Journal of Applied Testing Technology, Vol 23(Special Issue), 1-4, 2022

Online Remote Proctored Delivery of High Stakes Tests: Issues and Research

John A. Weiner[1]* and Dianne Henderson[2]

[1]Chief Science Officer, PSI Services LLC, 611 N. Brand Blvd, 10th Floor Glendale, CA 91203, USA; john@psionline.com
[2]Vice President, Research, ACT, Iowa City, IA 52243, USA; dianne.henderson@act.org

Abstract

Online administration of high stakes tests has been an increasing trend since the turn of the millennium, as assessment design and delivery have been undergoing a digital transformation. In 2020, the pandemic heightened the need for safe and socially distanced practices in education, work, and life activities. As a result, online Live Remote Proctored (LRP) testing rapidly became a necessary practice to enable credentialing to continue for professionals in essential occupations, and has continued as testing organizations adopt mixed modes of delivery for assessments in all markets. And while limited published research has been promising in support of LRP, there remain unanswered questions and a need for additional research. This special issue of the Journal of Applied Testing Technology includes several articles describing empirical studies that examine key measurement issues, including comparability of LRP and Test Center Proctored (TCP) scores on high stakes exams, detection of score anomalies and potential cheating for LRP and TCP exams, potential impact of testing modality on candidate experience ratings, and relationships between technology disruptions, candidate experience and test scores. A methodological article is included outlining an approach to quantifying candidate response similarity that may be indicative of cheating or other test fraud; and a literature review article is included providing background on the evolution of online testing and research in different assessment contexts.

Keywords: Candidate Experience, Online Remote Proctoring, Modality Effects, Test Fraud, Technology Disruption of Testing, Test Score Comparability, Test Security

1. Introduction

Online administration of high stakes tests has been an increasing trend since the turn of the millennium, as assessment design and delivery have been undergoing a digital transformation (Weiner & Foster, 2018). These advances have enabled test takers to complete online examinations from virtually any location while being monitored by remote proctors using live video, as well as other technologies and analytic methods to support secure administration. This trend was accelerated in 2020 with the pandemic and the need for safe and socially distanced practices in education, work, and life activities. Online LRP[1] testing rapidly became a necessary practice to enable credentialing to continue for professionals in essential occupations, and has continued as testing organizations adopt mixed modes of delivery in their programs. And while limited research has been promising in support of LRP for high stakes examinations (Weiner & Hurtz, 2017;

[1] Live remote proctoring, online remote proctoring, and remote proctoring are used synonymously in this special edition series of articles.

*Author for correspondence

Spence *et al.*, 2019), there remain unanswered questions regarding measurement quality and integrity, and, thus, a need for additional research.

Issues Covered in this Special Issue

This special issue of the *Journal of Applied Testing Technology* includes several articles that describe empirical studies examining key measurement issues, including comparability of LRP and Test Center Proctored (TCP) high stakes exam scores (Morin, Alves, & De Champlain, 2022; Hurtz & Weiner, 2022; Muckle, Meng, & Johnson, 2022), detection rates for score anomalies and potential cheating (Hurtz & Weiner, 2022), potential impact of test delivery modality on candidate experience ratings and their relationships to test scores (Hurtz & Weiner, 2022; Muckle *et al.*, 2022), and potential effects of technology disruptions on candidate experience and test scores (Morin *et al.*, 2022). An article is included that presents an approach to measuring response similarity for detecting potential cheating and collusion (Meng & Becker, 2022), and a literature review article is included providing background on the evolution of online testing and research in different assessment contexts (Langenfeld, 2022).

2. The Evolution of Online Testing

Online testing began well before the Covid-19 pandemic and has had a wide range of applications in distance learning and employment testing, and later in credentialing, with varying levels of stakes and security ranging from Unproctored Internet-based Testing (UIT) for low stakes exams to live human proctored high stakes secure exams. Langenfeld (2022), provides a summary of Internet-based testing and remote proctored testing trends and models and discusses extant research literature, issues and considerations for the use of online testing. Langenfeld traces the historical adoption of online Internet-based testing from the 1990s in secure test centers, to UIT for employment in the 2000s, to the seismic shift to online proctored high stakes credentialing testing in the 2020 pandemic. Published research for these various testing contexts is summarized, which, while limited, is generally supportive of the testing models for their intended applications. Issues and potential risks in online testing approaches are outlined, and approaches to secure test design and delivery technology considerations are discussed and summarized.

3 Empirical Research

Several key issues are examined in empirical studies published in this series, including cross-mode evaluation of psychometric quality and comparability of test scores, security analytics and detection rates for potential cheating and other response anomalies, technology disruptions, and candidate experience ratings as they related to testing mode and test performance.

Comparability: A fundamental consideration with online testing is the psychometric quality and comparability of test scores obtained in LRP versus TCP test administration. Cross-mode studies reported by Morin *et al.* (2022) in a study of a medical college admission examination, and Hurtz and Weiner (2022) in a study of six licensing and certification examinations, found that test score means, standard deviations, and reliability coefficients were comparable between modes. Muckle *et al.* (2022) reported between-mode differences in a pilot study of a pharmacy board exam that utilized both LRP and TCP administration during the pandemic, where they found TCP examinees scored higher. However, differences between examinee characteristics were noted that may account for the observed differences highlighting the need to capture examinee descriptive data in monitoring and evaluating cross-mode exam programs.

