The Canadian Journal for the Scholarship of Teaching and Learning

Volume 15 | Issue 2 Article 4

Summer 08-31-2024

How Do Course-Based Assessments Change in The Shift to Emergency Remote Teaching? Sustainable Assessment Strategies Through an Authenticity Lens

Justine Hobbins
University of Guelph, jhobbins@uoguelph.ca
Emilie Houston
McGill University, emilie.houston@mail.mcgill.ca
Kerry Ritchie
University of Guelph, ritchiek@uoguelph.ca

https://doi.org/10.5206/cjsotlrcacea.2024.2.15122

How Do Course-Based Assessments Change in The Shift to Emergency Remote Teaching? Sustainable Assessment Strategies Through an Authenticity Lens

Abstract
The shift from face-to-face (F2F) to emergency remote teaching (ERT) in response to COVID-19 has presented concerns for assessment in student learning. This study presents the comparison of a health science curriculum in F2F and ERT settings regarding assessments (count, type, authenticity) using our Authentic Assessment Tool and institutionally standardized course syllabi. Five hundred and seventeen assessments in 61 courses in ERT were inventoried (count, type) and subsequently categorized as 1 (low), 2 (moderate), or 3 (high) on core authenticity characteristics: realism, cognitive challenge, evaluative judgement criteria and feedback. These data were compared to a recent curriculum-wide F2F scan (457 assessments in 62 courses). Results show in the shift to ERT, the total number of both tests and assignments increased with a greater proportion of marks comprised of assignments (44% ERT versus 37% F2F). Curriculum-wide authenticity scores were similar (1.8 ± 0.4 ERT versus 1.8 ± 0.6 F2F), although this trend was because nearly an equal proportion of courses increased and decreased authenticity. The largest number of courses (*n*=30) making improvements on individual characteristics of authenticity did so regarding the dimension feedback. This work presents modest yet actionable items to achieve authenticity for consideration in assessment design as institutions begin to produce and consider policies regarding course structure and assessment design in the post-COVID educational context.

Le passage de l'enseignement en face-à-face (FàF) à l'enseignement à distance en situation d'urgence (EDSU), en réponse à la COVID-19, a suscité des inquiétudes quant à l'évaluation de l'apprentissage des étudiants et des étudiantes. Cette étude présente la comparaison d'un programme d'études en sciences de la santé dans des contextes d'enseignement en FàF et d'EDSU en ce qui concerne les évaluations (nombre, type, authenticité) en utilisant notre outil d'évaluation authentique et des descriptions de cours normalisées par l'établissement. Cinq cent dix-sept évaluations dans 61 cours enseignés à distance en situation d'urgence ont été inventoriées (nombre, type) et ensuite catégorisées comme 1 (faible), 2 (modéré), ou 3 (élevé) sur les caractéristiques principales d'authenticité : réalisme, défi cognitif, critères de jugement évaluatif et retour d'information. Ces données ont été comparées à une récente analyse d'enseignement en FàF à l'échelle du programme d'études (457 évaluations dans 62 cours). Les résultats montrent que lors du passage à l'EDSU, le nombre total de tests et de devoirs a augmenté, avec une plus grande proportion de devoirs dans les notes (44 % EDSU contre 37 % FàF). Les scores d'authenticité à l'échelle du programme étaient similaires (1,8 ± 0,4 EDSU contre 1,8 ± 0,6 FàF), bien que cette tendance soit due au fait qu'une proportion presque égale de cours a augmenté et diminué l'authenticité. Le plus grand nombre de cours (*n*=30) ayant amélioré les caractéristiques individuelles de l'authenticité l'ont fait en ce qui concerne la dimension du retour d'information. Ce travail présente des éléments modestes mais réalisables pour atteindre l'authenticité et les prendre en compte dans la conception de l'évaluation, alors que les établissements commencent à produire et à envisager des politiques concernant la structure des cours et la conception de l'évaluation dans le contexte de l'enseignement post-COVID.

Keywords
authentic assessment, emergency remote teaching, face-to-face, feedback; évaluation authentique, enseignement à distance en situation d'urgence, face à face, retour d'information

This research paper/Rapport de recherche is available in The Canadian Journal for the Scholarship of Teaching and Learning: https://doi.org/10.5206/cjsotlrcacea.2024.2.15122

In March 2020, COVID-19 disrupted the conventional face-to-face (F2F) learning environment that many higher education institutions had grown accustomed to. Instructors were faced with a universal catalyst for change, and typical teaching and assessment methods in higher education had to shift to emergency remote teaching (ERT) and learning as a result. Distinct from traditional online learning (Branch & Dousay, 2015), ERT is "a temporary shift to an alternative delivery mode due to crisis circumstances" (Hodges et al., 2020). ERT involves fully remote instructional methods that would otherwise be delivered F2F (or blended, hybrid), and would likely return to some variation of this format once the crisis subsides. However, as the pandemic has been ongoing for the duration of two complete academic years now (winter 2020–present), it is important to recognize that some elements of remote teaching and learning have the potential to stay as we transition to a new normal. It is therefore important to understand and consider how curriculums have changed in the shift from F2F to ERT, which changes should remain (i.e., were beneficial), and which changes were 'band-aid' (i.e., quick fix, unsustainable) solutions. Notably, ERT has been advantageous for some in terms of flexibility, accessibility, and creativity, but there has been an added challenge with respect to academic integrity.

