Tertiary Education Assessment Amidst COVID-19: Adjustments to Maintain Score Validity

Lyndon Lim
Singapore University of Social Sciences

Written with a tertiary educator in mind, this short paper aims to examine possibilities to maintain assessment score validity in light of the impact of COVID-19 on educational assessment. In doing so, this paper seeks to: (1) support some suggestions mooted by scholars or implemented by some educational systems, and (2) discuss how these may be applicable within tertiary education. The possibilities and a key threat to score validity (i.e., contract cheating) are discussed with respect to three of five facets of validity evidence (i.e., test content, internal structure and consequences of testing). The paper concludes by offering some views tertiary educators could consider when adjusting assessment to uphold score validity.

Keywords: educational assessment, higher education, score validity

INTRODUCTION

Understandably, as the virus that causes COVID-19 is location agnostic and unbiased toward any ethnic group, much has been discussed about the impacts, some perpetual, brought about or accelerated by the pandemic. Scholars from various domains such as educational assessment and measurement, and general/tertiary/continuing education have contributed to scholarship in areas related to how assessment might be re-imagined, and how certifications and qualifications could remain fit for interpretation in light of COVID-19 impacts.

Written with a practitioner (e.g., instructor, course assessment writer) on a broad spectrum (i.e., spanning from those who know about educational assessment thoroughly, to those who know less about educational assessment such as those who view it simply as testing) as an audience, this short paper seeks to: (1) support some of the suggestions mooted by scholars or implemented by some educational systems, and (2) discuss how these may be applicable within tertiary education. In doing so, the feasibility involving these actions are highlighted, anchored upon the validity of assessment scores.

KEY DRIVER OF ADJUSTMENT – VALIDITY OF SCORES

Given the sweeping reactions across the globe toward assessment including high stakes standardised assessment, it is of little doubt that score validity, a principle of assessment, remains a key driver of the changes seen since COVID-19 struck whether in general, tertiary or further/continuing education. Suffice to suggest, these changes have not been driven primarily by assessment innovations, meant to balance assessment security (e.g., online proctoring) or to deliver assessment via other modes (e.g., computer-delivered assessment), as some might suggest. Rather, assessment innovations and corresponding implementations could have been realised due to COVID-19 because assessment scores had to remain valid

and fit-for-purpose; some assessment innovations that had been actualised but implemented with a perfunctory attitude before COVID-19 struck would also be taken more seriously now.

Based on the unified concept of validity posited by Messick (1993), evidence supporting the validity of assessment scores could be gathered based on five facets: (1) test content (2) response processes (3) internal structure (4) relation to other variables and, (5) consequences of testing. Varying degrees of evidence sought from these facets would then render an assessment score interpretable and useful according to intended purposes. The following sections discuss possible assessment adjustments as a response to the impact of COVID-19 within tertiary education with respect to three of five of these facets.

Test Content – A Reduction?

Analogous to content appropriateness, test content includes the substance, wording, and format of the items in a test (American Educational Research Association, American Psychological Association & National Council on Measurement in Education [AERA, APA & NCME], 2014). Notwithstanding changes to assessment accelerated due to COVID-19, assessment scores are intended to reflect what students have been cued to do or learnt as part of the curriculum, so that claims about learning can be made about them. Assessment tasks should also follow a format that students would be familiar with. For example, a graded group presentation should remain as that as far as practicable, with the use of virtual meeting platforms; it should preferably not be transitioned into an individual paper-based online test as the intended assessment objectives of both assessment formats would be contrasting.

Time might be a valid concern owing to COVID-19 impacts when considering test content. In a situation where institutions are required to be physically closed by authorities making it less practicable to complete or enact parts of the intended curriculum (e.g., virtual instead of physical laboratory sessions), test content within assessment tasks designed to elicit evidence of learning should exclude topics taught toward the end of the curriculum, in part to be fair to students. Arguably, excluding test content is technically not a reduction. Rather, it is a reflection of maintaining validity evidence based on test content. If test content may not be excluded for some reason, facets of an assessment that is judged as less relevant for students in their future workplace could be considered for deletion.

