Does a classroom-based curriculum offer authentic assessments? A strategy to uncover their prevalence and incorporate opportunities for authenticity.

Justine O. Hobbins, Bronte Kerrigan, Niloufar Farjam, Ashley Fisher, Emilie Houston and Kerry L. Ritchie*

Department of Human Health and Nutritional Science, University of Guelph, Guelph, Ontario, Canada

Kerry Ritchie, PhD
Dept. of Human Health and Nutritional Science
University of Guelph
50 Stone Rd E, Guelph, ON, Canada
N1G2W1
email: ritchiek@uoguelph.ca
Does a classroom-based curriculum offer authentic assessments? A strategy to uncover their prevalence and incorporate opportunities for authenticity.

Authentic assessment is revered to support student learning, but it is typically described within the context of work-integrated learning and professional schools, leaving one to question whether a traditional classroom-based curriculum can offer authentic assessments. This study documented the prevalence of authentic assessments throughout a complete Health Science undergraduate curriculum in accordance with the four core dimensions of authentic assessment: realism, cognitive challenge, evaluative judgement criteria and feedback. Using a literature-informed authentic assessment tool and institutionally standardized course syllabi, 455 assessments in 62 courses were classified as low, moderate or high on core authentic assessment dimensions. Results show that few assessments scored high across all core dimensions (<1% of all assessments), as there was considerable variability across dimensions. Feedback was the weakest dimensional authenticity score (1.5±0.4). Authentic assessments were more prevalent within upper-year, small capstone courses, although they were not precluded from early-year, large classrooms. Assignments were significantly more authentic than tests (2.0±0.5 versus 1.5±0.4), though tests were more dominant in the curriculum (63% marks from tests vs. 37% from assignments). This work serves as a model for others seeking to review assessments in their curriculum and provides evidence from a large, representative BSc program to make practical recommendations to promote authenticity of assessments.

Keywords: authentic assessment; realism; evaluative judgement; feedback
Introduction

Assessment is central to and drives student learning, setting standards and intentions for a course (Gibbs and Simpson 2005). This impacts how students approach studying, the content they focus on, how they perceive courses and even how they select courses for their schedules (Brown 1997; Gibbs and Simpson 2005). While students typically experience their education as a curriculum (five concurrent courses/semester), assessment design is often conducted in isolation within a single course (Dawson, Carless, and Lee 2021). Curriculum mapping of learning outcomes has become more prevalent for programs to better understand and articulate their progressive learning goals, but less attention has been given to the quality of assessments that students experience across their curriculum to document these. Therefore, to offer a learning experience most effectively and comprehensively to students through assessment design, we must understand the curriculum as experienced by students (i.e. beyond one course).

A test-based environment focused in knowledge reproduction is often relied upon in undergraduate science assessments (Villarroel et al. 2018; 2019), primarily requiring the recall of disjointed facts, memorization and rote-problem solving (Wood 2009). This is particularly true for lower level study due to the focus on foundational knowledge and large courses as a result of resource constraints (e.g. time, teaching support, administrative load) (Kerr 2011). Consequently, graduates experience difficulties in applying knowledge in new, real-life situations and lack preparedness for employment (Wood 2009). As the responsibility for equipping students with adequate employability skills is increasingly placed on higher education institutions (Sokhanvar, Salehi, and Sokhanvar 2021), there is an increased demand in competencies for learners (e.g. critical thinking, problem solving, communication). Therefore, rote learning and regurgitation of facts or procedures associated with the commonly relied upon test-based environment is
insufficient (Wiggins 1990). Instead, the design of assessment processes where the focus is shifted from memorization of disjointed facts to supporting student learning may lead to the development of skills that graduates require for everyday life.

