Comparison of Standardized Assessment Methods Logistics, Costs, Incentives and Use of Data

Simper, Natalie; Frank, Brian; Kaupp, Jake ; Mulligan, Nerissa; Scott, Jill

Published in:
Assessment and Evaluation in Higher Education

DOI:
10.1080/02602938.2018.1533519

Comparison of Standardized Assessment Methods: Logistics, Costs, Incentives and Use of Data

*Natalie Simper[1], Brian Frank[2], Jake Kaupp[3], Nerissa Mulligan[2], Jill Scott[1]
[1] Office of the Provost,
[2] Faculty of Engineering and Applied Science,
[3] Office of Institutional Research and Planning,
Queen's University,

Keywords:
Assessment, Standardized test, Motivation, Incentive, Cost

Critical thinking, problem solving and communication are fundamental elements of undergraduate education, but methods for assessing these skills across an institution are susceptible to logistical, motivational, and financial issues. Queen's University conducted two research studies investigating the use of standardized tests to assess cognitive skill development across the institution. Synthesis of results from implementing the Collegiate Learning Assessment, the Critical Thinking Assessment Test, and the HEIghten test found that student test effort was a significant factor, and effort level correlated with performance at $r = .33$. Test incentives were also a significant factor; effort levels for the $25 financial incentive group were one standard deviation higher than effort for the in-class test group. A dedicated computer lab was the preferred option for computer-based testing. A paper-based test was found to be much simpler to administer, but test results were not available for a long time, therefore limiting the usefulness of data. The true cost of tests was greater than the price of the instrument; recruitment, training, proctoring, and marking costs need to be included in the calculation. Generally speaking, alignment of test objectives with student or course objectives, and timeliness of data, were key for participation and motivation.

Introduction

Higher education sectors have requirements for performance metrics, and over the past decade the "skills gap", driven by labour market forces, has identified the need for the development of 21st century skills (Johnson, 2009). Critical thinking, problem-solving and communication are fundamental elements of undergraduate education (Fadel, 2008; Liu, Frankel, & Roohr, 2014), and students consider them to be among the top three most important skills in undergraduate education (Canadian University Survey Consortium, 2015). It has long been argued that assessment drives learning (Boud, 1990; Tait, 2005; Boud & Falchikov, 2006). Therefore, it makes sense to know how well students in higher education are able to perform, and whether an undergraduate program is actually building these skills. The challenge however, is gauging the performance of these skills, such that it is possible to determine whether higher education institutions are adding to the development of them in their students. The Higher Education Quality Council of Ontario has been supporting universities in Ontario to investigate methods for assessing skills.

There are two main purposes for assessment, firstly, to improve student learning and secondly, to certify student achievement (Boud, 1995). In many countries, there are increasing demands

for consistent approaches when it comes to certification (Tertiary Education Quality and Standards Agency, 2017). In other countries, deciding on an assessment focus is a matter for strategic planning, and will depend on the needs and context of the institution. If leaders select specific skills as a focus for institutional assessment, then they need to formulate a method and comparable metric such that data can be aggregated. For this purpose, higher education institutions have turned to either standardized tests or institutional-level assessment rubrics, such as the Valid Assessment of Learning in Higher Education (VALUE) rubrics (Rhodes & Finley, 2013). Neither of these approaches is easy to implement, and depending on the scale, can be quite costly.

Standardized tests are generally administered to large cohorts, where all test takers answer the same questions or a selection of questions from a consistent question bank. They are scored in a consistent way to ensure comparability between individual test-takers. Historically, the use of standardized tests in the United States "reflects two fundamentally American beliefs about the organization and allocation of educational opportunities: fairness and efficiency" (Brown, 1992, p. 103). Standardized tests of academic ability were widely introduced in Elementary and Secondary schools in the United States as part of an educational reform movement following the No Child Left Behind Act of 2001. Standardized tests have been used as an accountability measure, to evaluate educational standards, gaps in performance, differences between student populations, and by extension to evaluate whether educational policy and funding were meeting the needs of the students. There has been an extensive debate, however, about the effectiveness of testing in high-stakes environments in achieving the desired goals (Hutchings, 2015). As Deming & Figlio (2016) point out, "the problem with high-stakes accountability is that the objective metrics are typically incomplete descriptions of performance" (p. 38).