Common Last Topics

By the same token, adjustments to test content as a response to COVID-19 should be taken by and communicated to stakeholders as an adjustment to maintain assessment score validity rather than a mere reduction. This has shown to be possible even at the national level. For example, common last topics (CLT) were removed from national high stakes examinations in Singapore due to a circuit breaker period in 2020, when schools in the country were closed and movements restricted (Ministry of Education & Singapore Examinations and Assessment Board, 2020). It is noteworthy that there had always been a deliberate effort to have CLT, so that national examinations could presumably be adjusted to maintain score validity owing to unforeseen circumstances such as COVID-19. Hence, it appears that rather than having to reduce or adjust test content as a response to impacts brought about by COVID-19 or the like, curriculum writers and course developers within tertiary education could consider having CLT that would be taught but non-examinable in times of exigencies. Such planning would likely still support validity evidence due to test content and hence, assessment score validity. As a practical suggestion, CLT could include topics that require students to demonstrate learning via complex assessment tasks (e.g., authentic performance tasks) as suggested by Koh (2017), as opposed to standardised tests.

Internal Structure – Reviewing and Refining

The internal structure of an assessment refers to the conceptual framework that indicates relationships among test items, and how this framework conforms to the construct intended to be assessed (AERA, APA & NCME, 2014). For instructors within educational institutions, this conceptual framework can be represented by utilising established learning taxonomies such as Bloom's Taxonomy (Krathwohl, 2002) within an assessment table of specifications as suggested by Leong (2018).

Adjusting Weights

A possibly less onerous approach of adjusting assessment in light of COVID-19 within tertiary educational institutions could be to adjust the weighting of each component within the table of specifications. For example, instead of a weighting such as written assignment (40%), presentation (30%), online discussion board (20%) and quiz (10%), course convenors could consider adjusting the weighting as written assignment (20%), online presentation (40%), online discussion board (30%) and online quiz (10%). The adjusted weighting can help students reframe their minds and focus on the presentation which can be based on a performance task. Further, this could discourage those who might be considering contract cheating services for their written assignments.

Reframing Assessment Focus

The advent and push by the United Nations Educational, Scientific and Cultural Organization to strengthen digital competencies and enhance computational thinking in education (UNESCO, 2018) have also provided directions and opportunities to review curriculum and corresponding assessments. COVID-19 and contract cheating seem to have been accelerants toward this cause, particularly with regard to the review of assessment tasks based on their corresponding internal structure so that assessment tables of specifications would remain fit-for-purpose. Such a review would better allow students to demonstrate higher order thinking and learning that reflect those required within workplaces; this also minimises the threat of contract cheating. For example, assessment reviews motivated by enhancements to include computational thinking or higher order thinking skills that were previously conducted perfunctorily could now be more thorough; this could include shifting an assessment focus from lower to higher levels of Bloom's Taxonomy since assessment tasks tied to the lower levels of Bloom's could be integrated within one or more larger tasks that reflect real world applications and allow students to demonstrate learning with skills required of them in the real world.

Further, professional discussions amongst instructors or assessment writers could be re-ignited, to get them thinking about re-aligning assessment tasks anchored on established learning taxonomies integrated with other considerations when reviewing assessment table of specifications as a response to COVID-19 impacts. One such consideration is that by Lemons and Lemons (2017), who found that in test design, test developers often considered unstructured categories that were viewed separate from the Bloom's Taxonomy. They summarised these unstructured categories into: (1) item difficulty – whether the item is challenging for students, (2) time required – the amount of time a student takes to provide a response to the item, (3) student experience – how familiar students are in addressing and responding to the item and, (4) correct answers – whether the item has more than one correct answer. While these four categories might be viewed separate from the Bloom's Taxonomy, considering them would almost logically fit within various levels of the taxonomy. As an example, an item, say, a performance assessment task, would likely borrow verbs from higher levels of Bloom's and naturally, would be seen as a task that requires more time and have higher cognitive demands on the student.

