Journal of Applied Testing Technology, Vol 23(Special Issue), 36-45, 2022

Comparability and Integrity of Online Remote Vs. Onsite Proctored Credentialing Exams

Gregory M. Hurtz[1,2] and John A. Weiner[3]*

[1]Professor, Department of Psychology, California State University, Sacramento 6000, J St. Sacramento, CA 95819-6007, USA; ghurtz@csus.edu

[2]Senior Research Scientist, PSI Services LLC, 611 N. Brand Blvd, 10th Floor Glendale, CA 91203, USA; ghurtz@psionline.com

[3]Chief Science Officer, PSI Services LLC 611 N. Brand Blvd, 10th Floor Glendale, CA 91203, USA; john@psionline.com

Abstract

Since the onset of the pandemic in 2020, many credentialing organizations have incorporated online remote administration of their examinations to enable continuity of their programs. This paper describes a research study examining several high stakes credentialing examination programs that utilized mixed delivery modes, including online remote testing at home, as well as testing in test centers. Candidates were monitored in real time by a test proctor, either remotely by video camera, or in person. The study examined the comparability of test scores, instances of irregular candidate testing behavior (potential cheating), and candidate test taking experience ratings across modalities. Overall, results of the study indicated that test scores were psychometrically sound and comparable across modes; rates of suspect test taking behavior were low and not significantly different across modes; and candidate experience ratings were favorable and unrelated to testing modality. Implications for future practice and research are discussed.

Keywords: Remote Proctoring, Remote Online Proctoring, Remote Invigilation, Equivalence of Proctoring Conditions, Test Security, Preventing Online Test Fraud, Data Forensics, High-Stakes Testing, Computer-based Testing, Internet-based Testing, Test-taker Perceptions

1. Introduction

Remote online delivery of high stakes tests has been steadily growing in acceptance in the educational and credentialing realms for some time, and quickly became the norm during the COVID-19 pandemic that struck globally in early 2020. The need for valid academic and professional assessments has not ended while much of the world has been in and out of lockdown status for nearly one year as of this writing. Academic institutions have continued to assign course grades and award degrees; state and national licensing bodies have continued to award licenses; and professional certification providers have continued to award occupational certifications. Halting these activities would only contribute to the strain on the economy that is already resulting from the pandemic and potentially contribute to shortages of properly educated, licensed, and credentialed individuals in the workforce for many occupations. At the same time, continuing these activities without proper safeguards to uphold their validity may lead to unqualified individuals receiving diplomas and occupational licenses/credentials, which could in turn put people and organizations at risk.

While some testing programs were already delivering remote tests before the pandemic others moved to remote delivery in response to the pandemic. It is likely that being thrusted into remote delivery has moved the

*Author for correspondence

industry rapidly in a direction it was already headed and has brought us to a "new normal" for testing that we would have eventually reached, just more gradually. From a research perspective this large-scale move to remote testing has provided a surge in the volume of data available for further evaluating the efficacy of remote testing. To this end, in this study we sought to evaluate the equivalence of tests administered in remote versus traditional delivery modes.

2. Literature Review

In recent years, increasing attention has been given to online delivery of high stakes tests in order to evaluate equivalence in terms of test validity, test security, and test-taker reactions. Weiner and Hurtz (2017) summarized the very limited body of previously published research on remote testing and noted that the research was limited to general commentaries with no data, surveys of user reactions, and comparisons of proctored to completely unproctored tests. They found no research directly comparing remotely proctored to onsite proctored tests. To fill this gap, they conducted a quasi-experimental study comparing psychometric properties and reactions of test-takers in remotely proctored versus traditional onsite high stakes test administration conditions. Across three professional licensure examinations administered in either remote kiosk locations or traditional brick-and-mortar test centers, with a total N of 14,623, they found the test score distributions and psychometric properties from both classical and item response theory analyses to be highly equivalent across modalities. Overall reactions from test-takers indicated a high degree of satisfaction with both modes of test delivery, and slight differences in survey responses were found only for ratings of testing conditions (which included noise, temperature, and distractions) in favor of traditional test centers. Questions regarding factors such as the clarity of the onscreen computer tutorial and the computer testing system suggested positive reactions in both modes, and survey ratings were found to have no relationship with test performance. Altogether, that study provided evidence supporting the soundness and equivalence of kiosk and test-center modes of delivery.