Test instruments are not the only means to assess these skills across multiple courses or departments. Many institutions in the United States use the VALUE rubrics (Rhodes & Finley, 2013) to assess course-based student work samples. The VALUE rubrics, however, were not developed as grading rubrics. They were "designed to be used at the institutional or programmatic levels in order to assess student learning overall and over time, not for specific assignments" (Rhodes & Finley, 2013, p. 6), so require expertise for grading. In addition, the grading of assignments that were not specifically aligned with assessment constructs present a challenge to the validity of their use.

If standardized testing is chosen (perhaps in addition to other strategies) as a pathway for institutional assessment, a decision needs to be made as to which instrument should be used. There are a number of tests designed to evaluate critical thinking, problem solving and communication, but an initial review of existing literature suggests that there was a lack of comparative data to inform such decision-making or weigh strengths of different instruments. The assessment constructs for three of the more widely used standardized tests currently available for assessment of the above-mentioned skills are displayed in Table 1.

Collegiate Learning Assessment (CLA+)

The Collegiate Learning Assessment (CLA+) was developed by the Council for Aid to Education (CAE) (S. P. Klein, Kuh, Chun, Hamilton, & Shavelson, 2005). It is a proctored web-based test including performance task and selected response questions administered through a secure browser that takes approximately 90 minutes to complete. The test prompts students to engage with real-world issues and come up with a solution or recommendation. The

performance task question bank contains five questions; each student is allocated a question at random. The selected response questions target scientific and quantitative reasoning, critical reading and evaluation and critiquing an argument. The CLA+ provides benchmarks for achievement and the data file for the test is available approximately two months after the test window closes.

Critical Thinking Assessment Test (CAT)

The CAT is a paper-based test designed to assess multiple areas of critical thinking and problem solving by engaging students in real-world problems. There are two different forms of the test, both aimed at STEM education (Science, Technology, Engineering and Mathematics) (Stein, Haynes, & Redding, 2016). The test comprises short answer responses that are scored on campus, using an assessment protocol facilitated by trained markers. After local scoring, the tests are returned to the test designers to substantiate marker reliabilities by scoring a random sample. The CAT is marketed as a catalyst for change, in a model where test marking is conducted as a faculty development activity.

HEIghten™ Critical Thinking Test

HEIghten is a 45-minute online test developed by the Educational Testing Service (ETS) designed for institutional assessment of critical thinking (Liu, Mao, Frankel, & Xu, 2016). The test is composed of a series of questions based on a shared multi-part stimulus that reflects real-world issues. Students are expected to analyze and evaluate an argument structure, and evaluate evidence and its use to develop a valid or sound argument. Two sub scores are reported, Analytical, and Synthetic, which combine to form a Total Score. Individual test scores are available immediately through the proctor portal, with a data download available within a matter of weeks. At the end of HEIghten, there is an exit survey, one of the questions asks about the amount of effort they put in to the test.

Table 1. Assessment constructs for each of the tests

Instrument	Assessment construct			
	Critical thinking	Problem solving	Written communication	Other
Collegiate Learning Assessment (CLA+)	Critical reading and evaluation	Analysis and problem solving	Writing mechanics	Scientific and quantitative reasoning
	Critique an argument		Writing effectiveness	
Critical Thinking Assessment Test (CAT)	Evaluation and interpretation of information	Problem solving	Effective communication	Creative thinking
HEIghten	Analysis			
	Synthesis			

Comparing test instruments

Continuous improvement, accreditation and accountability requirements mean that institutions are looking for practical, cost-efficient methods to evaluate student skill development. Referring to studies investigating the implementation of individual tests is of limited help for institutional decision-making, because without direct comparison, inferences need to be made as to the relative strengths. Further exacerbating the problem is that most of the reliability and validation studies are conducted by the test developers. As these test developers work toward commercial objectives, it can somewhat obscure an impartial, objective perspective.
As mentioned, there has been limited research on direct comparison between these methods for standardized assessment. Methods for assessing these skills across an institution are susceptible to logistical and motivational issues (Liu, Bridgeman, & Adler, 2012). As Klein, Liu, & Sconing, (2009) point out, it is difficult to directly compare quantitative test data derived without an experimental research design, but it can still be valuable to evaluate lessons learned from the logistical and methodological outcomes perspective.

