Canadian Medical Education Journal

Black Ice

Eleven ways to get a grip on the implementation of remote administration of high-stakes assessments

Onze stratégies pour optimiser la gestion des évaluations à enjeux élevés à distance

Christina St-Onge[1]

[1]*Université de Sherbrooke, Quebec, Canada*

Correspondence to: Christina St-Onge, PhD, Faculté de médecine et des sciences de la santé, 3001, 12e Avenue Nord, Local 1109 (CPSS). Sherbrooke, QC J1R 0B2; phone: 819-821-8000, 75047; email: christina.st-onge@usherbrooke.ca

Published ahead of issue: June 10, 2022; published: Aug 26, 2022. CMEJ 2022, 13(4) Available at https://doi.org/10.36834/cmej.73734

Abstract

The COVID-19 pandemic rushed licensure and certification institutions, as well as many university programs, to integrate Information and Communication Technologies (ICTs) in their practices to allow for remote administrations of their exams independent of distancing measures. The Black Ice covered in this manuscript is the integration of ICTs to allow remote administration of high-stakes assessments in terms of its development, administration, and monitoring with the aim to promote the validity of score interpretation.

Résumé

La pandémie de la COVID-19 a eu comme conséquence de rendre nécessaire le recours aux technologies de l'information et de la communication aux évaluations des ordre professionnels, institutions de certification ainsi que de nombreux programmes universitaires. Le terrain glissant exploré dans cet article est celui de l'utilisation des TIC pour permettre la réalisation d'évaluations à enjeux élevés à distance par rapport à leur élaboration, l'administration et le monitorage tout en assurant la validité de l'interprétation des scores.

Introduction

During the Spring 2020, many national licensure and certification exams were postponed because of the COVID-19 pandemic, leaving candidates in limbo with regards to the practice of their chosen profession. Many university programs had to pivot quickly to allow for off-site administration of their assessments. While the pandemic pushed these institutions to integrate Information and Communication Technologies (ICTs) in their practices to allow for remote administrations of their exams independent of distancing measures, these changes were done hastily. The Black Ice covered in this manuscript is the integration of ICTs to allow remote administration of high-stakes assessments, building on the lessons learned over the past two years and previous literature on this topic. To build on these lessons learned I offer considerations for the assessment development, its administration, and its monitoring with the aim to promote the validity of score interpretation[1] that could inform future integration of technologies to written- and performance-based assessment practices. Some of these recommendations may not apply nor be feasible for all assessments, but we could consider them as a goal to work towards.

Considerations while planning the assessment and before the administration

1. Use a stable and reliable platform.

The stability and reliability of the technology used can be a major threat to the validity of score interpretation.[2] For example, when the technology is not stable enough it can disconnect candidates during the assessment, undoubtedly hindering the quality of the experience. For example, a

stable and reliable platform would offer a user experience that has as little lag as possible, and a connection maintained through the examination. Beyond in-the-moment frustrations, issues of technology reliability can undermine the candidates' and public's trust in the examinations and what it represents. More importantly, this may void the assessment score altogether. Choosing a stable and reliable platform provider is of the utmost importance, and as such may require substantial piloting and testing to identify the best platform. When piloting a platform, one should aim to reproduce, as closely as possible, the conditions in which the exam would be administered (including number of potential candidates, number of items, length, etc.). While it may be impossible to reproduce all these conditions, mimicking the expected setting as closely as possible will provide a better sense of the stability and reliability of the platform.

2. Use a platform compatible with all (or most) Operating Systems and that requires minimum hardware, and internet bandwidth, to limit inequities between examinees.

Equity is more and more put at the forefront of assessment considerations,[3,4] especially in the context of e-assessment. Authors have commented, for example, on the negative consequences that the digital divide may have on low-income students.[2,5,6] Considerations for equity are enacted when one aims to "…*identify and remove construct-irrelevant barriers to maximal performance for any examinee. Removing these barriers allows for the comparable and valid interpretation of test scores for all examinees*."[1(p63)] Strategies to reduce inequity include providing candidates with the required technology such as providing computers, webcams or eventually Virtual Reality equipment.[7] Other strategies include using accessible technology (minimum hardware requirement and minimal bandwidth use) and, ensuring compatibility with multiple operating systems, screen sizes or even font size.