In a subsequent study, Spence *et al.* (2019) analyzed continuing education test scores and survey results from 1,217 certified nurses, approximately half in-person and half remotely proctored, who were also split across closed-book and two open-book conditions. Average test scores between remote and in-person conditions were equivalent, as were pass rates across all conditions of their study. These findings corroborated Weiner and Hurtz's (2017) earlier results revealing no performance differences between remote and in-person conditions. In slight contrast to Weiner and Hurtz, Spence et al. found more suggestions in their test-taker survey of technical problems with the online system, but also found test-takers to feel that remote testing would help reduce their anxiety. Together, these two studies found test performance to be equivalent across delivery modes and the issues revealed in test-taker surveys are easily addressable through technical improvements to the testing system and recommendations to test-takers for choosing their physical testing environments.

While no additional published research has been located, the results of these two studies are very promising for allaying trepidations that testing programs may have about moving toward remotely delivered and proctored tests either in place of, or in addition to, traditional onsite proctored examinations. Their research in 2017 and 2019 was quite timely given the onset of the COVID-19 pandemic in early 2020, when a large portion of testing activity was moved – even if temporarily in some cases – into the realm of remote test delivery and remote proctoring.

3. Present Study

The current study replicates and extends the prior pioneering research to continue exploring comparability of secure remote and onsite test delivery modes in several ways, in that it: compares online testing-at-home vs. test center conditions; provides a new comparison of test outcomes and candidate survey responses from a different set of professional licensure testing exams and a national certification exam program to build on the body of evidence; and extends the comparative analysis to data forensics indices designed to detect test-taker collusion that could potentially undermine the integrity of the examination.

The study was carried out with data from multiple licensing examinations offered by a state regulatory body where both online remote proctored and test center proctored administration had been offered since late 2019, and a national certification exam that began offering online remote testing in May 2020, during the

pandemic. For the former testing program, for the six months prior to the pandemic-related closures of test centers, candidates had a choice of delivery mode based on their selection of testing location. For approximately 10 weeks during the closures, this choice was severely restricted as most physical test centers were closed and candidates could only test remotely, or not at all. For the latter exam program testing was only offered onsite until the remote option was added in response to the pandemic. For both exam programs, multi-mode testing continues as of this writing with largely reinstated candidate choice in accordance with social distancing policies and possible local closures depending on a candidate's geographical location.

Research Questions

The study examined three main research questions pertaining to measurement quality, comparability, and integrity. The first research question was whether differences in test scores emerge between online remote (test at home) proctored versus onsite delivered and proctored exams. *Will the evidence continue to suggest no systematic effects of delivery mode on the psychometric properties of tests, consistent with the prior studies reviewed above?* The second research question was whether rates of candidate misconduct as indexed by data forensics indices would differ between delivery modes. *Are the detection rates for irregular test-taking behavior different when tests are delivered remotely online versus onsite?* The third research question was whether candidate reactions would differ across testing modes—*will the evidence continue to suggest high rates of positive reactions to both remote and onsite testing?*

4. Methods

4.1 Participants

The sample was comprised of examinees for four state licensing exams and one national certification exam. For the state licensing exams candidate test data and post-test survey responses were accumulated for 15 months (67 weeks) starting in October 2019, through the end of December 2020. For the national certification exam, the data for the study began when remote testing was introduced in May 2020 and continued through the end of December 2020. For all exams, first-time candidates who completed at least 90% of the test items were included in the analysis; retakes were excluded, as well as candidates with more than 10% missing responses. The analysis dataset included a total of 11,376 onsite candidates and 8,071 candidates whose tests were administered and proctored remotely, for a total of 19,447 candidates. Table 1 shows the percentage breakdown of remote versus onsite testing overall, and within three time-frames relative to pandemic-related test center closures. Candidate demographics were not available.