The Higher Education Quality Council of Ontario (HEQCO) provides funding as a part of "a system-wide commitment to measurement of learning outcomes… an opportunity for the Ontario system to show worldwide leadership" ("Learning Outcomes," n.d.). There is currently no legislated requirement in Ontario for assessment metrics in higher education, but there are in other places in the world. Learning outcome research initiatives are funded by HEQCO in an effort toward tackling some of the above challenges. If institutional-wide assessment were a mandated requirement, institutional leadership would like to be as informed as possible, such that the best decisions can be made for the institution. The following sections synthesize results from studies conducted comparing the logistics, costs, and utility of data derived from the CLA+, the CAT and HEIghten testing. The underlying purpose of the research is to be ready for potential changes in legislation, and to understand how such testing might help instructors be more aware of the relative strengths of their student populations.

Method

Researchers at Queen's University conducted two studies using a range of measures to evaluate cognitive skill development across the institution. The first research study utilized a longitudinal methodology implementing the Collegiate Learning Assessment, the Critical Thinking Assessment Test as well as the Valid Assessment of Learning in Undergraduate Education (VALUE) rubrics. The results from the VALUE rubric assessment are included in the report provided to the funding body (Simper, Frank, Scott, & Kaupp, 2018). The research questions primarily focused on the reliability of the tools to demonstrate the growth in skills between the beginning and the end of undergraduate education, but also asked whether data from instruments could be used to support skills development in courses. It was for the secondary reason that many of the course instructors agreed to be involved in the study. In the first study, participants were randomly allocated to the CAT or CLA+ in their first-year, and followed through the four years of their undergraduate education. Every attempt was made to match students to the same test in subsequent years, but due to logistical and ethical constraints, this was not always possible.

The second study collected cross-sectional student data from the HEIghten critical thinking test (Pichette, n.d.). This second study is ongoing, with ethical consent obtained to investigate achievement of critical thinking with future employment data. Between the two studies, there were multiple goals. While this paper compares the feasibility, and logistical concerns of the

use of standardized tests, it also touches on the utility of data for evidence-based decision making for curriculum improvement. Individual ethical consent was obtained for student participation in each year of the studies. The longitudinal study began in 2013 and the cross-sectional study began in 2016, so all of the final fourth-year testing was conducted in winter of 2017.

Implementing the tests

Within the two studies, there were two different testing contexts - in-class or out-of-class and four different incentive conditions as shown in Table 2. The first and fourth-year sample comprised students from 20 disciplines enrolled in a Bachelor of Arts degree program, 18 disciplines from Bachelor of Science, two from Health Sciences, three disciplines from the Bachelor of Computer Science, and 10 engineering disciplines from the Bachelor of Applied Science. Students in the second and third year were sampled in a range of engineering, psychology and physics courses. It was far easier to recruit participants from the large first-year courses that were involved in the studies. The first-year students were tested in lab or tutorial sessions in groups of between 40-80 students to a session. As the longitudinal testing went on, it became increasingly difficult to access students. The upper-year classes are smaller, therefore necessitating negotiations with a greater number of instructors for inclusion of testing. By the fourth year, the majority of testing was conducted outside class time. Fourth-year students were recruited by email, and test sessions had between 5-30 participants, scheduled at times that aligned with student availability. Invitations were sent to students for one of the selected tests, and students who attended a test session were subsequently invited to take an additional test. In the fourth-year test sample, there were 43 students who completed CAT and HEIghten, and 76 students who completed both the CLA+ and HEIghten tests.

Table 2. Testing context and sample sizes

Test Context: Incentive	First-year			Second-year		Third-year		Fourth-year		
	CLA	CAT	*HEI	CLA	CAT	CLA	CAT	CLA	CAT	*HEI
In-class: no incentive	326	105			8	84	25			
In-class: 1% attendance	252	165	1018	294	444	31	72			
In-class: 3% attendance							27			
In-class: food incentive						72	80			48
Out-of-class: food incentive						10	32			
Out-of-class: $25 incentive								139	139	285
Total	578	270	1018	294	452	197	236	139	139	333