Adjusting Assessment Duration and Item Types

Another example of adjustments to an assessment internal structure as a response to the COVID-19 pandemic include the May 2020 College Board's advanced placement examination. In the revised structure, selected response items were removed such that test-takers would see only constructed response items. Test duration was also significantly reduced from up to three hours to about 45 minutes, and test delivery transitioned from a face-to-face to a virtual setting that would be administered at the same time worldwide (College Board, 2020a). Some test-takers naturally took issue with these changes. For example, the selected response items some had been practicing would now be rendered inadequate in terms of earning college credits, noting that selected response items normally constituted more than half of the examinations. Further, some experienced technical glitches that presumably increased their test anxiety (Bavis, 2021; Wan, 2020). Nonetheless, the changes in assessment structure appear appropriate when manageability is considered. It is logistically demanding to administer face-to-face examinations during the pandemic, and online testing as an alternative would be physically demanding on test-takers if the test duration remained the same.

Of course, there had been more drastic changes to assessment other than revising the internal structure. Cambridge Assessment cancelled all secondary education high stakes assessment in 2020 as a response to COVID-19. In place of the actual assessment grades, an algorithm that considered student prior subject attainment, rank orders estimated by teachers and trends of school performances was used to predict grades for the high stakes assessments. Nonetheless, the predicted grades were not well-received by the public due to perceived unfairness in how the algorithm worked (Coughlan, Sellgren & Burns, 2020; Stewart, 2020).

Consequences of Testing… or Not?

Consequences of testing or assessment may be intended or unintended, and may not be limited to the assessment or corresponding scores. Whether intended or unintended, consequences can be categorised into positive or negative backwash effects; a negative backwash effect may not be all negative as it might lead to a positive backwash effect, as can be seen in the following discussions.

Positive Backwash Effects

Examples of positive backwash effects arising from the impact of COVID-19 include enabling students to acquaint and be more familiarised with computing devices that afford opportunities in learning. This is a step towards attaining relevant digital literacy competencies laid out by UNESCO in their Digital Literacy Global Framework (UNESCO, 2018). Further, with student familiarity with computing devices and COVID-19 as an accelerant, educators can be more motivated to engage in professional discussions, revise the internal structure of assessments and develop assessment tasks that better reflect real world applications as opposed to the more traditional paper-based modes. Some sectors have also benefited in terms of revenue from the consequences of changes of assessment due to COVID-19. For example, test waivers or the loosening of standardised testing requirements as part of programme applications due to COVID-19 led to a surge in master of business administration postgraduate programmes after years of declines (Thomas, 2020).

COVID-19 also accentuated views the public, including potential test-takers held about assessment. For example, the College Board reported that in a survey with 18000 advanced placement students, most did not want the examinations cancelled (College Board, 2020b). This indicated the value proposition of assessment or at least, how important test-takers viewed examinations as a fair means to determine achievement. This view also often reflects those by assessment regulatory agencies such as the Office of Qualifications and Examinations Regulation (Ofqual) in the United Kingdom that highlighted that exams are the fairest way to assess what students know and can do (Ofqual, 2020). Where high stakes examinations were cancelled (e.g., the 2020 high stakes assessments by Cambridge Assessment), a couple of positive backwash effects were also observed. First, alternative grading means were considered; this inadvertently involved multiple stakeholders in education (e.g., instructors, academic board in institutions) and this presented an opportunity to review how students had been graded all this while. Second, while the public outcry and dissatisfaction that arose from the grades that were awarded to students by means of an algorithm led to the resignation of key personnel within the United Kingdom Department of Education and the Ofqual (Stewart, 2020), the controversy again clearly demonstrated the depth of concern the public had about high stakes assessments and grading.

