Applied Psychological Measurement
2024, Vol. 0(0) 1–13
© The Author(s) 2024
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/01466216241248826
journals.sagepub.com/home/apm
Sage

Test Security and the Pandemic: Comparison of Test Center and Online Proctor Delivery Modalities

Kirk A. Becker[1], Jinghua Liu, and Paul E. Jones[1]

Abstract

Published information is limited regarding the security of testing programs, and even less on the relative security of different testing modalities: in-person at test centers (TC) versus remote online proctored (OP) testing. This article begins by examining indicators of test security violations across a wide range of programs in professional, admissions, and IT fields. We look at high levels of response overlap as a potential indicator of collusion to cheat on the exam and compare rates by modality and between test center types. Next, we scrutinize indicators of potential test security violations for a single large testing program over the course of 14 months, during which the program went from exclusively in-person TC testing to a mix of OP and TC testing. Test security indicators include high response overlap, large numbers of fast correct responses, large numbers of slow correct responses, large test-retest score gains, unusually fast response times for passing candidates, and measures of differential person functioning. These indicators are examined and compared prior to and after the introduction of OP testing. In addition, test-retest modality is examined for candidates who fail and retest subsequent to the introduction of OP testing, with special attention paid to test takers who change modality between the initial attempt and the retest. These data allow us to understand whether indications of content exposure increase with the introduction of OP testing, and whether testing modalities affect potential score increase in a similar way.

Keywords

test security, online proctoring, remote proctoring, risks to validity, high-stakes assessment

Introduction

The COVID-19 pandemic has disrupted testing programs and changed the outlook of the testing industry. One of the most notable changes is the rapid increase in remote online proctored testing.

[1]Pearson VUE, Chicago IL, USA

Corresponding Author:
Kirk A. Becker, Psychometric Services, Pearson VUE, 4627 N Leclaire Ave, Chicago, IL 60630, USA.
Emails: Kirk.becker@pearson.com; Kirk.becker@yahoo.com

Online proctored (OP) testing was developed to make test taking more available and convenient, while increasing the security of heretofore proctor-less online tests in employment and educational settings. By contrast, test center (TC) delivery involves dedicated locations where test takers provide physical ID and verify their identity in person and use computers provided by the test center in a controlled environment while being surveilled by an in-person proctor. TC environments may show variation based on the policies and technology requirements of the test delivery provider and variability in proctor quality across locations. OP environments, on the other hand, add more varieties, such as differences in the individual test taker's environment, computer hardware, and connectivity.

While not proven, it is easy to imagine a possible interaction between testing in public and private environments and one's inclination to try to cheat, one's attitude or motivation toward the test session, and other personal factors that could affect performance for good or bad. Given the limited published empirical research comparing the security of the two delivery methods, testing organizations must fall back on professional judgment to decide whether OP test delivery "eliminate[s] opportunities for test takers to attain scores by fraudulent means" (AERA et al., 2014, p. 116) to the same extent as TC delivery. It is reasonable to argue that a testing vendor or organization can eliminate more opportunities to cheat if they control the physical environment. Reviews of OP vulnerability to cheating behaviors (Woodley, personal communication, June 19, 2020) have sounded the need for caution here. This worry about greater vulnerability in the OP environment may explain why, until the pandemic, relatively few high-stakes programs had considered OP delivery given the technology and resources available at the time.

Of course, the goal of completely "controlling the testing room" is aspirational. Room control means that the proctor controls what and who is in the testing room, and limits the materials (cell phones, cheat sheets, etc.) that can be taken into or out of the room. The first practical objective for any proctor is to make it very difficult for a proxy to take the exam: either take the exam for the nominal test taker or communicate the answers to the nominal test taker in real time. While cheating doubtless occurs in testing centers where a proctor is surveilling several candidates in real time, a remote proctor surveilling several candidates with less room control and a narrower field of vision would likely be less successful at inhibiting or detecting cheating, even with the help of artificial intelligence.

Like individual cheating, exam theft through covert information gathering is logically more likely in a physical environment controlled by the test taker than in a testing center. A thief could easily hide many kinds of surveillance equipment in the remote test site and quickly steal all the items on an exam. For example, Foster and Marder (2020, April) showed that it is possible for hidden cameras to "outrange" a single computer web camera and, through Wi-Fi or Bluetooth connections to external computers, automatically process item text off the computer screen without being in the field of view of a proctor confined to a one-camera view of the room. Of course, every test presents the risk of item compromise through harvesting. Cell phones, miniature cameras (even contact lens cameras), and other spy technology also present a risk for recording content in the testing center. Lower-tech efforts to harvest through candidate memory will always exist. Whatever the method, the effects of item theft (typically on hard-to-contain social media) are even more serious than individual cheating, potentially undermining the interpretation of many test scores and rendering worthless item banks that may have cost hundreds of thousands to millions of dollars to produce.