4.2 Materials

All examinations were fixed-form, multiple-choice knowledge exams with 100 to 150 scored items plus blocks of experimental (unscored) items. The experimental items were excluded for the purposes of this study. For each examination, alternate equivalent forms were administered to candidates to enhance test

Table 1. Distributions of candidates between remote and onsite conditions across pre-closure, during closure, and partial reopening time periods for the data in this study

Examination Programs	Testing Mode	Pre-Closures (26 weeks)	During Closures (10 weeks)	Partial Reopening (31 weeks)	Total (67 weeks)
State Licensing Exams	Remote	9.9%	91.5%	78.8%	54.2%
	Onsite	90.1%	8.5%	21.2%	45.8%
National Certification Exam	Remote		49.4%	34.7%	36.6%
	Onsite		50.6%	65.3%	63.4%
Total	Remote	9.9%	70.7%	51.6%	45.8%
	Onsite	90.1%	29.3%	48.4%	54.2%

security. The number of alternate forms ranged from 3 to 5 for each exam. These examinations had been previously developed by testing experts according to professional standards (AERA *et al.*, 2014) for content validity, reliability, and psychometric quality. The tests were administered continuously via computer in one of two administration modes: (1) Onsite at a secure test center while being supervised by onsite proctors, or (2) Online at a location of the test taker's choosing that met organizational standards, while being supervised by a live remote proctor, using video communication and surveillance.

At the conclusion of the test, examinees completed a post-examination survey to rate their testing experience including 5 areas of focus in the current study.[1] Two questions were related to the testing software: "Was the testing software easy to use? (Yes/No)"; "Were the on-screen instructions easy to follow? (Yes/No). Two additional questions were related to the proctors: "Was the Assessment Center Proctor/Supervisor friendly? (Yes/No)"; "Was the Assessment Center Proctor/Supervisor helpful in solving any problems you encountered during testing? (Yes/No/Not applicable)." The final question was related to the testing environment: "How would you rate the noise level during testing" (Quiet; Occasional noise, but not distracting; Distracted during testing by the noise level; or Unacceptable for testing)." For analysis purposes, responses to the final question were dichotomized into positive and negative categories, as follows: the first two response categories were combined to indicate a positive environment, and the last two response options were combined to indicate a negative environment.

4.3 Analysis Procedure

The first research question was addressed with independent-groups comparisons of delivery mode differences in test score means (*M*), standard deviations (*SD*), and reliabilities (KR-20), as well as psychometric properties of test items, including difficulty (*p*-value) and point-biserial correlations (r_{pb}). The second research question regarding irregularities in test-taker behaviors was similarly addressed using between-mode independent-group comparisons to explore differences in rates of extreme response similarity, anomalous item response patterns, and abnormally fast response times. Finally, the third research question regarding between-mode comparisons in test taker experience ratings was evaluated with analysis of two-by-two contingency tables. For each comparison, a measure of effect size was provided and interpreted in conjunction with the result of a test of statistical significance. Many of the comparisons were based on very large *N*'s such that negligible effects were deemed statistically significant; conclusions from such effects were made based on the effect magnitude.

[1] Other questions were excluded because they were unrelated to test taking in both online and onsite modes of delivery.

5. Results

Research Question 1: Comparison of Test Scores and Item Properties

Table 2 provides summary statistics for each credentialing exam (I – V), for online remote versus onsite delivery modes, including: test score mean (M), standard deviation (SD), percent (%) passing, reliability (KR-20) and standard error of measurement (SEM); and item properties including mean item difficulty (*p*-value) and mean point-biserial correlation (r_{pb}). Values shown in the table are pooled across alternate equivalent forms to maximize sample size and minimize the number of repeated statistical tests. Each statistic in columns of the table is compared across test administration modes with an effect size index and a test of statistical significance, as defined in the table notes.