3. Create a lot of content (questions or stations) to decrease the potential effect of content sharing between candidates.

Using parallel forms of an exam reduces cheating in the context of multiple days of testing.[2] In addition, the development of bigger item banks minimizes the negative impact on items' psychometric properties associated with the frequent re-use of items.[5,8] In other words, using items less often reduces the risk that they become 'easier' after being shared between candidates. Collaborations between institutions, when possible, or the use of algorithm to generate items[9] are strategies that can potentially reduce the burden of content creation. However, when aiming for more content and more versions of the same exam there is a danger that these different exams may not be comparable. Consequently, strategies or processes ensuring comparable difficulty levels need to be put in place.

4. Leverage the technology to enhance the quality of the assessment

Moving to ICT-based assessment offers several opportunities, such as a purposeful and strategic use of multi-media, including audio and video clips. To be considered Technology Enhanced Assessment, the integration of ICTs to assessment practices should contribute to enhancing the validity of score interpretation -increasing the authenticity of assessment tasks- and the quality of user experience.[10,11] While integrating ICTs to assessment opens a world of possibilities, these changes should be done with considerations for platform stability and accessibility, as discussed previously. In addition, exam designers should consider questions that are long, how much scrolling is needed, and for students who like to take notes to help them organize their thoughts when answering questions, how is notetaking handled.

5. Offer simulation sessions and proper training for candidates, standardized patients, and examiners to familiarize themselves with the platform.

The technology can be a source of anxiety for candidates,[12–14] standardized patients and examiners. Given the usual distress associated with high stakes assessment, it is important to consider how to avoid the technology becoming a burden so big that candidates' true knowledge, skills and attitudes cannot be captured with the assessment. Providing practice opportunities for candidates, examiners and standardized patients to familiarize themselves with the technology can help to reduce the anxiety,[15,16] and favors a smoother administration because everyone involved knows the process and the technology.[17] It seems that just being exposed to the platform reduces the anxiety of all users (examinees, examiners and standardized patients).[16,17] This applies for both written and performance-based assessment. When simulation or practice sessions are not feasible, one could consider podcasts or step-by-step demonstrations to prepare candidates, standardized patients, and examiners.

Considerations during the administration of the assessment

6. Adapt the structure and process

Performance-based assessment may be the type of examination that require the most adaptation when conducted online. In the context of Objective Structured Clinical Examinations (OSCEs) for example, one might consider using a different approach to candidates moving from room to room. Lewandowski et al.[17] and Ryan et al.,[18] for example, favored an approach where examiners moved from candidate to candidate. These movements -from room to room—were facilitated by resource dedicated to the management of examinee and examiner location.

7. Ensure appropriate availability of resources.

While there is not one fail-safe way of doing a virtual OSCE, having sufficient resources available to manage and monitor moves between rooms, and to assist candidates, examiners, and SPs, is one way of contributing to a less eventful administration of the examination.[18] Similarly, one also requires sufficient available resources when conducting written exams on-line to ensure appropriate support and invigilation.

8. Use proper authentication and proctoring protocols while balancing for privacy considerations.

Authentication refers to the processes put in place to ensure that the correct candidate is sitting the exam.[6] This process can take on many forms from requiring a photo ID, some one-on-one questioning, and maybe testing the computer being used.[2] These strategies seem easy enough to implement and seem acceptable for most stakeholders. *Proctoring* refers to the surveillance put in place during the assessment to prevent—as much as possible—cheating behavior. Live remote proctoring, that is having an invigilator observe examinees by using their webcam, is a strategy used to mitigate the potential of cheating.[7] However, live remote proctoring has been criticized for creating additional test anxiety,[19-21] violating personal privacy,[19,20,22] and leading to test taker withdrawal from the assessment.[19] Measures and processes of authentication and proctoring need to be balanced out with issues of privacy. Aligned with principles of equity, fairness and responsible conduct of assessment, institutions implementing remote proctoring practices should be mindful, not only of privacy laws, but of how candidates perceive these strategies. In addition, strategies should be put in place to revisit any performance flagged as potential cheating behavior. Having an examiner doing live scoring can give the perception of an added invigilator, thus reducing opportunities for cheating behavior.