The rapid rollout of online proctoring with the onset of the pandemic created an inflection point in the attitudes of test sponsors towards testing at home. It also created a sudden change to the security environments of many exams. So far a very limited amount of data is available regarding the relative security of OP versus TC environments, and the purpose of this study is to begin to provide that quantitative data. Hurtz and Weiner (2022) compared OP and TC test takers from five different credentialing exams on exam performance and forensic indicators, including occurrence of high similarity responses, irregular response patterns, and response speed. While some small differences in the proportion similar responses by OP versus TC candidates and slightly higher percentage of irregular cases among OP candidates, two having slightly higher percentages among TC candidates, almost all cases the effect sizes in these comparisons were negligible. We hope the results we are reporting will add to the conversation about the appropriate use of OP.

Overview of Types of Test Security Analysis

Test security analyses involve the detection of behaviors or test results that are unusual in the general population and potentially indicative of dishonest behaviors or compromised content. Any such behavior that departs from legitimate test taking obviously threatens the validity of the test result interpretation and the downstream statistics used to construct subsequent exams. Examples of dishonest behaviors include illicitly accessing live test content for study, access to unauthorized information (e.g., cheat sheets) during testing, working with a proxy test taker, working with a dishonest testing location, and taking tests specifically to acquire test content.

Within this umbrella of illicit behaviors, “test collusion” refers to cheating in cooperation with another person, such as by answer copying, sharing test content, communicating during an exam, proxy test taking, and receiving content from instructors (Maynes, 2017). Forensic analyses can be conducted with the aim of identifying individuals, locations, or groups involved in collusion, as well as items and tests that have been compromised. These analyses are important aspects of test validity arguments in both TC and OP environments. Beginning at least as far back as the 1920s (Bird, 1927), detecting cheating on multiple-choice tests has involved finding unusual levels of agreement between test taker responses (response overlap), typically incorrect responses. When test takers are all taking the same test in the same location, copying answers is a relatively easy way to cheat on the test, either by stealing answers or collusion. The introduction of computer-based testing and randomized item order, the use of multiple different test forms, and the use of continuous rather than event-based testing has virtually removed the opportunity for answer copying in person. However, other methods of collusion such as preknowledge through access to illicit study guides and using proxy test taking, are susceptible to detection using response overlap.

The probabilistic nature of the testing situation means that even when two test takers have the same level of knowledge, they won’t necessarily know the answers to all the same items, although people who have studied in the same class might have more similar response patterns than those from very different teaching environments. As scores increase to near 100% correct, the sensitivity of response overlap decreases as pairs of high-scoring test takers will have all or nearly all item responses in common. Tests covering a wide range of content, or adaptive tests, may show different patterns of agreement than fixed form tests or those covering a narrow range of content. Multiple-choice items with three plausible distractors will have less agreement than those with only one plausible distractor. So as Saupe quipped in 1960, while perfect agreement of all correct and incorrect responses seems an obvious case of cheating, “there would exist a need for estimating the likelihood of the observed correspondence under the condition of no collaboration.” (Saupe, 1960, p.475). Numerous approaches have been used for determining significant response similarity for the purpose of identifying collusion (Becker & Meng, 2022; Maynes, 2012; Smith, 2019; Zopluoglu, 2017).

Other approaches to identifying cheating or security breaches include response time analyses, unusual response patterns, large retake score changes, item parameter drift, discovery of stolen content, pretest versus operational item performance, and a variety of other approaches (see Cizek & Wollack, 2017; Kingston & Clark, 2014; Wollack & Fremer, 2013). These analyses may supplement collusion analyses, identify situations outside of collusion, or apply to situations such as CAT or other pool-based testing that limit or preclude response overlap analysis.

This article reports on the results of two studies of potential test taker fraud. In both studies, we used collusion analysis to compare security between OP and TC settings. Study 1 examines high degrees of answer overlap across several hundred testing programs in 2022, while Study 2 examines high overlap and other security indicators in a detailed analysis of a single testing program prior to and following the introduction of OP in 2020.

Study 1

Data

The first dataset included a convenience sample of aggregated data from over 3 million test takers across 326 testing programs from January to July of 2022. These data were compiled by an automated process and were not subject to individual data cleaning procedures prior to aggregation. These programs included professional credentialing, admissions, and IT certification programs. Programs offering CAT, LOFT, or other pool-based exam administration, and programs with short tests (<50 items) were excluded.