A well-regarded approach to better support student learning and close the gap between the classroom and real world is authentic assessment (Wiggins 1990). Authentic assessment is reported to engage students in problem solving and critical thinking (Sokhanvar, Salehi, and Sokhanvar 2021). The term authentic assessment was first established by Grant Wiggins (1989), positing that knowledge and ability should be tested through the performance of exemplary tasks. As the level (or degree) of authenticity can vary, authenticity can be scaffolded to offer a range of assessment opportunities to students (Iverson, Lewis, and Talbot 2008; Bosco and Ferns 2014; Kaider and Hains-Wesson 2017). Notably, while many definitions of authentic assessment exist in the literature, consistent among these is the element of real-world experience, often while solving ill-structured problems (Wiggins 1990; Bosco and Ferns 2014). Authentic assessment is reported to close the gap between the classroom and the outside world by presenting students with an assessment task that mimics those performed by professionals in the field, and/or physically occurs in the workplace (Wiggins 1990). As such, authentic assessment is most often discussed in the literature within the context of work-integrated learning and professional schools (Gulikers, Bastiaens, and Kirschner 2004; Kaider and Hains-Wesson 2017). Limiting discussions and practices of authentic assessment to work-integrated learning and professional schools may be problematic given that these environments tend to be smaller in class size relative to a general degree programs (“Nursing” n.d.). Notably, work-integrated learning and professional school benefit from a common professional goal among most students (e.g. nursing) rather than a variety of career goals that would be typical in traditional, large classroom-based curricula.
However, several conceptualizations of authentic assessment do move beyond just the relevance of the task to professional life, and include relevance to daily-life outside of work (Bosco and Ferns 2014), consider the level of intellectual engagement with the task (Bosco and Ferns 2014), and the importance of judgement related to criteria (Gulikers, Bastiaens, and Kirschner 2004).

Most recently, Villarroel et al. (2018) comprehensively reviewed the literature to conceptualize authentic assessment for more traditional university-based learning. Rooted in existing authentic assessment literature, they outlined the design dimensions required to bring authenticity to classroom-based (versus work-based) learning. They present a practical stepwise model to incorporating authentic assessment into classrooms, outlining three core dimensions: realism, cognitive challenge, evaluative judgement to be implemented over four steps (1. realism, 2. cognitive challenge, 3. evaluative judgement criteria, 4. evaluative judgement feedback) (Villarroel et al. 2018; 2019). Realism describes the closure of the gap between the classroom and real-world by engaging students with problems or important questions that are relevant to everyday life beyond the classroom (Villarroel et al. 2018; 2019). This can be achieved by contextualized problems and/or by presenting the task in a setting beyond the classroom. Cognitive challenge describes the thinking skills required to use knowledge, process information, make connections and rebuild information to complete a task (Villarroel et al. 2018; 2019). Evaluative judgement describes the need for students to judge their own performance and regulate their own learning through the experience of judging quality (e.g. criteria) and providing and receiving information regarding their responses to assessments (e.g. feedback) (Villarroel et al. 2018; 2019).

While the framework by Villarroel et al. (2018) is comprehensive in describing dimensions of authentic assessment, there is limited research on the prevalence of authentic
assessment across a complete classroom-based curriculum. Bosco and Ferns (2014, 285) state that “authentic learning tasks are fundamental measures of a program’s distinctiveness and serves as an important criterion for attracting students”. Notably, the institution under study in this paper indicates a central commitment to authentic assessment within the Institutional Strategic Mandate Agreement, a document which outlines teaching and learning initiatives: “Through…authentic assessments, we aim to provide our students with a high impact education” (Hodgson 2017, 10). As institutions consider where and how to incorporate authentic assessment to support student learning, it is important to document authentic assessment opportunities that are currently offered across the complete curriculum as experienced by the students (i.e. beyond one course), particularly in classroom-based learning. To this end, we sought to firstly inventory all the assessments students are exposed to throughout a primarily classroom-based health sciences curriculum. Secondly, we sought to document the prevalence of authentic assessment throughout this curriculum in accordance with the core dimensions of authentic assessment presented by Villarroel et al. (2018). We further considered inventory and authenticity data in terms of year of study (first year – fourth year), class size (small, medium, large), and assessment type (test or assignment). While this is not likely to be the first complete assessment inventory of a broadly applicable health sciences curriculum, this is the first to be shared widely with practical suggestions for use beyond an institutional level. Notably, we provide specific evidence-based recommendations to improve authenticity of assessments based on a widely representative undergraduate science program, while providing a strategy for others who may want to review their existing curriculum. Finally, specific assessments from our curriculum are highlighted, which may serve as practical exemplars to motivate others.