Security Analytics: Another concern with online testing is in regard to security and potential cheating. While concerns in this area are as longstanding as the practice of high stakes testing, the increased use of online testing has pushed these concerns to the forefront as testing organizations establish best practices in this area. One approach to monitoring test security and integrity is the use of *data forensics* to detect anomalies in test-taker responses and scores that may be indicative of cheating through a variety of means, as well as potential theft of content and resulting widespread content exposure. Hurtz and Weiner (2022) examined test taker responses in the aforementioned licensure and certification exams. Using a suite of metrics, including a proprietary response similarity index (J_2), an index of response speed (tau-j), and an index of aberrant response patterns (modified caution index), they found no significant differences in detection rates for LRP and TCP modes.

One of the challenges associated with the rise of online remote proctoring is the increased opportunity for item harvesting and content sharing. A common approach to identifying similar response patterns is known as collusion analysis, and involves the computationally and time intensive process of sequentially comparing all response pairs across all test-takers. Meng and Becker (2022) propose the use of a matrix-based approach for quickly calculating exact overlap counts and for determining the associated flagging criteria of suspicious results as a potential new approach, improving the speed of this type of analysis.

Technology: Variability in technology environments raises the potential for construct irrelevant variance in examinee test scores – a potential threat to the validity of scores. One concern is when Internet connections are disrupted, which in turn interrupt the testing session and potentially impact the test taker's experience and performance on the test. Another concern is the delay in displaying test content – "dead time" – that may potentially impact test taker performance. Morin *et al.*, (2022) examined disconnections from the remote testing session and reported that nearly one quarter of test takers in the LRP condition were disconnected during the session and had to reconnect to complete the exam. The rate of disconnections for TCP candidates was only 4%. The study also looked at "dead time" due to delays in the items appearing on the screen. Using multiple regression analysis, Morin *et al.* found that neither disruptions, dead time, nor testing modality accounted for significant variation in test scores, when controlling for source of medical degree. That is, main effects for modality, disruptions, and dead time were non-significant and accounted for a trivial amount of variance in test scores.

Candidate Experience. The potential impact of test delivery modality on candidate experience ratings and their relationships to test scores was examined by Hurtz and Weiner (2022), who reported a high candidate favorability rating (>90%) on five factors related to software use, instructions, proctor interactions, and noise level, both for LRP and TCP modes. Candidate experience ratings were not appreciably associated with test scores or modality (correlation coefficients [r] were essentially zero). Similarly, Muckle *et al.* (2022) examined candidate experience ratings on seven factors related to scheduling, check-in, wait time, proctor interactions, and technical difficulty and two overall satisfaction scales, and reported a greater range in favorability ratings (71% to 98%). However, experience ratings were not significantly correlated with exam scores on 8 of 9 scales and in all cases accounted for less than 1% of the variance in test scores.

4. Discussion

The articles in this special issue summarize the evolution of online testing and present new empirical research studies, adding to the literature and providing support for the use of LRP in high stakes testing programs. Important measurement issues and questions are examined regarding comparability, security analytics, technology disruption, and candidate experience across test proctoring modes and their potential impact on test scores.

While the results of these studies are encouraging and should be helpful in supporting best practices in the application of LRP, especially in mixed mode programs, we encourage continued research and publication of additional studies to further guide best practices and guidelines as the application in high stakes testing becomes more widespread. This is especially important as online testing models evolve and, because not all approaches are the same, empirical results reported in these studies do not necessarily generalize to all online testing programs. Thus, high stakes online testing programs would be well advised to routinely monitor psychometric quality, security, technology disruptions and candidate experience ratings to ensure that potential issues are detected and addressed. To this end, active collaboration between psychometric, operations, and technology professionals is essential.

5. References

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

Langenfeld, T. (2022). Online remote high stakes testing: A solution for the new normal. *Journal of Applied Testing Technology*. Vol 23(Special Issue), 5-14.

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

Morin, M., Alves, C., & De Champlain, A. (2022). The show must go on: Lessons learned from using remote proctoring in a high-stakes medical licensing exam program in response to severe disruption. *Journal of Applied Testing Technology,* Vol 23(Special Issue), 15-35.

Muckle, T., Meng, Y., & Johnson, S. (2022). Quantitative evaluation of a live remote proctoring Pilot. *Journal of Applied Testing Technology.* Vol 23(Special Issue), 46-53.

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.

Weiner, J.A., & Foster, D. (2018). Ch 2. *Licensing and Certification*. In Scott, Bartram, & Reynolds, Eds., Next Generation technology-enhanced Assessment. Cambridge University Press.

Weiner, J. A., & Hurtz, G. M. (2022). A comparative study of online remote proctored versus onsite proctored high-stakes exams. *Journal of Applied Testing Technology,* Vol 23(Special Issue), 36-45.