In terms of cross-mode equivalence, no mean test score differences were found between online remote and onsite administration modes; none of the comparisons were statistically significant and the *d* effect sizes were trivial in magnitude (.01 to .09 in absolute value). Score variances were equivalent across modes in three of the five exams, but for two exams (I and V) Levene's test for the equality of variances was statistically significant, in both cases revealing slightly less score variance (by a fraction of .74 to .92) in remote conditions than onsite. In terms of the percentage of candidate scores exceeding the passing score cutoff, these percentages did not differ for four out of five exams; for one exam (V) the remote candidate pass rate (70.1) was statistically significantly higher than the onsite candidate rate (67.2), although the phi effect size index was small (.03). Altogether these findings do not suggest notable systematic differences in scores obtained by candidates in online remote and onsite conditions.

Likewise, for the comparisons of form properties no differences were found. Mean *p*-values were equivalent

Table 2. Comparisons of score distributions and psychometric properties across testing modes

Exam	Forms	Items	Mode	*N*	*M*	*SD*	% Passing	Mean *p*-value	Mean r_{pb}	KR-20	SEM
I	4	100	Remote	501	71.72	9.79	60.8	.72	.19	.82	4.16
			Onsite	274	70.81	11.39	59.5	.71	.23	.86	4.16
			Effect Size		0.09[a]	0.74[b]*	.01[c]	0.09[a]	-0.04[d]	.04[e]	1.00[f]
II	5	150	Remote	2233	72.25	11.78	61.7	.72	.26	.80	5.09
			Onsite	1880	72.81	11.57	62.7	.73	.24	.81	5.12
			Effect Size		-0.05[a]	1.04[b]	-.01[c]	-0.05[a]	0.01[d]	.01[e]	0.99[f]
III	4	100	Remote	1234	69.72	13.50	53.9	.70	.28	.90	4.19
			Onsite	932	69.52	13.80	54.8	.70	.29	.91	4.19
			Effect Size		0.01[a]	0.96[b]	-.01[c]	0.01[a]	0.01[d]	.01[e]	1.00[f]
IV	5	150	Remote	1508	69.11	10.70	50.8	.69	.21	.76	5.25
			Onsite	1580	69.21	10.65	49.6	.69	.21	.76	5.25
			Effect Size		-0.01[a]	1.01[b]	.01[c]	-0.01[a]	0.00[d]	.00[e]	1.00[f]
V	3	140	Remote	5900	67.51	12.11	70.1	.68	.24	.91	5.14
			Onsite	3405	66.76	12.66	67.2	.67	.26	.91	5.16
			Effect Size		0.06[a]	0.92[b]*	.03[c]*	0.06[a]	0.02[d]	.00[e]	0.99[f]

**p<.05.*

Notes: *N* = number of test-takers; *M* and *SD* are marginal means and standard deviations of percent-correct scores across three to five equivalent alternate forms per exam; *p*-value = item difficulty; r_{pb} = point-biserial item-total correlation adjusted for overlap.

[a] Cohen's *d* effect size; statistical significance evaluated with an independent-groups *t*-test.

[b] Ratio of larger to smaller variance; statistical significance evaluated with Levene's test of the equality of variances.

[c] Phi coefficient for the 2 x 2 contingency table; statistical significance evaluated with Fisher's exact test.

[d] Difference between independent correlation coefficients; statistical significance evaluated with a standard Z test for independent correlations.

[e] Difference between KR-20 coefficients; statistical significance evaluated with Bonett's (2003) Z test.

[f] Ratio of larger to smaller error variance (SEM^2); statistical significance not directly evaluated since its component parts – SD and KR-20 – were separately tested.