Methods
Research context

This research takes place at a Canadian research-intensive, comprehensive university. Health science majors (Biomedical Sciences, Human Kinetics and Nutrition and Nutraceutical Science) comprise ~40% of the Bachelor of Science program student population (n=~3900) (“Common University Data Ontario (CUDO)” 2019). The university is competitive enrollment, with incoming high school averages in the mid-80% and above for health science majors. Typically, students complete the BSc health sciences program over a total of eight semesters (~four years). During this time, students complete a selection of biology, chemistry, physics, math courses (i.e. core courses), a selection of science courses that concentrate on an area of health sciences (i.e. restricted electives), and a selection of other science or liberal education courses (i.e. free electives). Of note, courses completed in early years are common to all undergraduate BSc students at the institution of study, and largely at similar institutions across Canada, regardless of major (e.g. computer science, physical science, agriculture), therefore this subset of data can be applicable to students beyond health sciences.

Developing an assessment inventory

A list of core courses and restricted electives available to the three health science majors was collated based on the published course calendar and freely available standardized syllabi for each course was collected (no data collected for free elective courses). Year level (first, second, third and fourth year) and class size of each course was recorded and subsequently categorized as small (one-40 students), medium (41-240 students) or large (>240 students) based on the classification of Cash et al. (2017). Further, the number of assessments per course and their weighting towards a final grade in the course (out of 100%, i.e. a proxy for assessment size) was recorded. Assessments were categorized as ‘test’ (quiz, midterm and final exam) or ‘assignment’
(all other assessments, e.g. paper, project). Upon documenting the opportunities for assessments that students are exposed to throughout the curriculum as an inventory, the prevalence of authenticity among these was determined.

**Development of the authentic assessment tool**

Rooted in the core dimensions of authentic assessment identified by Villarroel et al. (2018, 2019), our research operated under the following evidence-informed definition: Authentic assessment refers to a formally evaluated assessment activity which engages students with problems or important questions that are relevant to everyday life beyond the classroom; prompts 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. Following this conceptualization of authentic assessment, we created the authentic assessment tool, a rubric that allows for scoring of individual assessments against each of the core dimensions proposed by Villarroel et al. (2018); realism, cognitive challenge, evaluative judgement – criteria and evaluative judgement – feedback (Table 1). Notably, our tool presents each dimension on individual rows, allowing one dimension to score different from the others, and specific descriptors range from low to moderate to high levels for each.

**Application of the authentic assessment tool**

Using the information collated in the assessment inventory and freely available course syllabi, individual assessments were scored according to each dimension of authenticity on the authentic assessment tool. This process involved three evaluators independently placing each assessment on the authentic assessment tool. Following the initial application of the authentic assessment tool,
tool, evaluators came together as a group to reach consensus, similar to methods previously suggested for applying the Bloom’s Taxonomy framework to assessments to identify their level or rank (Crowe, Dirks, and Wenderoth 2008). Following this, instructors from each course offering were invited to participate in a brief (15-minute) semi-structured in-person discussion with an evaluator to confirm details about their course’s assessment structure and design. This process allowed the confirmation of the initial categorical authentic assessment scoring (low, moderate, high) that was retrieved from the syllabi, or provided additional information necessary for scoring that was not present on the syllabi through a series of prompts and pointed questions (e.g. can you tell me a bit about how you provide feedback on the midterm?).

Each assessment was assigned a categorical score of low, moderate or high for each of the four authentic assessment dimensions, which were then converted to numerical scores of one, two, three, respectively. These 4 dimensions were then averaged to give individual assessments an average authentic assessment score (/3). This was repeated for all assessments in a course. To account for the different weighting of assessments (i.e. assessment size) within a course, each assessment was multiplied by its weighting (/100) and these scores were summed to produce a course-level authenticity score. For example, a weekly quiz worth 1% would contribute much less to the overall authenticity score of a course compared to a final term paper weighted at 40%. The same process was repeated on an individual dimension to produce the dimensional authenticity scores, allowing for the examination of differences between dimensions.