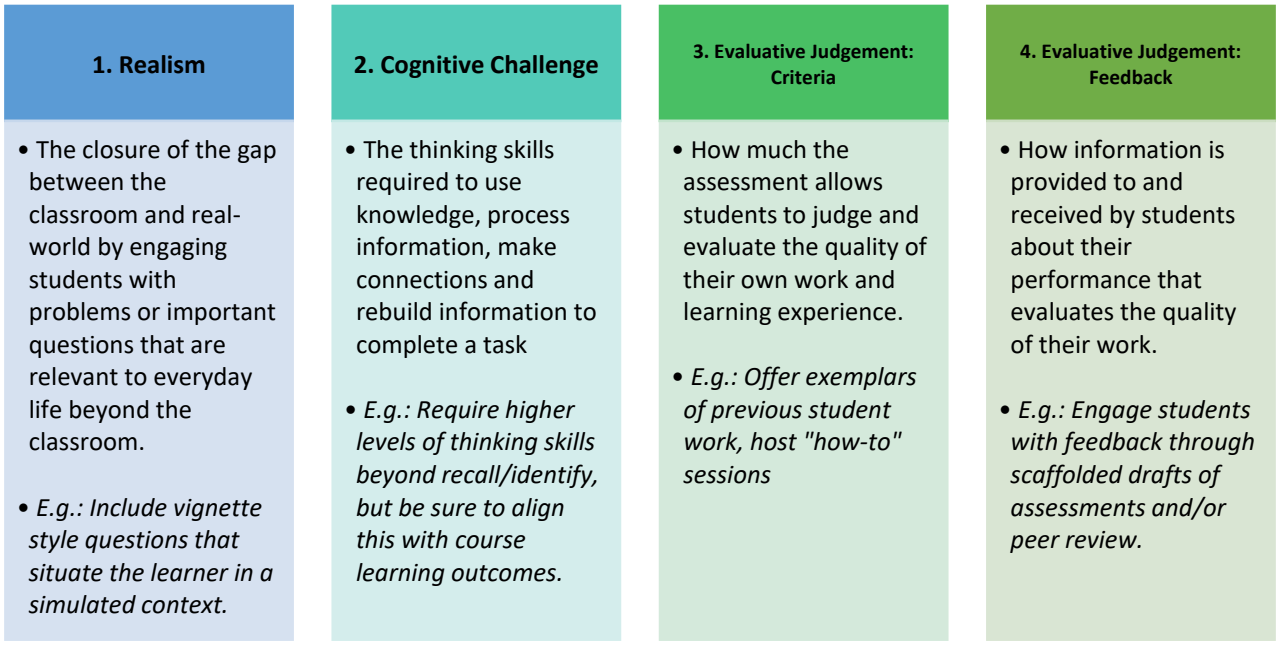
While very early discussions focused on the delivery of content and instructional method logistics in ERT, focus quickly turned to assessment strategies—either to modify and adapt existing assessments, or rethink structure and design of assessments all together (Slade et al., 2021). Bearman et al. (2017) propose a framework for the assessment design process, derived from qualitative analysis of interviews with educators regarding how they design assessments. Notably, a common impetus for change can initiate the (re)design of assessments. Impetus for change can be subject to environmental influences (e.g., class size, mode of delivery) and professional influences (e.g., sharing of best practices through professional development avenues and training) (Bearman et al., 2017). We propose that the COVID-19 disruption is a widespread impetus for change throughout higher education. Specifically, we were curious to see if the best practices that were being widely shared were adopted by instructors, and if these adoptions varied by environmental influences of class size and year level. As we consider new education policies and opportunities post-COVID, it becomes pertinent to systematically review what changes to assessments were made, and if these align with good assessment practices (suggesting an improved assessment structure) or if they had a detrimental impact on assessment practices.

An authenticity lens may be an important angle to consider assessment changes since authentic assessments are commonly prioritized by institutions (MCU & University of Guelph 2017; Sotiriadou et al., 2020). Authentic assessment, across all disciplines, can refer to a formally evaluated assessment activity which engages students with problems or important questions that are relevant to everyday life beyond the classroom; prompt students to use higher levels of thinking to extend knowledge and thinking, while also providing an opportunity to enhance self-regulated learning by engaging with grading criteria and providing and receiving feedback (Hobbins et al., 2021). Through informal discussions with colleagues (e.g., blogs, casual chats, YouTube learning) and more formal professional development avenues (e.g., recommendations from education developers at institutional teaching and learning centres), (Office of Teaching and Learning, n.d.), many recommendations that are well-aligned with characteristics of authentic assessment as described by (Villarroel et al., 2018) (Figure 1) have been circulated for higher education educators to consider in the implementation of their ERT courses. Some of the (anecdotally) commonly shared ideas by institutions broadly and instructors alike include assessments of alternative formats (e.g., open-book, "take-home"), comprised of questions that are contextualized beyond the classroom to increase their relevance (i.e., realism) ("Exploring Alternative Assessment Types"

n.d.). Further, increasing the level of thinking skills required by assessments was a proposed solution to protect academic integrity through the creation of "non-googleable" test questions (i.e., cognitive challenge). To relieve stress and uncertainty, clear communication of expectations and grading criteria between instructors and students was encouraged (i.e., evaluative judgement—criteria). Lastly, frequent-low stakes assessments allow for the provision of multiple points of feedback to students to gauge their progress in a course (i.e., evaluative judgement – feedback) ("Adapting Your Assessments," n.d.). As instructors rethink and revise their courses in accordance with widely shared recommendations of varied scope and scale for assessment design, it is to be expected that a variety of changes to assessments occurred.

Figure 1

The Four Characteristics of Authentic Assessment



Note. As described by Villaroel et al. (2018), with examples of how to execute each characteristic in practice (i.e., in a higher education classroom).

First reports of how instructors changed their assessments in the initial emergency transition have begun to emerge with the ongoing pandemic. Jankowski (2020) found 97% of survey respondents (i.e., instructors) made changes to assessments in response to COVID-19, including modifying assessment deadlines and shifting previously graded assessments to be pass/fail. Johnson, Veletsianos and Seaman (2020) found the majority of faculty reported making changes to their assessments in response to COVID-19, including dropping assignments or exams, and/or shifting to a pass/fail model for the semester. Further, Bartolic et al. (2022) surveyed and interviewed instructors responsible for a sample of courses, finding that the most frequent changes to assessments were with respect to assignments, with approximately half of sampled courses making changes to grading standards and weighting (Bartolic et al., 2022).

After two complete academic years online, institutions are beginning to develop policies and direct resources to guide instructors and communicate to students the expectations around teaching and learning (assessment in particular). We argue that before doing so, a complete curriculum-wide understanding (i.e., beyond a sample of courses) of how a program has transitioned from F2F to ERT would be helpful. To this end, we sought to investigate the following research questions (RQ):

- How did assessments change across a representative Canadian Bachelor of Science (BSc) program in the shift from F2F to ERT (i.e., assessment inventory)?
- Did these changes align with characteristics of authentic assessment which are prioritized by many institutions?

Ultimately, the goal of this work was to systematically report assessment changes through an authenticity lens to inform both individual assessment strategies at the instructor level, as well as decision making, policies and resource allocation at the program level.

Method

Study Context

This study takes place at a Canadian research-intensive, comprehensive university. Decisions for the fall (September–December) 2020 semester to be delivered by ERT format were confirmed in May 2020, therefore instructors were provided time to plan their teaching and assessment methods accordingly. Instructors offering the courses in this ERT health science curriculum were largely tenure-track or tenured faculty. All classes at the institution of study were offered remotely. These varied by synchronous (real-time) or asynchronous classes and seminars, with seminars and labs being entirely remote. This study considered the core foundational courses (e.g., biology, chemistry, math) and the restricted specialized electives (e.g., ergonomics, toxicology, pathology) that comprise the Bachelor of Science (BSc) health science curriculum, which would be considered representative of other undergraduate BSc programs.

RQ 1: ERT Assessment Inventory

We collated a list of all September 2019-April 2020 (F2F) and all September 2020-April 2021 (ERT) undergraduate health sciences courses (core and restricted elective). We then collected institutionally standardized and freely available course outlines (syllabi) for each course in the F2F and ERT settings. Course characteristics (e.g., year level, class size) and assessment characteristics (e.g., count, weighting, type) were recorded. We then compared F2F and ERT inventory data to each other to identify any changes to assessment structure (count, weighting, type).