*HEI=HEIghten

Timing of the test

In-class testing had to be conducted according to course schedules, so test sessions were occasionally conducted at sub-optimal times of the day or days of the week. For example, there was no other option but to conduct some test-sessions as late as 8:00pm; students can be tired after a long academic day. The Friday afternoon time-slot also proved to be less than ideal, with social or work activity conflicts. Arranging the implementation dates for the CLA+ was dependent on the available test-window. For example, the Spring test-window opens in March which is a poor time for Queen's University as mid-term course examinations take place in March. Researchers found that it was preferable to run the testing earlier in the term, which was when the CAT and HEIghten tests were conducted. Many of the participating courses offered their one-hour lab or tutorial slot for testing, which prevented the use of the 90-minute CLA+ in those classes. Researchers initially recruited students to test outside of class time, but attendance at out-of-class testing was very low, and sample bias was a concern. The only viable solution for a one-hour session was to run either the CAT or HEIghten.

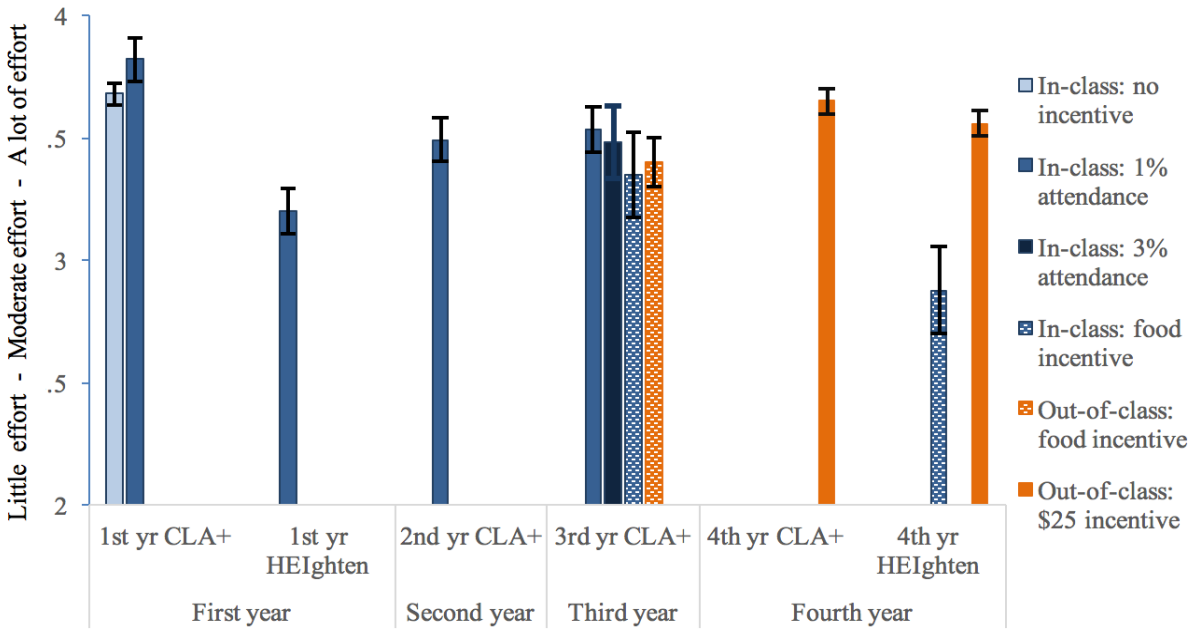
Technical issues for online tests (CLA+ and HEIghten)

Students who were scheduled for in-class testing as part of a course that was not in a computing facility were asked to bring their own laptop or required to move to an alternate environment. Moving proved to be time consuming and disruptive. In one first-year test session, researchers provided laptops but many were slow and some had connectivity issues which were also time-consuming. Some of the students tested in a lab with virtual machines (computer sessions that use a local monitor to connect to a session running on a central server) which were also problematic. The virtual machines prompted a security conflict, and a security bypass protocol had to be established by the test provider to access the test from these machines. Many of the students used their own laptops for testing, presenting a range of technical issues. Moreover, the HEIghten test did not run on the Firefox browser, or on tablet devices. There were issues with enabling "pop-ups" required by the test, but the main technical issue however related to anti-virus software on students' personal computers. With anti-virus programs running in the background, the test-browser (for both of the online tests) would not launch correctly. The majority of fourth-year test sessions were conducted in a stand-alone computer lab, presenting far fewer technical issues. The CAT is a paper-based test, as such it does not present similar technical issues. The test booklets were ordered from Tennessee ahead of time, which allows for administration without the need for a computer, connectivity or a particular web browser.