These data include a breakdown by testing location type and week but are otherwise aggregated by testing organization. Test location types included online proctoring (OP) and three types of TCs: (a) test centers owned and run by the testing vendor (company-owned), (b) test centers not owned by the testing vendor, but which met hardware and security standards (3rd party select), and (c) third-party test centers which had not indicated that they meet those standards (3rd party). Company-owned and 3rd party select test centers in this sample require direct line of sight for proctors, walk-throughs, and overhead cameras, while third-party TCs require at least one of those. For proctor to candidate ratios, the company-owned test centers required a maximum ratio of 15:1, while the other TCs are not specified. The OP exams made use of live real-time proctors with a single camera and up to 15 test takers assigned to each proctor using two computer monitors. Some testing programs included in these data tested in all test location types, while others tested in only a subset. There are large differences in sample size, as well as differences between the types of organizations making use of OP versus different types of testing centers. A subset of 33 organizations was selected, containing results from programs with $N > 100$ in all four test location types.

Methods

For this dataset, the total number of correct and incorrect responses in common was calculated for all test takers each week, compared with all other test takers from the current week and the previous time period (time period was typically 2 months, but could be shorter depending on exam volume). Because of the large number of programs and volume of test takers in this initial study, only extreme levels of overlap were flagged. Test taker pairs where each candidate's number-correct score was <90% correct and where response overlap was >95%, were flagged. Items that were not answered were not counted as matching, and high overlap for retake test takers with previous attempts were not flagged. These collusion criteria could be described as very stringent, even given the large number of comparisons being made, and are more likely to detect proxy test takers than preknowledge. The number of unique flagged test takers were then counted within each test location type by week. These counts, combined with the total number of test takers, were used to evaluate likely collusion by testing location. For the subsample including only testing programs with all four testing modalities, the percent of flagged test takers were calculated for each organization, and averages and standard deviations were computed across the 35 organizations for each test location type.

Results

Table 1 shows the total number of test takers and the percentage of test takers with a response overlap >95% with at least one other person and a percent correct <90%. The company-owned test centers had the lowest rate of collusion, with 3rd party select test centers showing a slightly increased rate, and 3rd party test centers having the highest incidence rate among test centers. OP test takers were flagged at a higher rate than any of the three types of test centers. These values are obviously influenced by the size and characteristics of the programs included, with large and highly valued programs potentially having an outsized impact. Many programs had no test takers with 95% overlap in a given test location type.

Figure 1 shows percentage of flagged candidates by test modality and type by week over the course of 25 weeks. While these weekly flagging rates show a great deal of variability, especially for the OP modality, the relative incidence of overlap by mode and test center type is consistent: OP shows the most overlap, followed by non-select 3rd Party test centers. The Company-Owned and 3rd Party Select test centers have the lowest incidence of high overlap.

Table 2 provides the mean and standard deviation of flagging rates for the 33 testing programs using all four test location types. The percent of candidates flagged (>95% response overlap and percent correct <90%) was calculated by organization and by location type, and these percentages were then used to calculate the mean percent flagged (Mean High Overlap %) and variance of percent flagged (SD %) reported in table 2. This breakdown partials out the effect of sample size and differences in the organizational makeup of the data provided in Table 1. While this sample includes only 33 organizations, over 1.9 million test taker results are represented. These results show a similar pattern to that in Table 1, with company-owned TC showing the lowest percentage of flag, whereas 3rd party TC having the highest percentage of flag. The percentage of OP flag is higher than any of the TC flag. The variance of flags is also higher in third-party and OP settings.

Study 2

Data

The second dataset included data from 69,353 testing instances from November 2019 through January 2021 for a single test program. The total data set includes incomplete tests and repeat test takers, although different analyses may include or exclude certain results. Between November 2019 and March 2020, six non-overlapping test forms were administered in company-owned and third-party select test centers. Test takers retesting after failing were administered a different form

Table 1. Percent of Flagged Candidates by Mode and by Test Location Type in 2022.

Modality	*N*	High overlap (%)[a]
TC: Company owned	>800k	0.5
TC: 3rd party – select	>200k	0.6
TC: 3rd party	>1.6 M	1.0
OP	>600k	3.5

[a] Percent of test takers with a 95% response overlap with at least one other test taker, and number-correct score <90% correct.