An unpaired t-test was performed to compare the average authentic assessment score of core courses and restricted electives. A series of one-way ANOVAs were performed to compare between dimensional authenticity scores of the complete curriculum (Figure 1), overall average authentic assessment score by year of study and dimensional authenticity scores by year of study.
and overall average authentic assessment score by class size and dimensional authenticity scores by class size (Figure 3). Tukey’s HSD post-hoc analyses were conducted for multiple comparisons. A linear regression was performed to analyze the relationship between average authentic assessment score and class size in fourth year specifically (Figure 4). A series of unpaired t-tests were performed to compare the overall average authentic assessment score and dimensional authenticity scores by assessment type (Figure 5). Results are presented with mean ± standard deviation.

Results

Assessment inventory

Table 2 provides an overview of the assessment inventory.

In the overall curriculum, there were 62 courses, taught by 45 instructors. Approximately 45% of courses considered in analyses were core courses for a given major while the remainder considered were restricted electives. Class sizes ranged from 1-on-1 research opportunities to a maximum of 600 students per section, translating to 13 small, 25 medium and 24 large classes. Notably, first year was comprised of only large classes, while fourth year was comprised of primarily small and medium classes.

There was a total of 455 assessments offered, with an average 7±5 assessments per course, ranging from 2-27 assessments per course. The average number of assessments per course decreased from first year (16±8) to fourth year (6±3). Regarding assessment type, 63% of possible marks comprising the curriculum were assessed by tests, and 37% of possible marks comprising the curriculum were assessed by assignments. There was also variability in
assessment type between years of study, such that tests dominated the curriculum in years 1-3 and assignments dominated the curriculum in year 4. Specifically, first year was comprised of only 22% assignments (78% tests), and fourth year was comprised of 57% assignments (43% tests).

**Prevalence of authenticity**

The overall average authenticity score for all 62 courses was 1.8±0.6. The average authenticity score for core courses (1.6±0.3) was significantly less authentic than restricted electives (1.9±0.5). The dimensional authenticity scores for all 62 courses were similar on realism (1.9±0.7), cognitive challenge (2.0±0.6), and evaluative judgement criteria (1.9±0.5), while evaluative judgement feedback was significantly less authentic than all other dimensions (1.5±0.4, p<0.01) (figure 1).

Of all 455 assessments considered, only 4 assessments (0.9%) scored high across all 4 dimensions, while 49 assessments (11%) scored low across all 4 dimensions. When considering authenticity of an individual assessment, there was often variability in how an assessment scored on each of the 4 dimensions as 80% of assessments earned different levels of authenticity across the 4 dimensions. 25% of assessments (n=114) scored high for realism, which was commonly coupled with high cognitive challenge (n=84, 74% of the time). When realism was low (n=173, 38% of assessments), no other dimension scored high. Further, only 3% of assessments scored high for feedback, while 57% of assessments scored low for feedback.

Authenticity of assessments significantly increased across year of study, with fourth year being more authentic than first year (2.0±0.4 versus 1.5±0.3, p=0.011) (figure 2). This holds true for dimensional authenticity scores of realism (2.2±0.6 versus 1.5±0.7, p=0.012), cognitive
challenge (2.3±0.6 versus 1.6±0.4, p=0.013) and evaluative judgement feedback (1.7±0.4 versus 1.2±0.2, p=0.031) (figure 2). No significant differences were shown between years one to four on evaluative judgement criteria (p=0.93).

There were significant differences in authenticity scores between class sizes. The authenticity scores for small (S) classes were significantly higher than medium (M) and large (L) classes (S: 2.2±0.5, M: 1.8±0.3 versus 1.6±0.4, p<0.05) (figure 3). This pattern holds true for dimensional authenticity scores of realism (S: 2.5±0.7, M: 1.9±0.5, L: 1.5±0.5, p<0.05), cognitive challenge (S: 2.5±0.6, M: 2.0±0.4, L: 1.6±0.5, p<0.01) and evaluative judgement feedback (S: 1.9±0.4, M: 1.5±0.3, L: 1.3±0.4, p<0.01) (figure 3). There were no significant differences in dimensional authenticity scores between class sizes for evaluative judgement criteria. To determine if class size was simply an artifact for year of study, we also considered the impact of class size within a year of study. Specifically, in fourth year courses (i.e. the only year with small, medium and large courses), there was still a negative relationship between authentic assessment score and class size, such that as class size increased, authentic assessment score decreased (F(1,26)=6.8, p<0.05) with an $R^2$ of 0.2 (figure 4).