Effort and incentives

Effort in the CLA+ and HEIghten tests was reported on a five-point Likert scale from "no effort at all' to 'my best effort'. For the CLA+, effort is reported separately for the performance task and the selected response questions. There is no option for an effort question on the CAT exit survey, so effort scores are not known for the CAT. Figure 1 displays the effort scores for the CLA+ performance task, and HEIghten test for each of the testing contexts and incentive conditions. The blue bars display the in-class test effort, and orange bars the out of class test effort.

Figure 1. Effort scores for testing contexts and incentive conditions (bars represent mean error)



There were 76 fourth-year students who completed both the HEIghten and CLA+ tests. To examine the consistency of reporting, Cronbach's alpha was calculated, based on the effort scores that these students reported for each of the tests. An alpha of .75 suggested that there was a high level of consistency, I.e. those who put effort into HEIghten, generally put similar effort into the CLA+. There were however differing levels of effort in the first-year samples. It needs to be recognized that the first-year HEIghten test-takers (n= 1018) were not the same students as the CLA+ test-takers (n= 578). The first-year mean effort on the HEIghten was 3.19 (SD .87), compared with 3.74 (SD .83) for the CLA+. Researchers speculated about the effect of the test length and complexity. HEIghten is approximately half the length of the CLA+, and perhaps the first-year students found the CLA+ more challenging and therefore, required more effort. Effort in the CLA+ dropped over the first three years of implementation, following the first-year mean of 3.74, the second-year mean was 3.49 and the third-year mean 3.45. This prompted the use of a monetary incentive for all of the fourth-year testing. The fourth-year mean effort for financially incentivized testing was 3.56.

Results

Results from the first-year testing suggested the tests may have required differing levels of effort, so to further investigate the effect of incentive, the fourth-year HEIghten test effort results were selected for analysis. The majority of fourth-year testing was conducted outside class-time, with a $25 cash incentive. There were 285 students who completed HEIghten and received the cash incentive. There was a comparison group of 48 students completed HEIghten during their regular class time, and pizza was offered to all the students at the end of the test. The effort question was not compulsory; there were four students in each group who did not report effort. There were no consequences for the result for either group, yet the students who received the cash incentive put significantly more effort into the test ($F(1,324)= 42,67$ $p= .001$), partial $\eta^2= .117$. The mean score for effort in the fourth-year HEIghten in-class test was 2.73, compared with an effort mean of 3.61 for the financial incentive group. The effect size

calculated using Cohen's d = 1.01 (M2-M1/ pooled standard deviation), indicated that there was a standard deviation difference in effort between the two groups. Pearson's correlations were calculated between the effort reported and the respective test scores. Reinforcing previous research suggesting that effort matters (Liu et al., 2012), we found a significant correlation between effort and score for both tests, CLA+ $r(1069)= .34$ $p< .001$, and HEIghten $r(1338)= .33$ $p< .001$ (as previously mentioned, the majority of, but not all students completed the post-test survey).

Costs

The costs for administering the tests included personnel time for organizational matters, recruitment, room bookings, technical support and system checks, and two test proctors for each test session. Researchers found it difficult to handle technical and login issues at the beginning of the test in sessions with more than 30 students. Room setup for testing needed to be done ahead of time, especially if there is research consent to be conducted also. Table 3 shows total administration cost in Canadian dollars, per 100 students, tested over four test sessions with an average of 25 students per test session. The tests were paid for in US$, so the falling exchange rate between Canada and the USA meant increasing costs for Canadian institutions. The average value of the Canadian dollar over the duration of these studies was approximately 0.8 US dollars.

Utility of data

Data from the CLA+ and CAT tests are not available for several months after testing. Delays in receiving data make it impossible to use the test score as part of a (low-stakes) course assessment, and also make it difficult to use the data for course improvement. The CLA and HEIghten test portals both allow proctors to view student progress throughout the test, but only HEIghten provides students with their report at the conclusion of the test. Moreover, the CLA+ includes a digital badging system for students, with reports sent directly to students at roughly the same time as the CLA+ institutional report is sent to the institution. The HEIghten test assesses fewer assessment areas than the other two tests. It provides sub-scores for analysis and synthesis; if institutions are looking to measure problem-solving or written communication, these skills are not reportable through this test.