Negative Backwash Effects

Negative backwash effects include an increase in contract cheating cases and the proliferation of contract cheating profit-making entities though these have also been accompanied by positive backwash effects. It is noteworthy that contract cheating, initially defined by Clarke and Lancaster (2006) as the process where students engage others to produce an original work which they then submit, often for a payment of a fee, goes beyond verbatim copying, ghost writing, plagiarism or collusion, and is considered academic fraud (Bretag, 2017). In his meta-analysis of contract cheating in higher education, Newton (2018) concluded that contract cheating had been increasing based on data from 2014 to 2018. Contract cheating was also observed to be in a declining trend from 2004 to 2014 based on studies in Australia (Curtis & Tremayne, 2019; Curtis & Vardanega, 2016). However, this declining trend was not observed in 2019 and, with the COVID-19 impact, contract cheating might well have stopped declining. In fact,

Lancaster and Cotarlan (2021) found that contract cheating did increase when they compared the use of a file sharing site "Chegg" during the April to August 2019 period with the April to August 2020 period, a period when COVID-19 impacted and many courses and assessments were delivered online.

The positive backwash effects brought about by contracting cheating, which ironically was a negative backwash due to COVID-19, include increased educator awareness of contract cheating and hence, motivated assessment writers to re-imagine assessment tasks to uphold academic integrity and minimise cheating. These cement the need to re-imagine assessment and assessment design particularly within tertiary education, for which contract cheating services would be more prevalent.

While COVID-19 presented an opportunity for re-imagining assessment, some educational systems have continued to emphasise the importance of fairness and not cheating. For example, administrators of college entrance examinations in China, better known as *gaokao*, have employed the use of artificial intelligence to detect cheating (Zhang, 2020). Undeniably, re-imagining assessment at this scale and the stakes involved would have posed challenges and might have increased test anxiety unnecessarily; just for 2021, 10.78 million students registered for the *gaokao* ("China holds college entrance exam with tailored Covid-19 measures in place", 2021).

Assessment Security or Academic Integrity?

Coupled with the re-imagination of assessment would be that of academic integrity – how might tertiary educators better instil academic integrity in the minds of students? Dawson (2021) suggested that assessment design is a cut above assessment security and undeniably, educators would need to seek the fine balance between academic integrity that is more values-based, with assessment security that is more adversarial. In terms of assessment design, assessment writers could, in addition to the suggestions discussed (e.g., setting assessment tasks that reflect real world applications): (1) include examples or cases that had been discussed during the course – the primary intent of this would be to have the student relate what s/he had learnt to what was being assessed, as opposed to deterring contract cheating (since the commercial entities might already not have access to the course lessons), (2) include within the assessment task a significant portion of how the course relates to a students' workplace – this might be more applicable for part-time students who work at the same time, so that the students would be better able to appreciate and link what had been taught to an application in the workplace, or (3) include viva voce as part of or a replacement to a written assignment – where viva is used as part of an assessment, it can also be viewed as a confirmation of what a student has learnt.

CONCLUSION

Possible adjustments to assessment including actual adjustments by educational systems were discussed anchored upon three facets of validity evidence (i.e., test content, internal structure and consequences of testing), in view of how COVID-19 had impacted educational assessment. It is evident that, to maintain score validity, educators or policy makers could: (1) postpone an assessment, (2) cancel an assessment or, (3) make adjustments so that assessment can still proceed. Some adjustments such as reframing thinking toward educational assessment in tertiary education, shifting the assessment focus from lower to higher levels of the Bloom's Taxonomy, and re-designing assessments to include tasks that reflect real world applications were highlighted. Unquestionably, these adjustments, for example, shifting the assessment focus to higher levels of the Bloom's Taxonomy should not lead to above-level testing as this would be detrimental to assessment score validity. Assessment tasks should remain aligned to the intended course learning outcomes regardless of any adjustments to mitigate the negative impacts brought about by COVID-19 on score validity.