Assignments displayed significantly higher authenticity scores than tests (2.0±0.5 versus 1.5±0.4, p<0.001) (figure 5). This trend holds true for dimensional authenticity scores of realism (2.3±0.8 versus 1.5±0.6, p<0.001), cognitive challenge (2.3±0.7 versus 1.5±0.5, p<0.001), evaluative judgement criteria (1.9±0.7 versus 1.7±0.5, p<0.001) and evaluative feedback (1.6±0.6 versus 1.3±0.5, p<0.001) (figure 4).

**Discussion**

This research inventoried assessments across a complete undergraduate health science curriculum and subsequently determined the prevalence of authenticity among these using a four
We present a tangible process that can be adopted by others looking to review their curriculum (or course) from an authentic assessment lens, and improve upon existing authentic assessment frameworks to facilitate application to a variety of assessment.

Overall, our program had few examples of authentic assessments, although prevalence varied by year of study, class size (small, medium, large) and assessments offered (type, number), and even across individual dimensions.

**The authentic assessment tool can be used to score typical assessments across a variety of courses**

Our tool was applied to 455 assessments in 62 courses spanning multiple disciplines (e.g. biology, chemistry, physics, math, physiology, nutrition, pathology etc.). We were able to score all assessments using the tool, largely from publicly available course syllabi (i.e. only 16% of authentic assessment scores were unable to be scored confidently from syllabi alone). From syllabi, realism could often be determined from assessment descriptions (e.g. weekly quiz to check for reading completion, social media posts extending course work to a new audience), and learning outcomes linked to specific assessments were commonly used for determining cognitive challenge (e.g. apply core concepts of physiology to a disease, synthesize novel ideas through a grant proposal). While some course syllabi clearly explained criteria (e.g. included rubrics, implicit expectations, description of review sessions), many did not communicate this. Similarly, while some clearly explained feedback processes (e.g. peer-evaluation, multiple draft submissions), this dimension tended to be lacking description on syllabi. 68% of authentic assessment scores were confirmed accurate upon discussions with instructors while 32% of authentic assessment scores were clarified. Of those that were clarified, most commonly (75% of the time) evaluators had underscored the assessments. For example, post-discussion with
instructor, a score commonly changed from low to moderate.

When faculty were contacted for follow up, we were encouraged by their willingness to engage in conversation about their assessment. 91% (n=41) of instructors participated in semi-structured in person discussions to confirm and clarify authentic assessment scores. Instructors demonstrated interest in what others in the curriculum were doing and expressed interest to learn about the outcome of the study and receive personalized feedback about their assessments. One of the greatest, albeit unanticipated impacts of this work, may be the optimism that most instructors are open to discussing their assessment design with others, and were receptive to immediate feedback (e.g. one instructor had never considered providing their rubric to students before the oral presentation). Upon reflection, we believe that our strategy of going out to individual faculty with specific knowledge of their course already, rather than sending a group email inviting interested faculty to join a workshop with general information, may be an effective model to promote faculty buy-in and encourage ongoing course review, which may be of value to others.