RQ 2: ERT Assessment Authenticity

Recently, we developed the Authentic Assessment Tool (AAT) according to the core characteristics of authenticity (i.e., realism, cognitive challenge, evaluative judgement – criteria and feedback) (Villarroel et al., 2018), which allows for scoring of individual assessments across the authenticity spectrum. Each assessment identified in the F2F and ERT inventory was scored on the AAT using information available in course outlines (e.g., learning outcomes, assessment

descriptions) by two independent evaluators. Each assessment was assigned a numerical score of 1 (low), 2 (moderate) or 3 (high) for each of the core authentic assessment characteristics. The scores across these characteristics for a given assessment were then averaged to provide an authenticity score out of three. The weighting of individual assessments (i.e., a proxy for assessment size) was considered by multiplying an assessment's score out of three by its weighting (out of 100). This calculation was repeated for all assessments in a course. These assessment-level scores were summed to produce a course-level authenticity score. As such, a weekly discussion post worth 2% would contribute much less to the overall authenticity score of a course compared to a final exam weighted at 50%. The same process was repeated for to produce the authenticity scores for individual characteristics, which allowed for comparison between different characteristics of authenticity.

To clarify and confirm information collected from course outlines about assessments with respect to inventory and authenticity investigations, we met one-on-one with instructors for an open discussion with a series of prompts and short questions (e.g., please confirm your assessment structure; how are guidelines and expectations communicated to students for the take-home test?) (see the Appendix). In these discussions, we also prompted instructors to comment on any factors they may have considered in assessment design in response to the pandemic. Inventory and authenticity data were considered by year of study (first – fourth), class size (small, medium, large) as defined by Cash et al. (2017) and assessment type (tests, assignments).

Statistics

Descriptive statistics were performed for the assessment inventory in the F2F and ERT setting (Table 1). Descriptive statistics for the changes in assessments from F2F to ERT were determined (Table 2). A paired t-test was performed to compare the F2F and ERT curriculums regarding overall average authentic assessment scores (Table 2). Figure 2 demonstrates the average changes in authenticity score for individual courses, calculated as a percentage of change.

Results

RQ 1: Curriculum-Wide Assessment Inventory

Table 1 provides an overview of the ERT assessment inventory, with F2F aggregate data included as a reference. In the overall ERT curriculum, there were 61 courses (versus 62 F2F due to one fourth-year elective course not offered in ERT). First- and second-year courses were largely core and common across majors, such that there are 8 courses in each of first and second-year. Third- and fourth-year are comprised of more restricted elective options, such that there were 18 third-year and 27 fourth-year courses offered. Average class size across all courses was slightly larger in ERT (272 ERT versus 222 F2F) with class sizes ranging from one-on-one research opportunities to a max of 1700 students per section (versus 1-on-1 to 600 F2F), translating to 7 small, 26 medium, and 28 large courses (versus 13 small, 25 medium, 24 large F2F). Average class size decreased from years one to four. Notably, first year was comprised of exclusively large classes, while fourth year was comprised of primarily small and medium courses.

Table 1

Assessment Inventory Characteristics of ERT Curriculum, with F2F as a Reference

	F2F Complete Curriculum	ERT Complete Curriculum	1st year ERT	2nd year ERT	3rd year ERT	4th year ERT
# courses	62	61	8	8	18	27
Class Size						
Range	1:1-600	1:1-1700	280-1700	185-1000	40-550	1:1-325
Average (χ±SD)	222 ± 184	272 ± 283	628 ± 452	540 ± 236	256 ± 152	98 ± 88
# SMALL courses (<40)	13 (21%)	7 (11%)	0	0	0	7
# MEDIUM courses (<240)	25 (40%)	26 (43%)	0	1	8	17
# LARGE courses (>241)	24 (39%)	28 (46%)	8	7	10	3
Assessment Structure						
# Assessments total	457	517	134	81	126	173
# Tests total	225	248				
# Assignments total	233	266				
Average # assess/course (χ±SD)	**7 ± 5**	**8 ± 6**	**17 ± 7**	**10 ± 5**	**7 ± 3**	**6 ± 5**
Range	*2-27*	*3-27*	*7-25*	*4-16*	*3-13*	*3-27*
% marks comprised of tests (T)	63%	57%	***76%***	***91%***	***68%***	*34%*
vs. assignments (A)	37%	43%	*24%*	*9%*	*32%*	***66%***

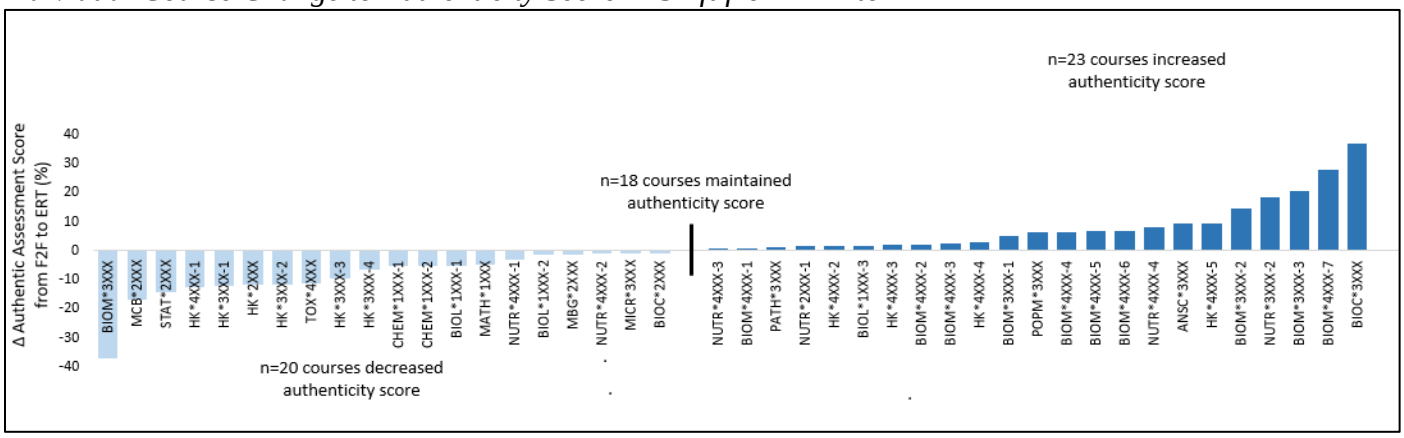
There were more total assessments in ERT at the curriculum level (517 ERT versus 457 F2F) and by course on average (8 ± 6 ERT versus 7 ± 5 F2F), ranging from 3 to 27 assessments per course (versus 2 to 27 F2F). More marks in the curriculum were due to assignments (43% ERT versus 37% F2F) and fewer marks due to tests (57% ERT versus 63% F2F). Early years (1-3) were dominated by tests, ranging from 68% to 91% of grades from tests, while fourth year was comprised of only 35% tests and 66% assignments. Many courses (*n*=33, 54%) increased the number of assessments given in their ERT offering (range: +1 to +9), while 13% of courses (*n*=8) decreased number of assessments (range: -10 to -1) and 33% of courses (*n*=20) did not change the total number of assessment touchpoints (Table 2). Specifically, 44% of courses (*n*=27) added assignments while 36% of courses (*n*=22) added tests. 20% of courses (*n*=12) had fewer tests, and 15% of courses (*n*=9) offered fewer assignments in the ERT offering compared to the F2F offering (Table 2). Notably, 34% of courses (*n*=21) decreased weighting of tests towards a greater emphasis on assignments (average: -24%, range: -88% to -3%).