($p > .05$; *d* ranged from .01 to .09) across conditions for all five examinations, as were point-biserial correlation coefficients ($p > .05$; differences in *r* of .00 to .04). Using a z-test by Bonett (2003), KR-20 reliability coefficients were found not to differ across modes ($p > .05$; differences in KR-20 ranged from .00 to .04). SEM values were virtually identical with ratios of squared SEMs ranging .99 to 1.00,

which is not surprising given that tests of their component parts – SDs and reliabilities – revealed equivalence.

Research Question 2: Comparison of Statistical Indices of Irregular Candidate Behaviors

Three types of indices were used to assess irregularities in candidate response behaviors: response similarity, item score patterns, and response speed. Results are summarized in Table 3 and discussed below for each index.

Response Similarity

The J_2 index of similarity (Weiner et al., 2013; Hurtz & Weiner, 2019) was used to measure excessive degrees of identical answer choices, as would be consistent with candidates either colluding directly with each

Table 3. Comparisons of data forensics index distributions across testing modes

			J_2			MCI			Tau-j		
Exam	Mode	*N*	*M*	*SD*	High	*M*	*SD*	High	*M*	*SD*	High
I	Remote	501	-0.27	0.92	0.8%	0.32	0.08	0.0%	-0.02	0.36	0.0%
	Onsite	274	-0.31	0.81	0.4%	0.32	0.08	1.1%	-0.12	0.37	0.0%
	Effect Size		0.04[a]	1.30[b]	0.03[c]	-0.08[a]	1.06[b]	-0.08[c]	0.25[a]*	1.02[b]	n/a
II	Remote	2233	-0.35	0.77	0.6%	0.32	0.06	0.5%	0.06	0.34	0.0%
	Onsite	1880	-0.40	0.63	0.0%	0.32	0.06	0.3%	0.01	0.34	0.0%
	Effect Size		0.06[a]	1.52[b]	0.05[c]*	-0.05[a]	1.02[b]	0.01[c]	0.13[a]*	1.00[b]	0.01[c]
III	Remote	1234	-0.28	0.74	0.2%	0.33	0.08	0.6%	-0.03	0.39	0.5%
	Onsite	932	-0.30	0.76	0.3%	0.33	0.07	0.2%	-0.14	0.38	0.0%
	Effect Size		0.02[a]	1.05[b]	-0.01[c]	-0.04[a]	1.11[b]	0.03[c]	0.28[a]*	1.05[b]	0.05[c]*
IV	Remote	1508	-0.28	0.80	0.1%	0.31	0.06	0.4%	0.03	0.31	0.0%
	Onsite	1580	-0.28	0.83	0.4%	0.30	0.06	0.4%	-0.03	0.32	0.0%
	Effect Size		0.01[a]	1.08[b]	-0.02[c]	0.03[a]	1.02[b]	0.00[c]	0.18[a]*	1.08[b]	n/a
V	Remote	3405	-0.07	0.95	0.5%	0.33	0.06	0.3%	0.03	0.23	1.0%
	Onsite	5900	-0.11	0.94	0.4%	0.33	0.06	0.4%	-0.01	0.22	0.6%
	Effect Size		0.03[a]	1.02[b]	0.01[c]	-0.03[a]	1.02[b]	-0.01[c]	0.17[a]*	1.13[b]*	0.02[c]*

*$p<.05$

Notes: J_2 is a measure of response similarity between pairs of candidates; MCI is a measure of irregularity in patterns of items scores relative to item difficulties; Tau-j is a measure of response speed; high values on all indices suggest test taking behavior that departs substantially from normal behavior.

[a] Cohen's *d* effect size; statistical significance evaluated with an independent-groups *t*-test.

[b] Ratio of larger to smaller variance; statistical significance evaluated with Levene's test of the equality of variances.