Advantages of our tool over other authentic assessment frameworks

While other conceptualizations and frameworks of authentic assessment exist, we posit that the authentic assessment tool is advantageous in several ways. Firstly, the tool conceptualizes authentic assessment on four core dimensions, whereas the single dimension of realism often dominates other authentic assessment definitions. For example, Kaider and Hains-Wesson (2017) applied a realism-focussed framework to all assessments from a subset of courses (n=40) across four (unspecified) faculties according to Unit Guides and found that 43% of assessments across all faculties were authentic, with all courses offering some form of authentic assessment. Had we considered only realism in our framework, 25% of all assessments would be considered highly
authentic, when in fact our conclusion is that the prevalence of highly authentic assessments in this classroom-based curriculum is low (i.e. 0.9% (n=4) of assessments scored high across all 4 dimensions). The focus on realism is not surprising, given that 71% of articles included in a recent systematic review of authentic assessment literature focused on realism, while only 55% of articles focused on cognitive challenge and 38% of articles focused on evaluative judgement (Villaroel et al. 2018). However, by considering all 4 dimensions, our tool was able to identify several highly realistic assessments that could still benefit by increased opportunities for evaluative judgement. For instance, a fourth-year occupational biomechanics course (n=75) has a highly realistic and cognitively challenging assignment where student teams enact the role of a consulting company to assist local business partners in developing an ergonomics plan for reducing musculoskeletal injuries in their workplace. However, the assignment criteria are focussed largely on the mechanics of the task (length, format, etc.) and the final project is presented to the class, with the instructor being the sole source of feedback at the end of the project. Therefore, there is room to increase engagement with evaluative judgement criteria and feedback through avenues such as early discussion of latent criteria using a variety of student exemplars, inclusion of personal reflection or peer-feedback and/or inviting the community partner to provide ongoing informal feedback throughout the semester for the students to act upon.

In addition to considering dimensions of authentic assessment beyond realism, another advantage of our authentic assessment tool is that it allows for an assessment to score variably between dimensions. Notably, our results demonstrated that 80% of assessments earned different levels of authenticity across the 4 dimensions. In comparison, Bosco and Ferns (2014) framework grouped dimensions of authentic assessment together in a single cell of a rubric and
required multiple dimensions of authentic assessment to be satisfied at the same level for classification within the given cell. The levels were then stratified based on proximity to the workplace. Bosco and Ferns specifically used their framework to analyze work-integrated learning summative assessments in a curriculum across programs (e.g. pharmacy, biomedical sciences) and it was concluded that the biomedical sciences program offers an “adequate range” of “highly authentic” tasks in diverse settings, while pharmacy administers a “high number” of authentic assessment tasks in limited workplace settings, although there were no values associated with this study with which to directly compare the results of the present study. Had we used this approach with our classroom-based curriculum, an assessment scoring low on even one dimension would have restricted the assessment to the low classification, essentially masking differences between dimensions. Therefore, our tool allows for visualizing nuanced strengths and limitations of assessment design that in turn provides more tangible and focussed recommendations (e.g. feedback is the weakest dimension).

Finally, though other frameworks exist and have been applied to assessments within a curriculum, information about the context of the programs (e.g. size, year level, number of courses or assessments) is lacking. The present study is the first to openly report on the application of a comprehensive framework across a complete curriculum, with reference to specific context (i.e. inventory) by which to interpret authenticity values within. This cross-curricular approach is a key distinction of our work.

**Practical take-aways**

*Instructors should orient efforts towards more authentic feedback:* Only 2% of assessments scored high for evaluative judgement feedback, while 57% of assessments scored low for
evaluative judgement feedback. Further, the weighted authentic assessment score for evaluative
judgement feedback was significantly less authentic than weighted authentic assessment scores
for realism, cognitive challenge, and evaluative judgement criteria. In higher education, feedback
most often takes the form of a ‘delivery’ model, where feedback is done by academics and given
to students, limited to a grade and/or comments in an isolated event (Carless et al. 2011). As
such, educators often consider students as passive recipients of pieces of information in response
to a completed assignment (Carless et al. 2011). Indeed, 1-on-1 discussions with many
instructors confirmed this. Descriptions of feedback were often limited to a grader’s comments
on an assessment, perhaps a detailed rubric or grading key being returned, and/or a grade
returned to students. A more authentic view of feedback is that it be regarded as an iterative
process to foster evaluative judgement. In this way, students are active in gathering and
responding to feedback and have access to feedback throughout their studies (Carless et al.
2011). Therefore, to be highly authentic, there must be student engagement with feedback. This
was not often seen in our classroom-based curriculum, though exemplars did exist. For instance,
a large (n=>350) second year human physiology course presents students with a practice open-
ended long-answer test question prior to each midterm that models the thought processes and
connection of concepts that will be seen on the midterm. Students complete the practice question
individually, and submit to an online peer evaluation, assessment and review (PEAR) tool for
grading. Upon provision of a marking guide (which includes implicit criteria, common
misconceptions) each student will grade and provide feedback for 2 other student questions as
well as their own. This can contribute to students developing the skills to judge what makes
quality work, while limiting the resources required from the instructor. Further, this assessment
models timely feedback as students engage with criteria and expectations for the midterm exam
as they receive their grade and peer evaluation while studying for the midterm exam.