Table 3. Comparative test costs

Cost	CLA+	CAT	HEIghten
Marker training/ on-site tech support	$100.00	$1,100.00	$100.00
Instrument fee (per 100 tests)	$4,200.00	$1,194.00	$1,500.00
Test proctoring (4 test-sessions)	$800.00	$560.00	$960.00
Marking costs (per 100 tests)	$-	$2,000.00	$-
TOTAL (per 100 sample)	$5,100.00	$4,854.00	$2,560.00
Cost per student	$51.00	$48.54	$25.60

CAT tests are required to be marked using a rigid marking protocol which necessitates trained markers with available time. Once the tests are marked, they need to be sent back to Tennessee for marking reliability checks. The CAT data file provides question-level breakdowns, and a total score, with guidelines for which questions address each of the assessment constructs, but no sub-scores are provided for the dimensions as shown on Table 1. Data from the tests are confidential and all three test providers have security protocols in place to ensure that the data are not inappropriately accessed. These tests provide data that can be aggregated across disciplines and used for comparisons to institutional averages, too.

Discussion

While recognizing that there are many ways of evaluating student learning, and there are challenges involved in any assessment, this paper focuses on the implementation of standardized tests to evaluate student learning across the institution.

Rationale for standardized testing

There have been arguments against the use of standardized testing in higher education. As summed up by a report for the Association of American Colleges and Universities, "most stakeholders in higher education are deeply interested in improving student learning outcomes, [but] there is disagreement about whether or to what extent standardized testing programs will contribute to this goal" (Miller, 2001, p. 3). However, assessment of student learning through course-based measures can be problematic for many reasons, for example there may be issues with alignment to high-level outcomes, necessary for data aggregation, or course-based tests may not be designed to assess the desirable constructs. As Coates (2015) points out, the course-based assessment can be 'perfectly aligned with an academic's curriculum and teaching, yet fail to contribute to the qualification-level information required for external professional accreditation' (p. 340). In addition, "course-specific tests also have glaring weaknesses, not only because they are often too low level and content heavy. They are rarely designed to be authentic tests of intellectual ability" (Wiggins, 1991, p. 345). Institutional desire to improve the quality of higher education means that potential arguments against standardized testing have been set aside in favour of increased adoption in higher education (Cumming & Ewell, 2017). The goal of any assessment is to evaluate the true ability of a student. With standardized testing, the true ability of students may be confounded by student's effort, or their incentive for taking the test and the results may be further complicated by technological or logistical problem Therefore, the most reliable, validated tests 'are meaningless unless they are feasible to implement' (Coates, 2018, p. 46).

Working with instructors to embed testing

One of the obstacles for the research team was overcoming instructor resistance. It was necessary to have senior leadership support for the institutional research projects (to secure funding, for research ethics and to ease facilitation of the initiatives), but this resulted in what many thought a "top-down" approach. There were instructors who objected to testing as an intrusion to their teaching, and some pushed-back simply as an assertion that they would not be told what to do. The initial group of instructors for the first study were a dedicated few who had been part of the discussions about the research. Beginning with these courses, the research team collected assessment data, and prepared reports detailing the achievement of their students. We found that inviting the second-year instructors to the debrief meetings was an effective strategy to recruit them in participating the following year. At these debrief meetings,

the results were shared, the opportunity was provided for discussion, and questions were answered. Another successful strategy was to appeal to the scientific nature of reluctant instructors. For example, there was an instructor who argued that standardized tests would not be an effective mechanism to quantify learning in their department. The researchers queried on what basis that judgement was being made, and explained that taking part in the study might provide them with the data to support their argument.

Implementation, timing and technical issues

The majority of standardized testing was embedded in the course environment, with students given the option to have their data included as part of the research study. The test environment, and available time for taking the test were important factors. Keeping the students in the room they were scheduled for meant that the students were comfortable with the environment, and time was not lost at the beginning of the lesson as students navigated to a new space. It was also preferable, because out of class testing required incentive for the students to attend and put effort in, and financial incentives add to the expense of the initiative. Moreover, keeping students in the scheduled room meant students needed to use their own laptop/computers for the online test if the room was not a computer lab. The CLA+ and HEIghten both have system check that could be run ahead of testing, but it is not feasible to run these checks on hundreds of different laptops. It also meant that the time available was fixed, so traditional 50-minute time slots ruled out the use of the 90 min CLA+.