RQ 2: Curriculum-Wide Assessment Authenticity

Most individual courses (70%) underwent a change in assessment authenticity in the shift from F2F to ERT. However, the direction and magnitude of change varied considerably. 38% of courses (*n*=23) increased authenticity (range: +2% to +37%, average: +9%), while nearly an equal proportion of courses (33%, *n*=20 courses) had a shift in the opposite direction (range: -37% to -1%, average: -9%) (Figure 2). Given the opposing responses, there was not a significant difference in curriculum level authenticity from F2F to ERT (1.8 ± 0.4 ERT versus 1.8 ± 0.6 F2F).

Figure 2

Individual Course Change to Authenticity Score in Shift from F2F to ERT



To better understand these different responses in authenticity, we categorized courses as ‘increasers,’ ‘decreasers’ or ‘maintainers’ and considered these groups by year of study (first – fourth), class size (small, medium, large), and assessment structure (count, type, weighting) (Table 2). Increasers, decreasers and maintainers did not demonstrate any significant differences between baseline (F2F) scores (Table 2). Courses which decreased, maintained or increased authenticity spanned each year of study and each class size category. However, many first-year courses (*n*=5, 62.5% of courses) and second-year courses *(n*=6, 75% of courses) decreased authenticity while several third-year courses (*n*=8, 44% of courses) and fourth-year courses (*n*=13, 48% of courses)

increased authenticity. Courses that decreased authenticity had the largest average class size (*n*=468), while increasers had a medium average class size (*n*=218) and courses which maintained authenticity were the smallest average class size (*n*=121), although considerable overlap exists. Typically, courses which decreased authenticity favoured tests (count: 7, weighting: 74%) over assignments (count: 4, weighting 26%). Conversely, courses which increased authenticity offered a larger number of assignments (count: 5) than tests (count: 3) per course on average and were approaching roughly equal marks assessed through tests and assignments (57% tests, 43% assignments). Finally, courses which maintained authenticity offered the fewest average number of assessments, divided equally between tests (count: 3) and assignments (count: 3) but favoured assignments slightly by weighting (weighting: 55%) rather than tests (Weighting: 45%). There was no significant difference in baseline (i.e., F2F) scores of courses that increased (1.8 ± 0.4), decreased (1.7 ± 0.4) or maintained authenticity (2.0 ± 0.5) in ERT (Table 2). However, those which increased and maintained ended with similar ERT scores (2.0 ± 0.4, 2.0 ± 0.5, respectively), while those which decreased authenticity were significantly lower (1.6 ± 0.3).

Table 2

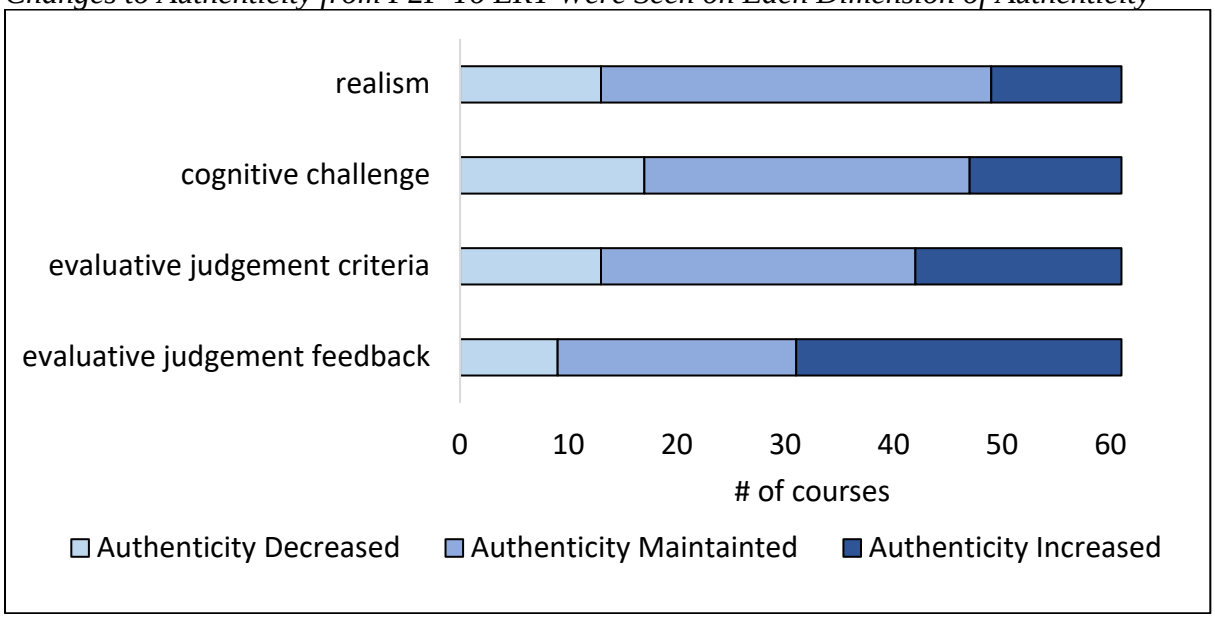
Course and Assessment Characteristics According to How Course Authenticity Changed in ERT

	Decreased Authenticity	Maintained Authenticity	Increased Authenticity
# courses (% of total curriculum)	20 (33%)	18 (30%)	23 (38%)
Year Level			
(n, % of courses in that year level)			
1st year	5 (62.5%)	2 (25%)	1 (12.5%)
2nd year	6 (75%)	1 (12.5%)	1 (12.5%)
3rd year	5 (28%)	5 (28%)	8 (44%)
4th year	4 (15%)	10 (37%)	13 (48%)
Class Size			
Range	30 – 1700	1:1 – 400	1:1 – 600
Average	468 ± 370	121 ± 123	218 ± 183
# SMALL courses *(<40)*	1	8	1
# MEDIUM courses *(<240)*	4	7	14
# LARGE courses *(>241)*	15	3	8
Assessment Structure			
Average **# assess**/course	**11 ± 6**	**6 ± 3**	**8 ± 6**
Average # **tests**/course	**7 ± 4**	**3 ± 3**	**3 ± 3**
Average # **assign**/course	**4 ± 4**	**3 ± 2**	**5 ± 7**
Average **% marks tests**	**74 ± 29**	**45 ± 42**	**54 ± 35**
Average **% marks assign**	**26 ± 29**	**55 ± 42**	**46 ± 35**
F2F AA Score	1.7 ± 0.4	2.0 ± 0.5	1.8 ± 0.4
ERT AA Score	1.6 ± 0.3	2.0 ± 0.5	2.0 ± 0.4