Importantly, some instructors in 1:1 discussion expressed frustration with experiences of feedback being time not well-spent, given that student participation during post-assessment review sessions is slim, and some graded assessments for return are not picked up by students. However, rather than the focus on provision of information to students post-assessment, we wish to highlight that this more authentic view of feedback can be incorporated into large courses successfully and manageably.

Efforts to include authentic assessment in large, introductory courses should be made: Weighted authentic assessment scores increased from years 1-4. This is not surprising given that in our context, first year offers exclusively large classes and more marks comprised of tests than assignments, whereas fourth year offers more small classes, and more marks comprised of assignments than tests. However, fourth year is comprised of primarily restricted electives, meaning that depending on course selections or caps on registrations (e.g. 1-on-1 research), some students may miss out on these authentic assessment opportunities. In our data, large class sizes were associated with lower authenticity scores (throughout the complete curriculum and fourth year specifically), and higher weighting of tests than assignments. This is consistent with the literature where large classes present greater challenges in fostering student engagement and deep learning (Cooper and Robinson 2000), and are typically associated with limited assessment schemes (midterms and final exams in multiple choice formats) (Cash et al. 2017). If programs value authentic assessment, consideration should be given to how authentic assessment can be incorporated into core, early year, large classes, even though it may seem easier to do so in small, upper year classes. In our curriculum, an intentional re-design of first year Biology nearly a decade ago introduced a novel assignment which incorporates authentic elements. Specifically,
3 large (n=>600 students/course) core first-year biology courses (genetics, health, environment) present students with an Interdisciplinary Project where students from each course are placed in groups of three to address a single real-world challenge (e.g. the flu, global obesity, sustainability through pollination) from their respective biological perspectives. Students brainstorm and integrate concepts to present a cohesive conclusion during a public-facing poster showcase. Students are provided with comprehensive rubrics and examples of previous projects at various levels, and receive peer-feedback, comments from attendees as well as detailed feedback from their graduate teaching assistant and peers. Although this structure requires considerable co-ordination, it ensures that all incoming students are exposed to an assignment with high realism early in their program of study.

Curriculum-wide knowledge can promote reflection on assessments: This research provides a tangible process to inventory and analyze the authenticity of assessments within the context of a primarily classroom-based undergraduate science curriculum. This information can be used by individual instructors wishing to reflect on their own assessment strategies, but importantly, can also be applied by curriculum committees across an institution to better recognize redundancies or gaps in scaffolding that may be impacting students’ performance at key bottleneck areas. We highlight the general finding of assignments being more authentic than tests, but also report limited exposure to assignments in early years to prepare students for this dominant assessment format in fourth year. This gap may help instructors of fourth year courses better appreciate the need for clear instruction including implicit expectations of quality since students may still be relatively novice in the task at hand. Alternatively, some early instructors may consider moving away from weekly comprehension quizzes and introducing low stakes, practical assignments to keep students engaged and excited about the application of course material beyond academia.
However, it should be noted that highly authentic assessments may not be appropriate for every learning context, such as earlier years of study when students are not yet comfortable collaborating or working on ill-defined problems (Herrington and Herrington 2006). Further, in instances where mastery of terminology and foundational concepts is critical, less authentic assessments (e.g. lower order multiple choice tests) may be most appropriate and effective. As such, it is critical to recognize that authenticity occurs on a continuum, and that moving towards authenticity is the underlying goal (Sokhanvar, Salehi, and Sokhanvar 2021). The authors also contend that it should be cautioned that more authentic may not necessarily result in improved learning. While authentic assessment is often shared as a best practice in assessment design, evidence which has evaluated its impact on student learning as a whole is lacking. Many of the elements of authentic assessment individually are associated with improved student outcomes, but there is a need to further investigate the impact of authentic assessment and the relative importance of the specific dimensions on student learning through future case studies and experimental work.