[c] Phi coefficient for the 2 x 2 contingency table; statistical significance evaluated with Fisher's exact test.

other or having advance access to the same items and answers. The J_2 index is a standardized residual from a regression analysis predicting a candidate's maximum number of matches with another candidate from their own test score. J_2 distributions were compared between modes to evaluate differences in the means, standard deviations, and percentages of cases flagged for extreme high similarity. Mean differences between testing modes were non-significant (*p*s > .05) with trivial effect sizes (*d*s ranging .01 to .06). Standard deviations likewise did not significantly differ (*p*s > .05; variance ratios ranging 1.02 to 1.52). Finally, an analysis of the percentage of extreme cases of similarity (J_2 > 3) resulted in nonsignificant chi-squared tests (*p*s > .05; phi coefficients ranging -.01 to .03) for four out of five exams; for one exam the chi-squared test was statistically significant ($p < .05$) but trivial in effect size (phi = .05). For this comparison the remote mode had more flagged candidates (0.6%) than the onsite condition (0.0%); however, scanning all other conditions in the table it appears that the detected effect here was more a function of the onsite condition being lower than normal for this exam, than the remote condition being higher than normal. The percent of candidates flagged across all conditions was never more than 0.8%. Overall, the pattern of results suggests no notable differences in response similarity distributions between test administration modes.

Response Patterns

The modified caution index (MCI; Harnisch & Linn, 1981) served as an index of abnormal response patterns from test takers. MCI is sensitive to response patterns that lead to item scores being inconsistent with the expected scores based on one's estimated ability and the items' difficulties. MCI is sensitive to multiple irregular patterns (Karabatsos, 2003) including responses stemming from advance knowledge of answers to difficult items (Hurtz & Weiner, 2019; Karabatsos, 2003). As seen in Table 3, there were no statistically significant differences (all *p*s > .05) between delivery modes in the means (*d*s ranging .01 to .06), standard deviations (variance ratios ranging 1.02 to 1.11), or percentages of flagged extreme (MCI > 3 SDs from the mean) cases (phis ranging -.08 to .03). Across all conditions the percentage of flagged cases ranged from 0.0-1.1%. Thus, the distributions of item response pattern irregularities were also equivalent between delivery modes.

Response Speed

The third comparison was on the tau-j index (van der Linden, 2006) which measures response speed, where high values indicate a pattern of response times that are systematically faster than normal for the items. Table 3 shows that for all five exams, candidates in the remote administration conditions on average had statistically significantly (*p*s < .05) faster response tendencies with effect sizes near what is commonly considered nontrivial but small (*d*s ranging .13 to .28). For one exam the standard deviations of tau-j were statistically significantly different ($p < .05$) yet the magnitude of the difference was trivial (variance ratio of 1.13). Likewise, for two exams the percentage of flagged extreme (tau-j > 3 SDs from the mean) cases was statistically significantly different but trivial in magnitude (phis = .02 and .05). The percentage of flagged cases ranged from 0.0-1.0%. Overall, the main pattern of findings suggests a tendency for remote candidates to respond slightly faster than onsite candidates.

The faster responding raises the question of whether this difference in response speed is associated with the other indices of irregularities in candidate behavior, or to test scores. Correlations of tau-j with these other measures revealed statistically significant but trivial correlations with J_2 ($r = .02$, $p < .05$) and MCI ($r = -.02$, $p < .05$), and a statistically significant and nontrivial but small correlation of speed with test performance ($r = .21$, $p < .05$). When isolating just the remote candidates, the respective correlations were similar if not slightly smaller ($r = .01$, $p > .05$; $r = -.01$, $p > .05$; $r = .17$, $p < .05$). Overall, these correlations do not raise any concern over speed being associated with misbehavior.

Research Question 3: Comparison of Candidate Reactions to their Testing Experiences

Candidate reactions on the five post-test survey items are summarized in Figure 1, which shows the percentages of candidates in remote versus onsite testing modes giving a positive response to each item. Prior to creating this figure, the percentages were compared across exams and found to be consistent in the direction of effects with only some negligible differences in magnitude, supporting the pooling of results across exams to present the overall findings. Overall, these results show a high level of favorable reactions (93–99%) to both remote and onsite proctored exams.