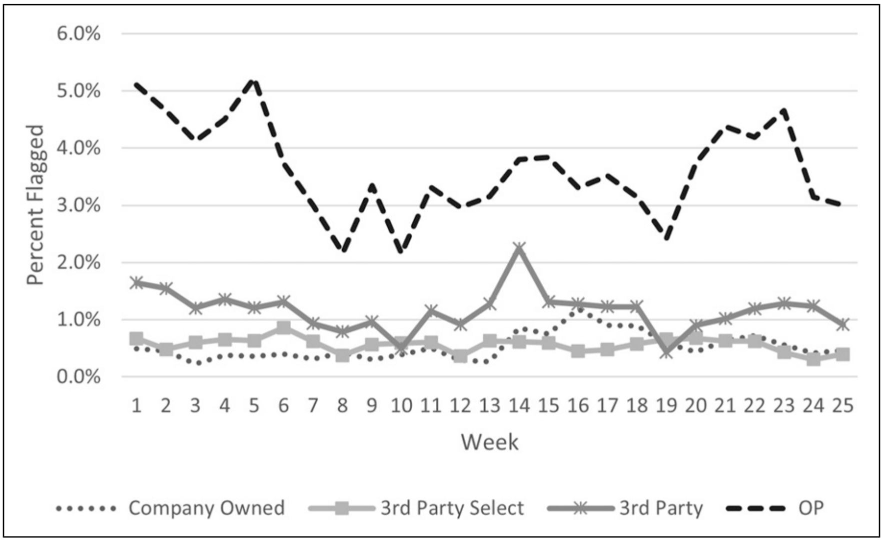

Figure 1. Percent of flagged candidates by mode and by test location across week of 2022.

Table 2. Flagged Candidates for 33 Programs Tested in all Location Types.

Location type	Candidates included	Mean high overlap %	SD%
TC: Company owned	>100k	0.7	0.02
TC: 3rd party – select	>1m	0.8	0.02
TC: 3rd party	>50k	1.4	0.04
OP	>600k	1.8	0.04

than previously administered when possible, and always a different form than the last one taken. This program began offering OP (live real-time proctors with a single camera and multiple test takers assigned to each proctor) as an option in April of 2020 during the pandemic shut down. For the first four months of OP administration, two of the six test forms that were previously administered in test centers were available in both OP and TC administration. Due to the shutdown only, OP exams were administered in April. During this period, an illicit study guide containing operational test items was discovered online. New test forms were then published in August of 2020, along with additional security protocols for OP.

Methods

While Study 1 looked only at test taker pairs with >95% response overlap and score <90% correct, Study 2 made use of a simulated null distribution (Becker & Jones, 2022; Becker & Meng, 2022) to determine flagging criteria. The use of a null distribution will avoid interpreting high levels of overlap as statistically normal if widespread cheating is present and empirical distributions are used to interpret overlap. The distribution of overlapping responses is determined by the number of items in common, the distribution of item parameters on the test, the probability of selecting each incorrect response on items, and the relative scores of a pair of test takers.

For Study 2, flagging values for significant response overlap were determined by simulating large numbers of responses with different scores for several test lengths and identifying overlap with $p < .00001$. While Study 2 involves a single testing program, we used a general method for flagging significant overlap that did not depend on the specific characteristics of this test. This is in line with our goal to have a general flagging rule that can be calculated ahead of time for the identification of potential misconduct. To calculate the significant overlap percentages, random normal distributions ($M = 0$, $SD = 1$) of IRT item difficulties (b values) were generated for test lengths of 25, 50, and 75. Three incorrect options were simulated with probabilities of 70%, 20%, and 10%, coinciding with the observation that real distractor options tend to be unequal in strength. Overlap criteria for these discrete test lengths were applied to pairs of response vectors of similar length, accommodating the fact that test takers did not necessarily take all items on the test, and it would be difficult to simulate every possible number of overlapping items. Thus, the flagging criteria, in terms of percent of overlapping responses, for 25-item exams were used to evaluate the overlap between response vectors containing 25 to 49 items, the flagging criteria for 50-item-exams were applied to vectors of between 50 and 74 items, and criteria for 75-item exams were applied to vectors of 75 items. We refer to the percentage of cases flagged for high overlap based on the null distribution as the *Response Similarity Index* (RSI).

Like the programs in Study 1, the detailed analysis of collusion for the single program involved calculating the overlap of correct and incorrect responses for all pairs of test takers. Instead of weekly analysis, this was done using monthly data, comparing test takers who tested within a given month with all other test takers in that month, plus those who tested during the previous month. For each test taker pair, the percent correct for each test taker was also calculated. The observed response overlap (both correct and incorrect) was then compared to the expected overlap for similar test taker pairs based on the simulated distribution of response overlap in the null condition. For example, on a 75-item test, when one test taker has a score of 80% and another test taker has a score of 90%, a total response overlap greater than 89.3% (67 responses in common) would appear with probability <.00001. A test taker was flagged if they had response overlap with at least one other test taker, with $p < .00001$, and that other test taker did not share the same person ID. Response overlaps of 100% for flagged pairs (all correct and all incorrect responses identical) were specifically called out.