**Future Directions**

Future directions for this research may consider student perceptions of authenticity. Since assessment ultimately drives student learning, defining what students regard as important in a course and thus how they will subsequently direct their efforts is critical (Brown 1997). While we consulted 1-on-1 with instructors and their course documents in determining the prevalence of authentic assessment in a health sciences curriculum, research suggests that instructors may perceive authenticity differently than students (Gulikers, Bastiaens, and Kirschner 2004). Additionally, students may perceive the relative of importance of authentic assessment
dimensions of authenticity differently (e.g. realism versus cognitive challenge), as was seen in
the authentic assessment framework by Gulikers, Bastiaens, and Kirschner (2004) (e.g. physical
context versus social context). Therefore, to gain a more comprehensive view of authentic
assessment in the curriculum, more research is required regarding how students perceive
authenticity with regard to the authentic assessment dimensions as outlined by Villaroel et al.
(2018, 2019)

While the present research was conducted in the context of face-to-face (F2F) learning,
there has been a recent shift to emergency remote teaching (ERT) due to COVID-19. As such,
instructors have faced a common impetus for change and subsequently have reconsidered their
courses and assessment design (Bearman et al. 2017). Research will continue to investigate how
assessments may have changed (number, type, authenticity) in ERT compared to the F2F setting
as presented in the current study, utilizing similar methodologies and analyses.

Lastly, this work was conducted in health science majors. It is worth noting that the
obsession with getting good grades seen in many higher education students may be exacerbated
in this particular student group by goals to continue on to professional school (e.g. veterinary,
medical, physiotherapy) as several instructors acknowledged this fact when explaining their
grading criteria and feedback strategies, and even in designing tests. Therefore, we intend to
apply the described research process to a broader range of disciplines with a greater variety of
typical assessments to increase the generalizability of these results.

Conclusion

To offer a learning experience most effectively and comprehensively to students through
assessment design, we have taken a cross-curricular approach to better understand the curriculum
as experienced by students (i.e. beyond one course). Using a literature-informed authentic assessment tool, we inventoried all 455 assessments across 62 courses in three primarily classroom-based health sciences majors in a Canadian comprehensive university. Our results suggest that instructors do not intuitively design classroom-based assessments to be authentic across the dimensions of realism, cognitive challenge and evaluative judgement. Notably, feedback was the weakest dimension, as it was most often passively provided to students rather than requiring meaningful student engagement. Our broadly representative program provides evidence-based rationale to focus training and resources to improve feedback methods and consider replacing tests with assignments, particularly in early years of study where class sizes tend to be large. Assessment exemplars provided may encourage readers to reflect on their own assessment strategies, serving as a guide for varying authentic experiences in a classroom-based curriculum. Overall, this research provides a tangible tool and clear process to inventory and analyze the authenticity of assessments in accordance with core dimensions, which can be used by instructors, curriculum committees and institutional review. Future research may consider the student voice in determining the prevalence of authenticity, and contexts such as remote teaching and those beyond the sciences.

**Disclosure Statement**

No potential conflict of interest was reported by the authors.
References


Table 1. Authentic assessment tool.

Table 2. Assessment Inventory.

Figure 1. Overall authentic assessment score and dimensional authenticity scores for the complete curriculum. To account for different weightings of assessments within a course, each individual assessment score was multiplied by its respective weighting before contributing to the overall authentic assessment score. Dimensions not sharing a letter are significantly different from each other ($p<0.01$).

Figure 2. Overall authentic assessment score and dimensional authenticity scores for the complete curriculum by year of study. Within the overall curriculum and each dimension, years of study not sharing a letter are significantly different from each other ($p<0.01$).

Figure 3. Overall authentic assessment score and dimensional authenticity scores for the complete curriculum by class size. Within the overall curriculum and each dimension, class sizes not sharing a letter are significantly different from each other ($p<0.05$).

Figure 4. Overall authentic assessment score for fourth year courses by class size. As class size increases, authenticity decreases ($F(1,26)=6.8$, $p<0.05$), $R^2$ of 0.2.

Figure 5. Overall authentic assessment score and dimensional authenticity scores for the complete curriculum by assessment type (test, assignment). Overall and by each dimension, assignments were significantly more authentic than tests ($p<0.001$).