9. Implement fair accommodation strategies.

Additional time to complete an exam is by far the most common accommodation requested and offered.[23,24] Remote assessment could facilitate having settings that are less distracting if conditions put in place by the testing institutions are respected. While these accommodations are easy to implement in remote assessments, other forms of accommodation may be more challenging. Some candidates may require larger prints or read-aloud test directions or questions (which may or may not be possible according to the platform used).[25] Some candidates may require a sign interpreter in the context of performance-based assessments,[25] or additional breaks.[26] Further research is required to understand the consequences on performance of these accommodations in the context of remote assessment.

10. Record performances as a safety net.

Saving candidates' answers as they move along in the process of a written exam is a common practice. Having a recording of candidates' performances offers many possibilities.[27–33] If an examiner is disconnected from the platform, the recording allows for scoring of the performance later thus not penalizing the candidate. In addition, the recording can be used in the case of a candidate contesting their score.[17]

Considerations for detecting cheating behaviors

11. Use advanced statistical modeling to determine the probability of individual cheating.

Advanced statistical modeling, such as Person-Fit Statistics, offer the opportunity to establish the probability of cheating behavior on Multiple Choice Exams.[34,35] These statistics have been tested to detect unusual and unexpected assessor behavior (i.e., leniency and stringency) in a performance-based assessment, such as an OSCE.[36-38] Future research could be undertaken to explore if Person-Fit Statistic have any merit or potential use in detecting cheating behavior in candidates in assessments other than MCQs.

Conclusion

While there is no question that licensure and certification institutions, as well as university programs, need to integrate ICTs in their high-stakes assessment practices and be prepared to conduct reliable and valid remote assessment, during the transition to remote e-licensure assessment there are bound to be some hits and some misses. The question then becomes, "how many fail safe and contingency plans should be put in place?" In addition, if these institutions play their cards well in integrating ICTs to their assessments, this could go beyond "being prepared," but also enhanced the quality of their assessments and validity of the score interpretation.

Conflicts of Interest: The author has no conflict of interest to declare. This work was initially presented as a white paper to the Medical Council of Canada's Assessment Innovation Task Force. The work was conducted within the scope of the *Paul Grand'Maison de la Societé des Médecins de l'Université de Sherbrooke* Research Chair in Medical Education.