Test security flags were also calculated based on test and item time anomalies, as well as indicators of unusual responses. While these flags are not conclusive evidence of security violations, they are suggestive, especially when they are used collectively. These flags include:

- Slow correct responses: Number of correct responses in slowest 5%. Flagged if number is < 3 SD below mean.
- Fast correct responses: Number of correct response with time <15 seconds. Flagged if number is > 3 SDs above mean fast correct count.
- Fast Pass: Candidates passing the test in fewer than 15 minutes.
- Negative Response Time Correlation: Correlation between response time and mean response time <0.
- Low Easy versus Hard Performance Delta: Performance difference between percent correct on easy and hard items (items above or below the median item p-value). Flagged if difference <3 SD below the mean difference.
- High Operational versus Pretest Performance Delta: Performance difference between percent correct on pretest and operational items. Flagged if difference >3 SD above mean pretest/operational percent correct difference.

In addition to the response overlap and security flags mentioned, these data also included interesting within-candidate evidence bearing on exam security. Within the timeframe studied, there were 7,259 test takers who failed their first attempt and then retook the exam a second time. The results for these test takers allow us to compare the security environments for OP versus TC test delivery using differential number-correct score gains and pass rates between the first and second attempt. In analyzing these cases, we distinguished between exams that were "fully delivered" or "not fully delivered." A test that was "fully delivered" meant that either the test taker used all available time or answered all questions on the exam. A test that was not "fully delivered" in OP could have been stopped by the proctor, stopped due to technical issues, or stopped by the test taker, although that information is not available to the authors.

Results

Figure 2 shows the RSI by month in both OP and TC environments. The average monthly volume for the period was around 3,000 test takers. We divided the time frame into three periods: Pre-OP, OP begins, and new forms published. In the five months prior to the introduction of OP, four candidates out of 17,000 tested were flagged for significant overlap. This flagging rate for over 17,000 test takers (~200 billion pairs compared) offers strong evidence that the flagging criteria are not detecting chance levels of overlap. With the introduction of OP, we see not only a higher number of flagged candidates in that modality, but also the appearance of flagged candidates in TCs. The general pattern of results was the same in both modalities, albeit to different degrees: the flagging rate started to climb with the introduction of OP, reached its peak sometime in the summer, and then declined when new forms and security protocols were introduced in August. After August, the OP channel still shows greater numbers of flags than TC, and TC shows small but non-zero numbers of flags, when prior to OP introduction there were effectively no flags.

Table 3 shows the monthly test taker counts and RSIs. The OP test takers are further broken down to show the percentage of cases equal to 100% response overlap and the percentage less than 100% overlap. TC flags are not broken down this way because a total of only 16 TC test takers had 100% overlap with at least one other test taker (where the other test taker was either TC or OP).

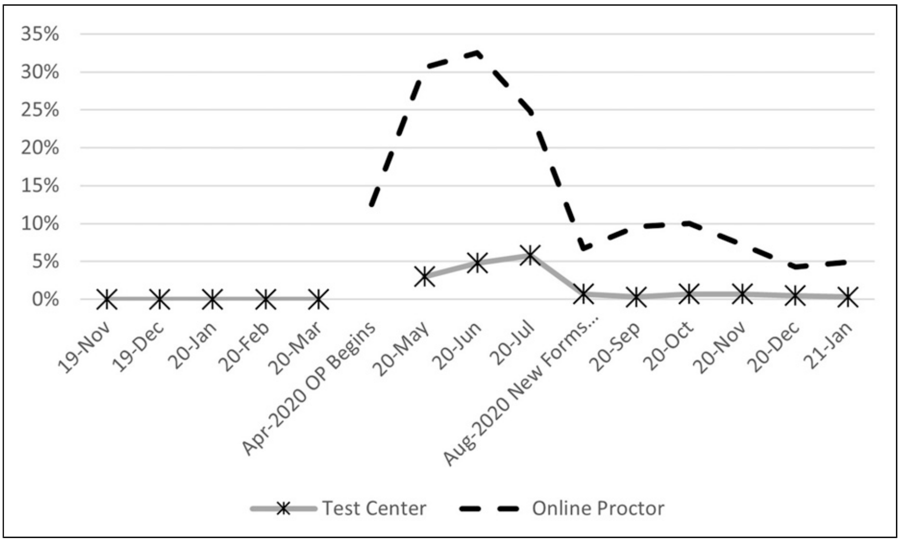

Figure 2. Monthly RSI during 3 time periods.