Changes to authenticity occurred on all core authenticity characteristics (Figure 3). For each characteristic, there were courses that made improvements, courses which decreased authenticity, and courses that maintained authenticity. Notably, the most frequent decrease to authenticity occurred for cognitive challenge as 17 courses declined in this regard. The most frequent increase to authenticity was seen for evaluative judgement feedback as 30 courses improved in this regard. Through one-on-one interviews, instructors provided insight into assessment design elements that they considered in the shift to ERT, which aligned with specific characteristics of authenticity. These considerations included resources (time available, class size, monetary cost), stress levels and workload (of both their students and they as instructors), practices regarding frequent low-stakes assessments and academic integrity, among others. For example, one instructor commented, "A suggestion from teaching folks was multiple, low stakes assessment, so we tried to go with this as much as possible. For example, we split up the midterm into two tests so the content was broken down and non-cumulative to give students a break, and a chance for feedback earlier in the semester." Another instructor commented, "There was a desire to give students an eye into expectations for the written and practical tests, therefore we wanted to provide multiple opportunities to do so via quizzes." Interestingly, increases and decreases in a characteristic of authenticity for a course were seen even when the same best practice was considered (e.g., frequent, low-stakes assessments) due to differences in how this was operationalized between courses.

Figure 3

Changes to Authenticity from F2F To ERT Were Seen on Each Dimension of Authenticity



Discussion

With the shift from F2F to ERT due to COVID-19, all instructors were afforded an opportunity, albeit with constraints, to take stock of their courses and think about what has historically worked in their courses, what colleagues were doing, and best assessment practices to prepare for remote teaching. This study documented how assessment structure and authenticity changed in matched courses in the shift from F2F to ERT. Most courses made changes to

assessment structure (count, type, and/or weighting) in the shift from F2F to ERT. Many courses (54%) increased the number of assessment touchpoints. Increased emphasis on tests was commonly seen in early-year courses, while increased emphasis on assignments was commonly seen in fourth-year courses. Many courses improved authenticity, suggesting this common impetus promoted positive change, although a roughly equal proportion of courses decreased authenticity. Decreased authenticity was more commonly seen in early-year, large classes relying on testing as a dominant assessment format whereas improvements were more commonly seen in medium-sized third- and fourth-year courses that increased reliance on assignments, though notable exceptions did occur. There were the greatest number of improvements on the characteristic of evaluative judgement feedback in the shift to ERT. Interestingly, instructors adopted a variety of widely shared best practices, though adoptions varied by environmental influences. From this comprehensive assessment scan during ERT, coupled with a matched data set from a pre-COVID-19 F2F offering, we continue to deepen our understanding of assessments in a complete, representative science curriculum, while providing data to inform priorities and decision making in the future.

Best Practices Were Operationalized Differently, but Instructors Faced a Variety of Course Constraints

In the complete curriculum, there was no change to authenticity from F2F to ERT. However, most courses made changes to assessment structure, and most individual courses changed authenticity (increased, decreased) while fewer maintained authenticity (Figure 2). When speaking with instructors, we asked them to describe any factors they considered when shifting their course(s) from F2F to ERT. While there was certainly variability in instructor responses, there were several common elements that informed if and how instructors made changes in the shift to ERT. Common elements considered included widely circulated best practices such as protecting academic integrity and frequent low-stakes assessments, as well as the realities of the course contexts they teach in (year level, class size). Interestingly, some of these considerations were operationalized differently depending on the course, representative of the diverse responses to the shift to ERT occurring across the globe (Crawford et al., 2020).

Best Practice Operationalization

Academic Integrity

Slade (2020) suggests assessments were shifted from F2F to remote without academic integrity as a primary focus such that there was little or no assessment redesign in this regard. In contrast, we found that concerns around academic integrity encouraged instructors to make changes to assessments through innovating test questions to require higher levels of thinking (and thus be “non-google-able”), while some opted to introduce more contextualized assignments in addition to, or instead of, tests. Many instructors opted to implement open-book or take-home assessments. For example, an instructor commented, “I wanted to make the course as accessible as possible—all open book, all double time to write the assessments. I wanted questions to be ungoogle-able, therefore I increased the cognition level required by questions.” In these cases, courses consistently shifted towards higher authenticity by improving realism and cognitive challenge. Conversely, some instructors expressed concern over the release of content (test

questions, answer keys) following an assessment, fearing that current students would share this information with incoming students. Therefore, some instructors opted to not host review sessions or office hours to go over the assessment, resulting in a decrease to authenticity, specifically evaluative judgement.

Frequent, Low-Stakes Assessments

The best practice of frequent, low-stakes assessments (Schrank, 2016) was implemented both in courses which decreased and increased authenticity. In both cases, instructors expressed the desire to break down heavily weighted assessments into smaller pieces to reduce student stress and cognitive load at the time of each assessment. In cases of decreased authenticity, this best practice was often operationalized by decreasing the weighting (and/or number) of highly weighted summative assessments (e.g., midterm/final exam) in order to include several non-cumulative low-stakes quizzes. This inclusion of non-cumulative low-stakes quizzes often resulted in a decreased realism and cognitive challenge score for an assessment. However, these often served as practice questions for higher stakes summative assessments and provided students with multiple points of information regarding their progress in the course, therefore improving evaluative judgement criteria and feedback. This method of providing criteria and feedback is supported by available literature which suggests that frequent low-stakes assessments (e.g., quizzes) can help students become familiar with remote assessments (Rahim, 2020). Alternatively, frequent low-stakes assessments were also operationalized by shifting a test-based course in F2F to include several assignments in ERT. We have previously shown assignments are more authentic than tests (Hobbins et al., 2021). This trend held true in ERT, and therefore it is unsurprising that those who maintained or improved authenticity favoured assignments rather than tests (Table 2). Our data suggests that a shift towards assignments is a positive addition to courses with respect to authenticity.