In the adjustments discussed and to uphold assessment score validity, it would be judicious to appreciate when students would be more likely to engage contract cheating services. In their significant research work with 14086 tertiary students across eight Australian universities, Bretag et al. (2019) found that the proportion of students who perceived that contract cheating was more likely was higher for assessment tasks with short turnaround time, heavily weighted tasks, and series of small graded tasks. Conversely, this proportion was much lower for assessments involving viva, reflection on practicum, personalised and

unique tasks, and in-class tasks. Further, Bretag and colleagues found that over 70% of educators surveyed in the same study reported at least moderate use of assessment that appears relevant to higher levels of the Bloom's Taxonomy (e.g., integrates knowledge and skills, develop research, analysis and critical thinking skills), though relatively fewer educators used assessments that students would be less likely to engage in contract cheating (e.g., viva).

Appreciating the works of Bretag and colleagues provide initial directions to how educators might make adjustments to assessments such that assessment scores remain valid in the insidious situation of contract cheating, particularly in light of COVID-19. Faculty development and support should follow, so that assessment score comparability and hence validity remains, for adjustments should preferably not lead to a drastic change in assessment scores when compared with previous cohorts'. In supporting faculty whether through professional development sessions or publications/bulletins to motivate assessment adjustments in the face of COVID-19, some of the following critical questions should be addressed: (1) how might assessments be adjusted so that score validity remains (2) how manageable would the adjustments be, for students, professional staff administering the assessment and the faculty in-charge (3) should the assessment be postponed (4) is there a robust grading method in the event the assessment is cancelled.

COVID-19 has been a catalyst to accelerate assessment innovations or adjustments, indisputably to defend score validity. All the discussions and adjustments seen thus far point to how concerned the public and educators are with regard to learning and assessment, and that the conferral of a certification must be anchored upon valid scores, complemented with multiple sources of evidence, as there would be an impact on the economy at large, say, if graduates gain employment based on a certificate that was obtained due to contract cheating.

REFERENCES

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. American Educational Research Association.

Bavis, P. (2021, March 10). AP exams can't be business as usual this year. *Education Week*. Retrieved from https://www.edweek.org/teaching-learning/opinion-ap-exams-cant-be-business-as-usual-this-year/2021/03

Bretag, T. (2017). Special collection: The rise of contract cheating in higher education – academic fraud beyond plagiarism. *International Journal for Educational Integrity*. Retrieved from https://www.biomedcentral.com/collections/cche

Bretag, T., Harper, R., Burton, M., Ellis, C., Newton, P., van Haeringen, K., . . . Rozenberg, P. (2019). Contract cheating and assessment design: Exploring the relationship. *Assessment & Evaluation in Higher Education*, *44*(5), 676–691. DOI: 10.1080/02602938.2018.1527892

Clarke, R., & Lancaster, T. (2006). Eliminating the successor to plagiarism? Identifying the usage of contract cheating sites. In *Proceedings of 2nd plagiarism: Prevention, Practice and Policy Conference 2006*. JISC Plagiarism Advisory Service.

College Board. (2020a). *2020 AP testing guide*. Retrieved July 15, 2021, from https://apcentral.collegeboard.org/pdf/ap-testing-guide-2020.pdf

College Board. (2020b). *Taking the AP exams*. Retrieved July 15, 2021, from https://apcoronavirusupdates.collegeboard.org/educators/taking-the-exams

Coughlan, S., Sellgren, K., & Burns, J. (2020, August 13). A-levels: Anger over 'unfair' results this year. *BBC*. Retrieved from https://www.bbc.com/news/education-53759832

Curtis, G.J., & Tremayne, K. (2019). Is plagiarism really on the rise? Results from four 5-yearly surveys. *Studies in Higher Education*. DOI: 10.1080/03075079.2019.1707792

Curtis, G.J., & Vardanega, L. (2016). Is plagiarism changing over time? A 10-year time-lag study with three points of measurement. *Higher Education Research & Development*, *35*(6), 1167–1179. DOI: 10.1080/07294360.2016.1161602

Dawson, P. (2021). *Defending assessment security in a digital world: Preventing e-cheating and supporting academic integrity in higher education*. Taylor & Francis. DOI:10.4324/9780429324178.