References

1. American Educational Research Association., American Psychological Association., National Council on Measurement in Education., Joint Committee on Standards for Educational and Psychological Testing (U.S.). Standards for educational and psychological testing. Washington, DC: American Educational Research Association; 2014.
2. Camara W. Never let a crisis go to waste: large-scale assessment and the response to COVID-19. *Educ Meas Issues Pract*. 2020;39(3):10–18. https://doi.org/10.1111/emip.12358c
3. Chilisa B. Towards equity in assessment: crafting gender-fair assessment. *Assess Educ Princ Policy Pract*. 2000;7(1):61–81. https://doi.org/10.1080/713613318
4. Cumming JJ. Legal and educational perspectives of equity in assessment. *Assess Educ Princ Policy Pract*. 2008;15(2):123–135. https://doi.org/10.1080/09695940802164168
5. Wiley A, Buckendahl CW. Your guess is as good as ours. *Educ Meas Issues Pract*. 2020;39(3):49–52. https://doi.org/10.1111/emip.12366
6. Langenfeld T. Internet-based proctored assessment: security and fairness issues. *Educ Meas Issues Pract*. 2020;39(3):24–27. https://doi.org/10.1111/emip.12359
7. Evans J, Knezevich L. Impacts of COVID-19 on the law school admission test. *Educ Meas Issues Pract*. 2020;39(3):22–23. https://doi.org/10.1111/emip.12367
8. Joncas SX, St-Onge C, Bourque S, Farand P. Re-using questions in classroom-based assessment: an exploratory study at the undergraduate medical education level. *Perspect Med Educ*. 2018;7(6):373–378. https://doi.org/10.1007/s40037-018-0482-1
9. Pugh D, De Champlain A, Gierl M, Lai H, Touchie C. Can automated item generation be used to develop high quality MCQs that assess application of knowledge? *Res Pract Technol Enhanc Learn*. 2020;15(1):1–13. https://doi.org/10.1186/s41039-020-00134-8
10. Joint Information Systems Committee (JISC). *Effective assessment in the digital age* [Internet]. UK: HEFCE; 2010 Available from: https://www.webarchive.org.uk/wayback/archive/20140613220103/http://www.jisc.ac.uk/media/documents/programmes/elearning/digiassass_eada.pdf [Accessed Aug 19, 2021].
11. Sweeney T, West D, Groessler A, et al. Where's the transformation? Unlocking the potential of technology-enhanced assessment. *Teach Learn Inq*. 2017;5(1):1–16.
12. Hewson C. Can online course-based assessment methods be fair and equitable? Relationships between students' preferences and performance within online and offline assessments. *J Comput Assist Learn*. 2012;28(5):488–498. https://doi.org/10.1111/j.1365-2729.2011.00473.x
13. Hewson C, Charlton J, Brosnan M. Comparing online and offline administration of multiple choice question assessments to psychology undergraduates: do assessment modality or computer attitudes influence performance? *Psychol Learn Teach*. 2007;6(1):37–46. https://doi.org/10.2304%2Fplat.2007.6.1.37
14. Hewson C, Charlton JP. An investigation of the validity of course-based online assessment methods: the role of computer-related attitudes and assessment mode preferences. *J Comput Assist Learn*. 2019;35(1):51–60. https://doi.org/10.1111/jcal.12310
15. St-Onge C, Ouellet K, Lakhal S, Dubé T, Marceau M. COVID-19 as the tipping point for integrating e-assessment in higher education practices. *Br J Educ Technol*. 2022;53(2):349-366. https://doi.org/10.1111/bjet.13169
16. Donn J, Scott JA, Binnie V, Bell A. A pilot of a virtual objective structured clinical examination in dental education. A response to COVID-19. *Eur J Dent Educ*. 2021;25(3):488–494. https://doi.org/10.1111/eje.12624
17. Lewandowski R, Stratton A, Gupta TS, Cooper M. Twelve tips for OSCE-style Tele-assessment. *MedEdPublish*. 2020;9. https://doi.org/10.15694/mep.2020.000168.1
18. Ryan A, Carson A, Reid K, Smallwood D, Judd T. Fully online OSCEs: a large cohort case study. *MedEdPublish*. 2020;9. https://doi.org/10.15694/mep.2020.000214.1
19. Karim MN, Kaminsky SE, Behrend TS. Cheating, reactions, and performance in remotely proctored testing: an exploratory experimental study. *J Bus Psychol*. 2014;29(4):555–572. https://doi.org/10.1007/s10869-014-9343-z
20. Lilley M, Meere J, Barker T. Remote live invigilation: a pilot study. *J Interact Media Educ*. 2016;1(6):1–5. http://dx.doi.org/10.5334/jime.408
21. Stowell JR, Bennett D. Effects of online testing on student exam performance and test anxiety. *J Educ Comput Res*. 2010;42(2):161–171. https://doi.org/10.2190%2FEC.42.2.b
22. Weiner JA, Hurtz GM. A comparative study of online remote proctored versus onsite proctored high-stakes exams. *J Appl Test Technol*. 2017;18(1):13–20. Available from: http://jattjournal.net/index.php/atp/article/view/11306 1