Course Constraints

It is important to acknowledge that instructors work within the context of a variety of environmental influences (e.g., year level: foundational core course versus upper year specialized course, class size: large versus small class, etc.). Courses which maintained and improved authenticity were primarily smaller, upper year research opportunities and project-based courses. These courses tend to be restricted elective courses, which classically allow for more instructor autonomy and flexibility in course design – a single instructor designs and delivers the course and it is completed by a smaller group of students. For example, the range of change to authentic assessment score for third-year was large (-12% to 37%), suggesting that such courses may lend well to higher authenticity, or have the greatest ability to improve authenticity. In contrast, core foundational courses typically are perceived as inflexible service courses, having more immovable parts as several instructors and/or course coordinators design and deliver the course over multiple sections to a large group of students. Further, the first-year range of change was small (-5% to 3%), such that these large first-year courses decreased authenticity to a smaller degree (average: -4%), arguably nearly maintaining authenticity, compared to other years of study. For example, most second-year courses (n=6, 75%) decreased authenticity but did so at an average decrease of 14% (range: -37% to -1%). As most second-year courses (n=7, 88%) are comparable in size to first-year i.e., large, this result suggests second-year courses were impacted most negatively in the

shift to remote. The reasons why this negative impact on second-year courses occurs are unclear and should be investigated further. These data may serve as a reminder that small incremental changes may be more realistic in large early year courses, while small senior courses may be more likely to see significant shifts for the better.

Sustainable Assessment Strategies to Support Core Characteristics of Authenticity in Any Setting

The mean authenticity score of the F2F and ERT curricula were similar (M: 1.8). However, many individual courses (67%, *n*=41) maintained or improved authenticity in the shift to ERT. Notably, we documented a variety of positive changes to each characteristic of authenticity in the ERT context. Below are five key practices that resulted in changes to authenticity on all core characteristics in this health science curriculum that readers may consider for their own curricula.

Shift Focus from Tests to Assignments

Frequent low-stakes assessments do not have to be in the form of weekly check-in multiple-choice quizzes, rather, you might consider low stakes *assignments* in the form of discussion board prompts to interact with course content in action. Realism (closing the gap between the classroom and real-world) was improved in 12 courses in the shift to ERT largely due to instructors placing a greater emphasis on assignments (either in absolute number or greater weighting). For example, a fourth-year course shifted four multiple-choice tests (100% tests) in F2F to two quizzes (12% of marks) and seven take-home assignments (88% of marks due to assignments) in ERT. The increased emphasis on assignments leading to improved authenticity is consistent with the previous finding that assignments are more authentic than tests (Hobbins et al., 2021). Further, previous research has demonstrated students are more likely to employ a surface approach to learning when presented with multiple-choice questions, compared to a deep approach to learning when preparing for assignments, as assignments are perceived to be assessing higher levels of cognitive processing (Scouller, 1998). Therefore, this greater emphasis on assignments may be an assessment design choice that could be considered moving forward. Notably, courses that were assignment-dominant to begin with (F2F) were most resilient to the stresses of a shift to ERT, and therefore may allow for a more flexible course structure regardless of teaching context. The inclusion of more assignments also resulted in improvements to evaluative judgement feedback, where rubrics are returned to students with comments beyond the corrective feedback typical of tests, notably where there are multiple drafts of an assignment. Additionally, courses with improved evaluative judgement feedback scores had decreased the weighting of a high-stakes final exam which historically does not provide so much as a grade to students (little opportunity for feedback). In cases where there was a shift towards increased emphasis on tests, typically in the form of low-stakes quizzes or multiple midterms, realism declined. This resultant decrease in realism suggests there may be value in avoiding low-stakes tests to protect against low realism. Shifting focus from tests to assignments is consistent with literature which posits that students are more engaged in their learning online, resulting in effective learning, when they complete meaningful, authentic work, such as assignments rather than quizzes (Bailey, Hendricks, & Applewhite, 2015).

Incorporate Higher Order Vignette Style Assessment Questions

Consistent with literature which reports a substantial shift away from memory-based exams to higher-level thinking assessments (Jankowski, 2020), the 'non-googleable' focus of remote assessments in our context saw many instructors shift traditional lower-order multiple-choice questions to open-book tests coupled with vignette style questions, often resulting in improved cognitive challenge. These questions required students to engage with higher levels of thinking to a greater degree than traditional, content-focused questions, which may more easily be answered through Google in an unsupervised assessment. Therefore, instructors may consider including vignette-style, open-ended assessment questions in their courses, whether these be in F2F or remote. This item for consideration is proposed with the caveat that instructors ensure alignment between the course learning outcomes and instructional activities to offer the most appropriate learning environment for students (i.e., higher order questions of this nature may not be appropriate for a lower level, foundational course). An additional consideration is open-ended, constructed response assessment items can be labour intensive to grade. However, some courses were able to achieve this higher order, vignette-style question through multiple-choice, thereby simplifying grading. For example, a fourth-year biomechanics course offered clinical patient scenarios followed by multiple-choice questions which required application and critical thinking.

Include a Review or "How-To" Session for Assessments

With years of experience instructing the same course, it is possible instructors may have begun to gloss over expectations for a routinely offered assessment that is well established in a course. The new focus on assessments in ERT encouraged many instructors to revisit tried and true assessments, and/or develop new, unfamiliar assessment opportunities. 19 courses improved evaluative judgement criteria by including a review or "how-to" session with explicit instructions (e.g., rubric) and the provision of exemplars and practice questions for assessments in places where this wasn't done previously, to prepare students for what to expect in an online test. In parallel, circulated best practices encouraged instructors to provide students with clear expectations and practice opportunities (e.g., drafts). In these cases, evaluative judgement criteria were improved. Often, the shift to focus on assignments in the curriculum was coupled with the provision of a rubric (ahead of the submission deadline) in many cases when it wasn't previously provided for the same assignment in F2F, or where there were 12ssignents to provide such a rubric for in the course in F2F. Frequent low stakes assignments in the form of smaller drafts serve as chunks or practice for the final submission allowing students to engage with what is expected of them on the assessment, thereby providing students with transparent expectations for the larger assignment (Rahim, 2020).

Remote Technologies May Facilitate More Frequent Feedback

One could argue there is a sense of isolation and reduced personal interactions in the remote setting. However, it has previously been suggested that the remote environment provides more opportunities for one-on-one feedback as students have more frequent access to the instructor (Beebe et al., 2010). Indeed, positive changes at the individual characteristic level were most frequently seen for feedback in our data as 31 courses improved evaluative judgement feedback authenticity in the shift to remote. Slade et al. (2021) note the importance of providing feedback

to students, but highlight a loss in real-time feedback in the shift to ERT as opportunities for skill-based (e.g., practical) learning were limited. However, we contend not all is lost as there are multiple means of providing remote feedback to students, evidenced by ~50% of instructors improving authenticity on this characteristic in this study. Some instructors opted to implement tailored automated feedback on lower stakes assessments (e.g., quizzes), to direct students' focus for future similar assessments. Though this finding suggests the remote environment may better facilitate multiple points of feedback for students, the quality of feedback must also be considered. Şenel and Şenel (2021) report that while most students agreed that assessment results and feedback were instant in ERT, they also asserted that this feedback did not significantly increase the effectiveness of their learning. However, it was demonstrated that students reported infrequent opportunities to engage with peer and self assessment during the pandemic (Şenel & Şenel, 2021). In our context, while low stakes quizzes may allow for frequent points of information regarding performance (i.e., grades) or corrective feedback, this feedback was not as authentic as intentionally scaffolded drafts which included reflection or peer review. Therefore, opportunities wherein students are not only receiving feedback, but also providing feedback, are encouraged (e.g., peer review).