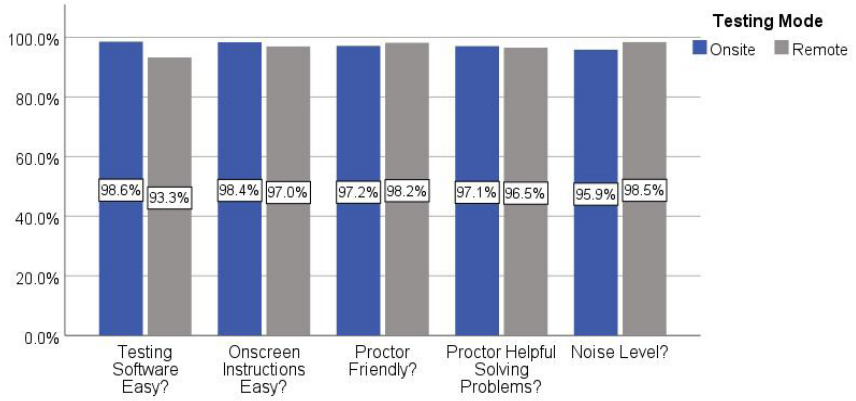

Figure 1. Percentages giving positive responses to reaction items, by mode.

Table 4. Test of association between candidate reactions and test scores

Reaction	N Responding	Bivariate Association		Controlling Exam & Mode Variance		Test of Reaction × Mode Interaction	
		β	r^2	β	ΔR^2	β	ΔR^2
-Software easy to use?	18,785	.01	.00	-.02*	.00	-.02	.00
-On-screen instructions easy?	18,784	-.01	.00	.00	.00	.06	.00
-Proctor friendly?	18,888	-.03	.00	.02*	.00	.13*	.00
-Proctor helpful, if problems?	10,238	.03*	.00	.03*	.00	-.02	.00
-Noise level during testing?	16,270	.00	.00	.00	.00	.18*	.00

Note. Reaction × Exam interactions were also tested which likewise had trivial effect sizes, and none were statistically significant.

While reactions overall were very positive in both modes, some slight differences in patterns across the five items are worth noting. For the first two items regarding ease of the software and onscreen instructions, somewhat more positive ratings were given to the onsite condition than remote (phi = -.15 and -.02 respectively, *p*s < .05). The difference in ease of the software was slightly stronger with a 5.3% difference in the percentages giving a positive reaction, while the difference in onscreen instruction ratings was very small at 1.4%. On the other hand, a slightly higher percentage of positive ratings was given to the online remote proctored condition for proctor friendliness (phi = .05, $p < .05$) and noise in the testing environment (phi = .10, $p < .05$). These differences were again quite small with 1.0% and 2.4% difference in percentages, respectively.

Despite these slight differences in patterns, Table 4 reveals that candidate reactions were virtually uncorrelated with

[2]While some exam comparisons were statistically significant indicating that the remote vs. onsite differences varied across exams, this was driven by the high statistical power to reject the null hypothesis for practically insignificant effects. The nature of the differences amounted to slight variations in magnitude but not direction of group effects, and the impact on the mean percentages shown in Figure 1 was negligible.

test scores in that none of the reaction variables explained enough variance in performance to register beyond an r^2 of .00. Some beta weights in a series of regressions presented in the table were statistically significant, in the simple bivariate analysis as well as an analysis partialling out exam and mode differences, and also an interaction model testing for different effects of reactions across modes; however, this is clearly driven by the very large *N* leading to high statistical power for deeming negligible relationships to be different from zero.

6. Discussion

The purpose of this study was to replicate and extend the limited body of past research comparing remote to onsite testing modes. The first research question focused on delivery mode comparisons of psychometric properties of test scores and items, and results showed no such differences. The second research question focused on comparisons of data forensics indices designed to detect irregularities in candidate response patterns and results again showed no such differences between online and onsite proctored exams. The third research question focused on comparisons of candidate survey responses between modes, and revealed that reactions were very positive for both testing modes with small differences between modes; remote online proctored candidates had slightly lower rates of indicating the software was easy to use, slightly higher rates of indicating that the proctors were friendly, and slightly higher rates of indicating that their testing environment was free from noise distractions.