23. Lovett BJ. Extended time testing accommodations for students with disabilities: answers to five fundamental questions. *Rev Educ Res*. 2010;80(4):611–38. https://doi.org/10.3102%2F0034654310364063
24. Lovett BJ, Leja AM. ADHD symptoms and benefit from extended time testing accommodations. *J Atten Disord*. 2015;19(2):167–72. https://doi.org/10.1177/1087054713510560
25. Katsiyannis A, Zhang D, Ryan JB, Jones J. High-stakes testing and students with disabilities: Challenges and promises. *J Disabil Policy Stud*. 2007;18(3):160–7. https://doi.org/10.1177%2F10442073070180030401
26. Lin P-Y, Lin Y-C. Examining accommodation effects for equity by overcoming a methodological challenge of sparse data. *Res Dev Disabil*. 2016;51–52:10–22. https://doi.org/10.1016/j.ridd.2015.12.012
27. Bautista JMD, Manalastas REC. Using video recording in evaluating students' clinical skills. *Med Sci Educ*. 2017;27(4):645–650. https://doi.org/10.1007/s40670-017-0446-9
28. Sturpe DA, Huynh D, Haines ST. Scoring Objective structured clinical examinations using video monitors or video recordings. *Am J Pharm Educ*. 2010;74(3):1–5. https://doi.org/10.5688/aj740344
29. Kiehl C, Simmenroth-Nayda A, Goerlich Y, et al. Standardized and quality-assured video-recorded examination in undergraduate education: informed consent prior to surgery. *J Surg Res*. 2014;191(1):64–73. https://doi.org/10.1016/j.jss.2014.01.048
30. Vivekananda-Schmidt P, Lewis M, Coady D, et al. Exploring the use of videotaped objective structured clinical examination in the assessment of joint examination skills of medical students. *Arthritis Care Res*. 2007;57(5):869–876. https://doi.org/10.1002/art.22763
31. Kumar RV. Videotaped OSPE: is this a right procedure to assess health science students' performance?--A pilot study. *Int J Inf Educ Technol*. 2016;6(3):211–214. https://doi.org/10.7763/IJIET.2016.V6.686
32. Driscoll PJ, Paisley AM, Paterson-Brown S. Video assessment of basic surgical trainees' operative skills. *Am J Surg*. 2008;196(2):265–272. https://doi.org/10.1016/j.amjsurg.2007.09.044
33. Nickel F, Hendrie JD, Stock C, et al. Direct observation versus endoscopic video recording-based rating with the objective structured assessment of technical skills for training of laparoscopic cholecystectomy. *Eur Surg Res*. 2016;57(1–2):1–9. https://doi.org/10.1159/000444449
34. Meijer RR, Sijtsma K. Methodology review: evaluating person fit. *Appl Psychol Meas*. 2001;25(2):107–135. https://doi.org/10.1177%2F01466210122031957
35. Wood TJ, St-Onge C, Boulais A-P, Blackmore DE, Maguire TO. Identifying the unauthorized use of examination material. *Eval Health Prof*. 2010;33(1):96–108. https://doi.org/10.1177%2F0163278709356192
36. Iramaneerat C, Yudkowsky R, Myford CM, Downing SM. Quality control of an OSCE using generalizability theory and many-faceted Rasch measurement. *Adv Health Sci Educ Theory Pract*. 2008;13:479–493. https://doi.org/10.1007/s10459-007-9060-8
37. McManus I, Thompson M, Mollon J. Assessment of examiner leniency and stringency ('hawk-dove effect') in the MRCP(UK) clinical examination (PACES) using multi-facet Rasch modelling. *BMC Med Educ*. 2006;6:42. https://doi.org/10.1186/1472-6920-6-42
38. Aubin A-S, St-Onge C, Renaud J-S. Detecting rater bias using a person-fit statistic: a Monte Carlo simulation study. *Perspect Med Educ*. 2018;7(2):83–92. https://doi.org/10.1007/s40037-017-0391-8