Instructors needed to be on-board to run testing in courses. Instructor buy-in was fostered with high-level institutional support and ongoing instructor consultation, but not all of the instructors were agreeable to standardized testing. Some instructors supported the need for investigating test instruments but could not spare the class time for testing, and other instructors refused any involvement at all. Some of this resistance was because the standardized tests do not specifically contain content directed at a particular discipline. This factor can also be a detriment to student motivation. Additionally, since the CAT test questions are specifically tailored to STEM fields, these may discourage students from the Humanities or Arts. Data reports were prepared for all of the instructors involved, and debrief meeting were arranged to discuss student achievement on the aspects of student performance. Where welcomed, this debrief was extended to committees in departments, providing the opportunity to discuss assessment, and areas of student achievement at a wider level.

Incentives

First-year (freshmen) students who tested were rewarded with a course percentage mark. We found the first-year students to be very keen test-takers, but interest waned in second and third-year, as was reflected by the dropping effort levels being reported. As effort was significantly correlated with the test score, researchers looked for methods to incentivize fourth-year students. We were going into the fourth-year of the longitudinal study, and were conducting an additional cross-sectional study, so a significant investment of time and resources had been made. The prospect of either not recruiting a viable sample of fourth-year test-takers, or deriving scores that did not accurately represent the skill capacity of students, was a threat to the research. The decision was made to pay students to test and the financial incentives did yield improved effort with effort levels almost at the first-year level as shown in Figure 1. This was reassuring for the research, but the additional costs over and above those presented in Table 3 will likely be prohibitive for long-term sustainability.

Costs

Gathering reliable and valid data on student achievement of learning outcomes can be quite costly, and so it is an important part of any large assessment project to keep track of expenditures and ensure that activities and instruments are evaluated for their relative benefit. Throughout this project, we tracked the full cost of the various instruments, including both direct and in-kind costs, and found that the CLA+ was the most expensive test. If student incentives are needed, it further increases the cost. The costs presented in Table 3 do not include the $25 per student incentive; the fee for the CLA+ was $5,100 for a sample of 100, adding the $25 incentive makes a total cost of $7,600. Our findings indicate that course-embedded assessments provide the greatest value for money, and ensures that sample bias does not play a part in the resulting data because all students in the particular course take the test, not just those who opt-in for financial incentive. This becomes more challenging in upper years as the relatively few large courses required working with a large number of course instructors.

Finally, it is important to consider how many students to test; if every test costs $30-$50CAD, then it makes sense to test enough students to have a random, representative sample but not more. For a representative sample, and alleviate any potential concerns around selection bias, an institution would need to test more than 100 students. Based on the financial incentive condition, the cost for 500 students would be $38,000; testing 1000 students would require $76,000. It is also important to remember that these amounts do not include any overhead or administration costs. For this reason, it is essential to ensure the value of the resulting data. It is very likely that a variety of stakeholders, including government, students, parents and administrators will continue to be interested in data demonstrating student achievement of learning, and in order to ensure that assessment initiatives are sustainable over time, tracking costs and finding efficiencies will be an ongoing part of continuous improvement.

Utility of data

Jonson, Guetterman, & Thompson (2014), present a framework for the possible effects of assessment, including instrumental (an action taken impacting practice or policy), conceptual/cognitive (new understandings or ways of thinking), affect (impact on a person's disposition or beliefs regarding assessment), and affirmation (confirming appropriateness of existing practice). Instrumental effects, i.e. practice or policy changes, are the most commonly expected but infrequently observed effect. Blaich & Wise, (2011) found that after conducting institutional learning outcome assessment activities, "most Wabash Study institutions have had difficulty identifying and implementing changes in response to study data" (p. 3). This speaks to not only both a need for patience and recognition of the other effects of assessment, but also a need to think carefully about the utility of data emerging from assessment.