Overlap of 100% is likely a result of proxy test taking (test takers are remarkably consistent in their item responses over time) which, while not impossible in a test center, appears to be very difficult in that modality. Approximately 47% of the flagged cases in OP had 100% response overlap with at least one other test taker.

Table 4 reports the additional security flags for OP versus TC during the three time periods, as well as the test pass rates. There is a greater incidence of fast correct responses and low response time correlations among OP candidates in the initial OP and later OP time periods. Pass rates do show differences, however, there is no control over the comparability of the populations in OP versus TC. The 15 second correct and response time correlation flags also show larger differences between test takers flagged for possible collusion (i.e., significant overlap) versus those not flagged, which is shown in Table 5 along with the pass rates for these groups. It is important to note that while the data does include repeat test takers, RSI flags exclude comparisons between a retest test taker and their own previous results.

Next, we look at subsets of repeat test takers, focusing on differences between Time 1 and Time 2. These test takers failed the Time 1 attempt and then either passed or failed on the Time 2 attempt. The data we have reviewed so far suggests that collusion is more frequent for this exam in the OP mode, and perfect overlap between responses is exceedingly more frequent in OP. Thus, we might expect TC- > OP Time 2 scores and pass rates to be elevated for test takers with a propensity to cheat, but who couldn't do so under TC restrictions. We might expect less elevation among OP- > OP test takers who failed Time 1 either because they had less propensity to cheat or less ability. These persons might still be elevated more than OP- > TC or TC- > TC test takers who had more restrictions during Time 2. Obviously, we would expect greater score gains among test takers who were flagged for collusion than among those having no evidence of collusion.

Table 6 provides the Time 1 and Time 2 modality based on whether Time 1 was complete (excluding those for whom Time 2 was incomplete), as well as the score change and pass rates for

Table 3. Monthly Response Similarity and 100% Overlap by Testing Modality.

	TC		OP		OP RSI breakdown	
Month	*N*	RSI (%)	*N*	RSI (%)	<100 (%) Overlap	100 (%) Overlap
Pre-OP						
Nov-19	3798	0.0				
Dec-19	3389	0.0				
Jan-20	3481	0.0				
Feb-20	4130	0.0				
Mar-20	2382	0.0				
OP begins						
Apr-2020	7	0.0	1327	12.5	8.7	3.8
May-20	1167	3.0	5141	30.6	15.0	15.6
Jun-20	1747	4.8	5203	32.5	16.2	16.3
Jul-20	1616	5.8	3948	24.8	12.6	12.3
New forms published						
Aug-2020	1543	0.7	2980	6.7	4.7	2.0
Sep-20	2123	0.3	4224	9.6	6.0	3.6
Oct-20	2306	0.7	4379	10.0	5.4	4.6
Nov-20	1615	0.7	3203	7.2	4.1	3.1
Dec-20	1823	0.5	2846	4.3	2.9	1.3
Jan-21	1778	0.3	3479	4.9	3.3	1.6

Table 4. Security Flags by Modality and Time Period.

Mode	*N*	Pass rate (%)	Slow correct resp. (%)	Fast correct resp. (%)	Fast pass (%)	Neg RT correl (%)[a]	Low easy versus Hard perf. Delta (%)	High oper. versus Pre. Perf. Delta (%)
Pre-OP								
TC	17176	62.0	0.8	0.2	0.0	0.1	0.1	0.4
OP begins								
OP	18599	75.4	0.9	5.4	0.3	4.0	0.1	0.4
TC	5856	59.5	0.7	0.6	0.0	0.4	0.0	0.7
New forms published								
OP	18131	76.7	1.0	2.4	0.1	1.6	0.0	0.1
TC	9596	68.5	0.9	0.2	0.0	0.1	0.1	0.2

[a] Negative correlation between item response times and average response times.

Table 5. Select Security Flags and Collusion Flags.