Incorporate Ungraded, Informal Check-in Points Throughout the Course

Given that a goal of low-stakes assessments is to help students stay on track and measure progress in a course, instructors may consider implementing more frequent ungraded check-ins (e.g., polling students). As evaluative judgement is rooted in student ability to judge their own work, grades may not always be necessary, especially since instructors mentioned instances of stress due to overloading students and teaching assistants with graded works. However, it worth noting that health science students are often high achievers and tend to be in a state of heightened awareness when it comes to anything presented as an "assessment," regardless of its weighting. Therefore, instructors are encouraged to present these check-ins as ungraded and informal. As ~50% of instructors improved authenticity for feedback in our data, there is evidence that multiple points of (graded) check-ins were occurring in the curriculum.

Contributions, Limitations and Future Directions

The above key practices incorporated in this health science curriculum may provide educators with direction as courses continue to be offered with elements of ERT. While these tangible practices may not be ground-breaking ideas for those familiar with and fluent in the Scholarship of Teaching and Learning, our curriculum wide review (with instructors at varying levels of engagement) are likely reflective of many Canadian institutions. Therefore, our curriculum wide scan provides an evidence base to inform program wide policies and recommendations and serves as a reminder that simple recommendations (e.g., offer a how-to session prior to an assessment) can improve authenticity of a course's assessments. However, it is important to note that instructors face a variety of competing interests and priorities such as their research programs, graduate student training programs, and their funding structures, while also balancing home and personal life. Therefore, COVID-19 is unlikely to represent the ideal rethink/redesign of course assessments. It is thus possible that other challenges may have dampened the impact of the potential opportunity to rethink assessment, and our results therefore need to be interpreted within that wider reality.

Even so, our data are consistent with early emerging reports in the literature, as we show most courses (85%) made an assessment-related change (count, type, weighting, and/or authenticity). Notably, our data consider a complete curriculum comprised of all instructors, rather than being limited to the confines of survey data (self report, subsample). Further, available literature does not specify how these changes presented (e.g., how assessments changed weighting or how many tests or assignments were reduced/increased per course by count). In our curriculum, we demonstrate there was an increase in both the total number of tests (249 ERT versus 224 F2F) and assignments (268 ERT versus 233 F2F) in the complete curriculum, but an increased proportion of marks came from assignments in the shift to ERT (43% assignments ERT versus 36% assignments F2F). Perhaps most importantly, we acknowledge that even where courses did not change authenticity score, instructors were thoughtful and intentional in transitioning elements of their courses to a remote setting. Instructors in these courses often expressed a desire to continue offering a course that had historically been successful with students in F2F with an assessment structure that seemed to work.

The challenge now will be how to use this common impetus of mass level course rethink and redesign to highlight the potential for continued course improvement, without the constraints and stress that couple COVID-19. Further, given that instructors considered and variably implemented best assessment practices in this shift to ERT, the authors also suggest that more work is needed to clearly communicate examples of how best practices can be implemented on a complete curriculum level. Possibly, uncovering more examples of how best assessment practices were operationalized across a variety of curricula (i.e., beyond health sciences) and by a variety of instructors (i.e., of varied teaching experience) may lend to more clear, widespread communication of how to best implement assessment practices. Lastly, should elements of ERT remain in our course delivery, there is a marked need to evaluate what elements of ERT are best for student learning. Student perception of assessment design elements, with a focus on characteristics of authenticity, would be a valuable addition to this work and the wider body of literature.

Conclusion

The rapid shift to an ERT mode of teaching and learning has presented concerns for assessment in student learning, as online learning can be perceived as lesser quality than F2F learning (Hodges et al., 2020). This study compared a complete health sciences curriculum in the F2F and ERT settings with respect to assessments (count, type, authenticity) using our literature-informed AAT. Our results demonstrate most instructors changed their assessment structure from F2F to ERT, but the operationalization and output of these changes varied considerably. Despite the common narrative that ERT is lesser quality than F2F, coupled with the overwhelming challenges of ERT, our work shows most courses maintained or even improved authenticity of assessments. Senior courses were more likely to see improvements, and the introduction of assignments to many courses was a positive addition. Another promising finding was more authentic feedback was seen in ERT versus F2F. However, there is still a need to direct training and resources to support feedback methods (those which are often labour intensive), as this was still the weakest characteristic of authenticity. Introducing alternative forms of assessment, such as open book tests and assignments, may be a relatively straight forward method to improve authenticity in a course by moving away from compartmentalized content testing towards more contextualized applications.

Overall, this research provides evidence that courses in a broadly representative curriculum were able to maintain or improve authenticity in the shift from F2F to ERT. Importantly, while it was suggested that there has been more focus on instructional methods than assessment methods (Jankowski, 2020), we wish to acknowledge and commend instructors for their reflective and thoughtful efforts in assessment design in response to the pandemic, despite the large number of competing demands within and beyond academia. Future research may consider how to best communicate and operationalize best assessment practices on a wider curriculum level in a sustainable way for instructors and students, beyond health sciences, as well as student perceptions of which assessment design elements impact their learning.

References

Adapting your assessments for remote teaching and learning. (n.d.). University of Guelph, Office of Teaching and Learning. Retrieved April 21, 2023, from https://otl.uoguelph.ca/system/files/Adapting%20Your%20Assessments%20for%20Remote%20Teaching%20and%20Learning_2.pdf

Bailey, S., Hendricks, S., & Applewhite, S. (2015). Student perspectives of assessment strategies in online courses. *Journal of Interactive Online Learning*, *13*(5), 112–125. https://www.ncolr.org/jiol/issues/pdf/13.3.2.pdf

Bartolic, S. K., Boud, D., Agapito, J., Verpoorten, D., Williams, S., Lutze-Mann, L., Matzat, U., Moreno, M. M., Polly, P., Tai, J., Marsh, H. L., Lin, L., Burgess, J.-L., Habtu, S., Rodrigo, M. M. M., Roth, M., Heap, T., & Guppy, N. (2022). A multi-institutional assessment of changes in higher education teaching and learning in the face of COVID-19. *Educational Review*, *74*(3), 517–533. https://doi.org/10.1080/00131911.2021.1955830

Bearman, M., Dawson, P., Bennett, S., Hall, M., Molloy, E., Boud, D., & Joughin, G. (2017). How university teachers design assessments: A cross-disciplinary study. *Higher Education*, *74*(1), 49–64. https://doi.org/10.1007/s10734-016-0027-7

Beebe, R., Vonderwell, S., & Boboc, M. (2010). Emerging patterns in transferring assessment practices from F2F to online environments. *Electronic Journal of e-Learning, 8*(1), 1–12. https://files.eric.ed.gov/fulltext/EJ880094.pdf

Branch, R. M., & Dousay, T. A. (2015). Survey of instructional design models. In D. Walling (Ed.), *Survey of instructional design models* (5th ed., pp. 20-39). Association for Educational Communications and Technology.