Koh, K.H. (2017). Authentic assessment. *Oxford Research Encyclopedia of Education*. Retrieved from https://oxfordre.com/education/view/10.1093/acrefore/9780190264093.001.0001/acrefore-9780190264093-e-22

Krathwohl, D.R. (2002). A revision of Bloom's Taxonomy: An overview. *Theory Into Practice*, *41*(4), 212–218. https://doi.org/10.1207/s15430421tip4104_2

Lancaster, T., & Cotarlan, C. (2021). Contract cheating by STEM students through a file sharing website: A Covid-19 pandemic perspective. *International Journal for Educational Integrity*, *17*(3). https://doi.org/10.1007/s40979-021-00070-0

Lemons, P.P., & Lemons, J.D. (2013). Questions for assessing higher-order cognitive skills: It's not just Bloom's. *CBE – Life Sciences Education*, *12*(1), 47–58. https://doi.org/10.1187/cbe.12-03-0024

Leong, S.C. (2018). Principles of assessment: Validity, reliability and fairness. In P.G. Toh & S.C. Leong. (Eds.), *Assessment in Singapore Volume 3: Concepts and tools for classroom assessment* (pp. 48–58). Singapore Examinations and Assessment Board.

McCartney, M. (2013). Testing at a higher level. *Science*, *339*(6126), 1361. DOI:10.1126/science.339.6126.1361-c

Messick, S. (1993). *Foundations of validity: Meaning and consequences in psychological assessment*. ETS Research Report No. RR-93-51, Series 2.

Ministry of Education and Singapore Examinations and Assessment Board. (2020, April 21). *Mid-year holidays brought forward as schools adjust academic calendar; institutes of higher learning to extend home-based learning* [Press release]. Retrieved from https://www.seab.gov.sg/docs/default-source/publiccommunications/press-releases/2020/final-moe-seab-joint-press-release.pdf

Newton, P.M. (2018). How common is commercial contract cheating in higher education and is it increasing? A systematic review. *Frontiers in Education*, *3*(67), 1–18. https://doi.org/10.3389/feduc.2018.00067

Office of Qualifications and Examinations Regulation. (2020). *Ofqual welcomes DfE announcement on 2021 exams* [Press release]. Retrieved from https://www.gov.uk/government/news/ofqual-welcomes-dfe-announcement-on-2021-exams

Stewart, H. (2020, August 26). Boris Johnson blames 'mutant algorithm' for exams fiasco. *The Guardian*. Retrieved from https://www.theguardian.com/politics/2020/aug/26/boris-johnsonblames-mutant-algorithm-for-exams-fiasco

The Straits Times. (2021, June 7). *China holds college entrance exam with tailored Covid-19 measures in place*. Retrieved from https://www.straitstimes.com/asia/east-asia/china-holds-college-entrance-exam-with-tailored-covid-19-measures-in-place

Thomas, P. (2020, September 29). Applicants flock to elite business schools to ride out the coronavirus pandemic. *The Wall Street Journal*. Retrieved from https://www.wsj.com/articles/applicants-flock-to-elite-business-schools-to-ride-out-the-coronavirus-pandemic-11601409456

United Nations Educational, Scientific, and Cultural Organization. (2018, June). *A global framework of reference on digital literacy skills for indicator 4.4.2* (Information Paper UIS/2018/ICT/IP/51). Retrieved from http://uis.unesco.org/sites/default/files/documents/ip51-global-framework-reference-digital-literacy-skills-2018-en.pdf

Wan, T. (2020, May 14). Online AP testing glitches force some students to retake exam. *EdSurge*. Retrieved from https://www.edsurge.com/news/2020-05-14-online-ap-testing-glitches-force-some-students-to-retake-exam

Zhang, Z. (2021, July 7). No cheating in gaokao, AI is watching: China Daily columnist. *The Straits Times*. Retrieved from https://www.straitstimes.com/asia/no-cheating-in-gaokao-ai-is-watching-china-daily-columnist