		No collusion			Collusion			
Mode	*N*	Fast correct resp. (%)	Neg RT correl (%)	Pass rate (%)	*N*	Fast correct resp. (%)	Neg RT correl (%)	Pass rate (%)
Pre-OP								
TC	17172	0.2	0.1	62	4			
OP begins								
OP	13988	0.9	1.5	67.8	4611	19.2	11.5	98.2
TC	5632	0.4	0.3	58.2	224	6.3	4.5	92.4
New forms published								
OP	16767	0.7	0.7	74.9	1364	23.0	12.9	98.5
TC	9547	0.2	0.1	68.3	49	8.2	0.0	89.8

the different sequences of modalities. The first thing to point out is that very few repeaters changed modality between Time 1 and Time 2. Test takers who began in OP but switched to TC had a much higher incomplete rate for Time 1 than those in the OP- > OP group (26% vs 7%), suggesting that they may have had technical problems and switched to test centers when able. The table also shows that the TC- > TC test takers had the lowest Time 2 pass rate (39%) and the smallest score gain (5.3), while the TC- > OP test takers had the highest pass rate (53%) and largest score gain (11.6), as expected. Finally, those with incomplete Time 1 results have a similar Time 2 pass rate to the overall first-time pass rate on the test (81%), strongly indicating that they should be excluded from further retest analyses, as the score gains are likely due to legitimately finishing the test.

Table 7 incorporates the collusion flags and excludes test takers with an incomplete Time 1 or Time 2 test. Prior to OP there were no repeat candidates flagged for collusion, the retest pass rate was 40%, and the average number-correct score change was 4.9 points. In both initial OP and later OP, retest pass rates are much higher for test takers flagged for collusion than they are for those not flagged and were even higher than the first-time complete test taker population of 81%. Score changes are largest for the TC- > OP group, and retest pass rates are also highest for that group, along with the handful of OP- > TC cases. This is true in both initial OP and later OP samples.

When comparing just the No Collusion OP- > OP and TC- > TC groups, the OP group shows about a 1-point higher score change, but a 10% higher pass rate than the TC group. While changes in the population, or changes to training and preparation may have resulted in the initial months of the pandemic, such changes would not seem to explain the differences between modalities.

Discussion

Two salient findings emerge from these two studies. Study 1 shows that the magnitude of high answer overlap across many programs corresponds to the relative level of security presumed in each delivery mode; however, the magnitude of this difference is related to the composition of the samples evaluated. Study 2 shows that, within the context of a single program, the onset of the OP delivery option resulted in increased high answer overlap and that high answer overlap is accompanied by corroborating indicators of cheating, namely, higher incidence of rapid response times and lower correlation between observed and expected response times. Retest score gains in Study 2 also indicate that Time 1/Time 2 modalities are associated with differences in repeat performance and pass rates, as are test takers flagged for possible collusion. Results such as these should be taken seriously when evaluating the suitability of delivery mode for a particular testing application.

Table 6. Retest Performance by Mode and Completion Status.

	Time 1 complete			Time 1 incomplete		
Time 1 modality/Time 2 modality	*N*	Pass %	Score change	*N*	Pass %	Score change
OP- > OP	2949	52	7.6 (10.6)	218	81	38.2 (15.2)
OP- > TC	212	50	7.8 (7.8)	75	80	39.7 (17.8)
TC- > OP	748	53	11.6 (14.7)	1		
TC- > TC	3056	39	5.3 (6.7)			

Table 7. Retest Changes by Mode and Time Period for Complete Tests.

	No collusion			Collusion		
Seq	*N*	Pass rate (%)	Mean score change (SD)	*N*	Pass rate (%)	Mean score change (SD)
Pre-OP						
TC- > TC	1385	40	4.9 (6.0)	0		
OP begins						
OP- > OP	1126	46	6.8 (8.6)	192	94	21.2 (10.9)
OP- > TC	76	47	8.1 (8.1)	6		
TC- > OP	393	41	8.3 (11.5)	108	100	29.1 (10.5)
TC- > TC	663	36	5.6 (7.7)	27	93	18.1 (8.6)
New forms published						
OP- > OP	1499	49	6.7 (8.2)	85	94	20.2 (12.4)
OP- > TC	128	48	6.8 (6.8)	2		
TC- > OP	188	47	8.7 (11.8)	36	100	27.5 (10.1)
TC- > TC	974	37	5.4 (6.4)	7		

An analysis of collusion flags from the various test sponsors represented in Figure 1 (not reported for confidentiality reasons) shows different testing programs are subject to different levels of cheating for different reasons. A self-assessment, for example, is unlikely to elicit cheating as there are no stakes attached to the results and the test takers are interested in learning from the valid results of the test. If grades or other rewards were attached to the results, it is possible that some level of cheating might begin to occur. Testing programs with financial rewards attached to passing status (promotion or salary incentives), especially those with course-sized prerequisites, as opposed to curriculum-sized prerequisites, might inspire high levels of cheating. This is especially true when, as is often the case, such programs test in diverse locations in the interest of market penetration. While the results reported in this article cover 326 different testing programs and over 3 million test takers, the generalizability of these results to a specific testing program will depend on the characteristics of that program. These results do indicate that collusion is more likely in OP settings than TC overall. This pattern is maintained when controls for sample size and comparability of test centers offered is controlled for, although the magnitude changes. Moreover, as noted above, while all TC results showed lower levels of collusion than OP, test centers with fewer controls over hardware and security showed more flags than those with greater controls.

