerized adaptive testing. Psychometrika, 68, 251–265.
- van der Linden, W. J., & Guo, F. (2008). Bayesian procedures for identifying aberrant response-time patterns in adaptive testing. *Psychometrika*, 73(3), 365–384. <https://doi.org/10.1007/s11336-007-9046-8>
- van der Linden, W. J. (2009). A bivariate lognormal response-time model for the detection of collusion between test takers. Journal of Educational and Behavioral Statistics, 34, 378–394.

- 
- Wollack, J. A., & Maynes, D. D. (2017). Detection of test collusion using cluster analysis. In G. J. Cizek & J. A. Wollack (Eds.), *Handbook of quantitative methods for detecting cheating on tests* (pp. 124–150). New York, NY: Routledge.
- 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* (pp. 25–46). New York, NY: Routledge.