Cash, C. B., Letargo, J., Graether, S. P., & Jacobs, S. R. (2017). An analysis of the perceptions and resources of large university classes. *CBE—Life Sciences Education, 16*(2), ar33. https://doi.org/10.1187/cbe.16-01-0004

Crawford, J., Butler-Henderson, K., Rudolph, J., Malkawi, B., Glowatz, M., Burton, R., Magni, P. A., & Lam, S. (2020). COVID-19: 20 countries' higher education intra-period digital pedagogy responses. *Journal of Applied Learning & Teaching, 3*(1), 9-28. https://doi.org/10.37074/jalt.2020.3.1.7

Exploring alternative assessment types for remote proctored assessment alternatives. (n.d.). University of Guelph, Office of Teaching and Learning. Retrieved April 21, 2023, from https://otl.uoguelph.ca/system/files/Exploring%20Alternative%20Assessment%20Types%20-%20for%20remote%20proctored%20assessment%20alternatives.pdf

Hobbins, J., Kerrigan, B., Farjam, N., Fisher, A., Houston, E., & Ritchie, K. (2021). Does a classroom-based curriculum offer authentic assessments? A strategy to uncover their prevalence and incorporate opportunities for authenticity. *Assessment & Evaluation in Higher Education, 47*(8), 1–15. https://doi.org/10.1080/02602938.2021.2009439

Hodges, C., Moore, S., Lockee, B., Trust, T., & Bond, A. (2020). *The difference between emergency remote teaching and online learning*. EDUCAUSE. https://er.educause.edu/articles/2020/3/the-difference-between-emergency-remote-teaching-and-online-learning

Jankowski, N. A. (2020). *Assessment during a crisis: Responding to a global pandemic*. University of Illinois and Indiana University, National Institute for Learning Outcomes Assessment. https://www.learningoutcomesassessment.org/wp-content/uploads/2020/08/2020-COVID-Survey.pdf

Johnson, N., Veletsianos, G., & Seaman, J. (2020). U.S. faculty and administrators' experiences and approaches in the early weeks of the COVID-19 pandemic. *Online Learning, 24*(2), 6-21. https://doi.org/10.24059/olj.v24i2.2285

Office of Teaching and Learning. (n.d.). Shifting your assessments to the remote environment. Retrieved March 11, 2022, from https://otl.uoguelph.ca/teaching-remotely/remote-assessments/shifting-your-assessments-remote-environment

Ontario Ministry of Colleges and Universities, & University of Guelph. (2017). 2017-2020 strategic mandate agreement: University of Guelph. Retrieved January 6, 2024, from https://www.ontario.ca/page/2017-20-strategic-mandate-agreement-university-guelph

Rahim, A. F. A. (2020). Guidelines for online assessment in emergency remote teaching during the COVID-19 pandemic. *Education in Medicine Journal, 12*(3), 59–68. https://doi.org/10.21315/eimj2020.12.2.6

Schrank, Z. (2016). An assessment of student perceptions and responses to frequent low-stakes testing in introductory sociology classes. *Teaching Sociology, 44*(2), 118–127. https://doi.org/10.1177/0092055X15624745

Scouller, K. (1998). The influence of assessment method on students' learning approaches: Multiple choice question examination versus assignment essay. *Higher Education, 35*(4), 453–472. http://www.jstor.org/stable/3448270

Şenel, S., & Şenel, H. C. (2021). Remote assessment in higher education during COVID-19 pandemic. *International Journal of Assessment Tools in Education, 8*(2), 181-199. https://doi.org/10.21449/ijate.820140

Slade, C., Lawrie, G., Taptamat, N., Browne, E., Sheppard, K., & Matthews, K. E. (2021). Insights into how academics reframed their assessment during a pandemic: Disciplinary variation and assessment as afterthought. *Assessment & Evaluation in Higher Education, 47*(4), 1–18. https://doi.org/10.1080/02602938.2021.1933379

Sotiriadou, P., Logan, D., Daly, A., & Guest, R. (2020). The role of authentic assessment to preserve academic integrity and promote skill development and employability. *Studies in Higher Education, 45*(11), 2132–2148. https://doi.org/10.1080/03075079.2019.1582015

Villarroel, V., Bloxham, S., Bruna, D., Bruna, C., & Herrera-Seda, C. (2018). Authentic assessment: Creating a blueprint for course design. *Assessment & Evaluation in Higher Education, 43*(5), 840–854. https://doi.org/10.1080/02602938.2017.1412396

Appendix
Guiding Discussion Prompts for One-on-one with Instructors

1. COGNITIVE CHALLENGE

The first element I am going to ask you about is called "cognitive challenge", which refers to the way students are required to use knowledge. This occurs on a spectrum of understanding → application/interpret → evaluate/critique.

Q: Where does each assessment fall on this spectrum? Can you provide an e.g.?

Prompts: Based on your learning outcomes, this assessment requires students to use skills:

→ *Can you provide an example?*

→ *Can you ballpark the % of how much of this assessment is based on learning outcomes x, y, z?*

2. REALISM

The next element I am going to ask you about is called "realism", which refers to how assessments might engage students with problems or challenges they may face beyond the classroom.

Q: Are Qs on these assessments presented in any kind of context? [let them answer] Tell me a bit about that i.e. what kind of context/what does the context look like. Can you provide an example?

Prompts: To clarify, are core concepts presented mainly in a disciplinary context or beyond?

Can you provide an example? Does this represent the majority of the assessment Qs?

3. CRITERIA

Q: How are guidelines and expectations communicated to students for each assessment?

Prompts: Leading up to each assessment, do you talk to students about what they could expect to see/be required to do? Do you talk to them about how they could best prepare?

Do students engage with grading criteria before and/or after the tests? Can you provide an e.g.?

4. FEEDBACK

Q: Can you tell me a bit about how feedback is provided to students for each assessment?

Prompts: Following this assessment, do you talk to students about common errors, thought processes, etc. How do you do this?

(key words to note: timing, what FB is provided for (content vs skills), courselink, rubric, grade-only, corrected content, skill, actionable steps to improve, general statements to class)

5. INSTRUCTOR INSIGHT

Q: You were able to largely retain the same assessment structure in the remote setting from the F2F setting. Overall, how did you decide on these assessments for this remote semester?

OR

Q: You made several changes to the assessment structure from the F2F setting to the remote setting. Overall, how did you decide on these assessments for this remote semester?

Prompts: What factors or considerations did you make in this decision?