Altogether the findings support the equivalence of assessments delivered and proctored remotely, as compared to onsite exams, and suggest some benefits in terms of positive proctor interactions and control over the testing environment. Occasional software issues may occur in self-service applications using the candidates' own equipment but even so, 93% of remote candidates said the software was easy to use. Nevertheless, test taker experiences in the five questions addressed by the reaction survey were unrelated to their test scores, suggesting that the conditions did not hinder performance.

The findings of this study are consistent with the limited body of past research (Weiner & Hurtz, 2017; Spence *et al.*, 2019). Weiner and Hurtz likewise found no differences in test score distributions or psychometric properties of tests delivered onsite versus remote kiosks at specific locations, and Spence *et al.* similarly reported consistent means and standard deviations of scores when comparing remote to in-person tests. The current study replicates these findings of equivalence between onsite and remote testing. In addition, the finding of no differences in distributions or rates of extreme values on statistical indices designed to detect irregularities in candidates' item response similarities and patterns lends further evidence to the cross-mode comparability of assessments delivered onsite versus at a location of the candidate's choosing. Consistent with both Weiner and Hurtz, and Spence et al., the survey results here suggest positive candidate reactions to remote testing in a professional testing environment. The mounting evidence appears to suggest no systemic drawbacks to leveraging computer and audiovisual technology for test delivery with online remote proctoring.

This study provided new research examining irregularities in test taking behavior and comparing detection rates between online remote proctored and onsite proctored exams. The findings in this research were encouraging in that the rates of extreme similarity, irregular response patterns, and excessive speed were small, and were not significantly different between delivery modes.

7. Conflict of Interest Disclosure

None.

8. Acknowledgement

None

9. References

AERA, APA, NCME (2014). *Standards for educational and psychological testing.* Washington DC: American Educational Research Association.

Bonett, D. G. (2003). Sample size requirements for comparing two alpha coefficients. *Applied Psychological Measurement,* 27, 72-74. https://doi.org/10.1177/0146621602239477

Harnisch, D. L., & Linn, R. L. (1981). Analysis of item response patterns: Questionable test data and dissimilar curriculum practices. *Journal of Educational Measurement,* 18, 133–146. https://doi.org/10.1111/j.1745-3984.1981.tb00848.x

Hurtz, G. M., & Weiner, J. A. (2019). Analysis of test-taker profiles across a suite of statistical indices for detecting the presence and impact of cheating. *Journal of Applied Testing Technology,* 20(1), 1-15.

Karabatsos, G. (2003). Comparing the aberrant response detection performance of thirty-six person-fit statistics. *Applied Measurement in Education,* 16, 277–298. https://doi.org/10.1207/S15324818AME1604_2

Spence, D., Ward, R., Wooden, S., Browne, M., Song, H., Hawkins, R., & Wojnakowski, M. (2019). Use of resources and method of proctoring during the NBCRNA continued professional certification assessment: Analysis of outcomes. *Journal of Nursing Regulation*, 10(3), 37-46. https://doi.org/10.1016/S2155-8256(19)30147-4

van der Linden, W. J. (2006). A lognormal model for response times on test items. *Journal of Educational and Behavioral Statistics,* 31(2), 181-204. https://doi.org/10.3102/10769986031002181

Weiner, J., Saiar, A., & Granger, E. (2013). An empirical method for the detection of potential test fraud. *Presented at the 2nd annual meeting of the Society for the Detection of Potential Test Fraud in Madison*, Wisconsin.

Weiner, J. A., & Hurtz, G. M. (2017). A comparative study of online remote proctored versus onsite proctored high-stakes exams. *Journal of Applied Testing Technology,* 18, 13-20.