The analysis of a single test program over time in Study 2 also showed higher levels of collusion in OP than in TC. The baseline information provided for the program indicates that test content was likely not widely available prior to OP. Following the onset of OP, there was an increase in indications of preknowledge in test centers. The increase in RSI combined with the lack of 100% overlap suggests that some TC test takers were memorizing the exposed content. The high RSI along with high levels of 100% overlap suggest that OP colluders feature a combination of those memorizing exposed content and proxy test takers. It is possible, but unlikely, that numerous test takers would memorize the exposed content so well that their responses would be identical. It is more likely that individuals testing for others would respond in the same way over multiple tests (Becker & Makransky, 2011).

The additional test security flags also showed an informative interaction with collusion flags. Pass rates were much higher for flagged test takers in the overall population and the first-time retest population. Additionally, fast correct responses and negative response time correlations were also much more common in the collusion-flagged group. The authors suggest that extreme levels of response overlap combined with non-extreme scores may be indicative of proxy test takers. Research in this area is ongoing. Relatively short tests, relatively shallow content domains, and relatively exposed content will lead to high levels of correct overlap, but not necessarily exact incorrect overlap. As organizations identify verified test takers who are memorizing material versus verified proxy test takers, we will hopefully increase our knowledge of their characteristics.

This report is intended to show how one may usefully analyze operational data to evaluate the security of online proctored test administrations relative to test center test administrations. We do not think that the instances of high overlap or the test security flags reported here constitute the "true rate of misconduct." Not every cheating instance may trigger a flag. While chance overlaps to the degree reported here would be extremely unlikely, a few could happen, while other more idiosyncratic efforts to cheat might go undetected. Continued research on cheating detection, as well as effective responses to cheating, continue to develop. Many factors contributing to the likelihood of cheating and the ability to detect it could not be included here, such as the value or stakes of programs, geographic considerations, exposure, test length, or content breadth. Given the concentration of interest in these issues by test delivery providers and sponsors, it is also likely that OP technology and program actions regarding security issues will continue to change the test security environment going forward. The authors encourage individual programs using OP, and testing vendors offering this service, to evaluate the security of their testing modalities.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

ORCID iD

Kirk A. Becker https://orcid.org/0000-0002-4141-8740

Supplemental Material

Supplemental material for this article is available online.

References

American Educational Research Association. (2014). American psychological association, national council on measurement in education. *Standards for educational and psychological testing*. AERA. https://www.testingstandards.net/uploads/7/6/6/4/76643089/standards_2014edition.pdf

Becker, K. A., & Jones, P. (2022). Varying item parameters and collusion detection. *Paper presented at the national council on measurement in education annual meeting*. ERIC

Becker, K. A., & Makransky, G. (2011). Verifying candidate identity over time: Candidate response consistency for repeated test items. *Presented at the association of test publishers annual conference*. Association of Test

Becker, K. A., & Meng, H. (2022). Identifying statistically actionable collusion in remote proctored exams. *Journal of Applied Testing Technology*, *23*(Special Issue), 54–61.

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

Cizek, G. J., & Wollack, J. A. (2017). *Handbook of quantitative methods for detecting cheating on tests*. Routledge.

Foster, C., & Marder, A. (2020). *Stealing test content with hidden cameras*. Presentation at the Association of Test Publishers annual meeting, Online.

Hurtz, G. M., & Weiner, J. A. (2022). Comparability and integrity of online remote vs. onsite proctored credentialing exams. *Journal of Applied Testing Technologies*, *23*(Special Issue), 36–45.

Kingston, N. M., & Clark, A. K. (2014). *Test fraud*. Routledge.

Maynes, D. (2012). Detection of non-independent test taking by similarity analysis. *Paper presented at the statistical detection of potential test fraud conference*. University of Wisconsin–Madison

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*. Taylor & Francis

Saupe, J. L. (1960). An empirical model for the corroboration of suspected cheating on multiple-choice tests. *Educational and Psychological Measurement*, *20*(3), 475–489. https://doi.org/10.1177/001316446002000304

Smith, R. (2019). Comparing B3 to answer similarity Index for detecting collusion. *Paper presented at the annual meeting of the conference on test security*. University of Miami

Wollack, J. A., & Fremer, J. J. (2013). *Handbook of test security*. 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*. Taylor & Francis