The most rapid instrumental effects occur when the course instructor uses the data to make a change in the course due to assessment data. This can arise from both assessment data from course activities (Salem & Frank, 2017) or use of standardized tests. In the first study, several instructors observed gaps in student knowledge that allowed them to make rapid but relatively small changes in their course. Slower but more widespread changes can occur when the data is used by a department, faculty, or institution as a result of assessment. Assessment played an influential role in significant faculty-wide curriculum change at this institution prior to this study (Frank & Strong, 2010; Frank, 2013), and during this study influenced a significant curriculum revision in one department.

The results of CAT testing in a large undergraduate first-year course suggested that the students were struggling with *identifying additional information to evaluate a hypothesis*. The cohort breakdown in the CAT Institutional report, suggested that the overall effect size, compared with the US National Lower division average for that group, was +.37 (mean difference divided by pooled standard deviation), whereas the effect size for the *identifying additional information* question was -.85. These results were concerning for the department, so steps were taken to support student's development of critical thinking. A new critical thinking activity was developed for a large second-year course in the department which had not previously assessed critical thinking. The second-year activity was designed as an authentic assessment of critical thinking (Chun, 2010). Testing continued, and discernible improvement was made. By the third-year, the overall effect size for the cohort was +1.49, and there was no significant difference between the US National average for Upper Division on the *identifying additional information* question for the cohort in question. While it should be recognized that the student sample differed from year-to-year with annual consent, and there was a possibility that the sample became more selective over the years (with program and course selections), the department was still spurred by the results. In the year following the study, the first-year course instructor introduced a new three-step critical-thinking activity aimed at building to conceptual understanding. The activity included an individual pre-lab example with video debrief modeling desirable thinking, an example worked on through class-based group-work, with facilitated discussion, and a final individual example.

It is suggested that lack of alignment is a common problem with using standardized tests (Coates, 2018), yet the above example provides evidence assessment data being used to inform changes in practice. This happened because the data were meaningful to the course instructor and the test constructs were somewhat aligned with program outcomes. Valid conclusions require representative participation in the test, which requires student willingness to both participate and put in meaningful effort. This can be influenced by a course instructor's perception of the test's alignment to their own course, and the student's perception of test value. As reported in Simper et al., (2018), students suggested that while food may be an incentive to some, monetary gifts may be more effective, and professor's positive approach towards the test would have been a great source of motivation to many.

Limitations

The results presented here were based on data from a medium-sized research-intensive university, and vary in alternate contexts. The itemized costs were presented in Canadian dollars and specific rates, based on the needs of the institution for the implementation of testing. As such are intended to provide an indication of financial commitments, and not a literal sense. There may be specific needs for alternative contexts that are not captured here.

Conclusions

Test implementation of the CLA+, CAT and HEIghten were investigated and the HEIghten test was found to be the most useful, cheapest and feasibly practical option for our institution's goals. The CLA+ had the most comprehensive data set, but delays in the return of data files limited the usefulness of the test for course improvement. The CAT test was the only test where pre-post test condition could be controlled, but it primarily targets students in STEM fields. Running testing avoids potential sample bias, but the level of effort that students put into the test had a significant effect on scores, reinforcing past work showing student motivation is a key consideration for implementing testing. Key conclusions include:

1. Computer labs with dedicated computers present far fewer problems for in-class computer-based testing than using standard classrooms where students bring laptops.
2. Paper-based tests are much simpler to administer than computer-based tests, but test results may not be available until long after the course is over.
3. The true cost of tests is made up of a combination of recruitment, training, instrument fee, proctoring, and marking costs.
4. Alignments of test objectives with student or course objectives, and timeliness of data, are key for participation and motivation.

Given the high cost and logistical challenges of implementing assessments like the CLA+, CAT and HEIghten, institutions using the tools need to be confident that the data is valid and actionable. However, the reliability and validity of the data is called into question by the impact of student effort. There is a need for continued research on the implementation of assessment tools in alternate contexts, with a focus on logistical, motivational and cost factors.

ACKNOWLEDGEMENTS

The authors would like to acknowledge support from Higher Education Quality Council of Ontario (HEQCO). None of the authors on this paper are affiliated with the test organizations; CAE (CLA+), ETS (HEIghten) or Tennessee Tech (CAT). Queen's University paid for the tests in full and did not receive compensation from the organizations.

References

Blaich, C., & Wise, K. (2011). From gathering to using assessment results. *National Institute for Learning Outcomes Assessment*.

Boud, D. (1990). Assessment and the promotion of academic values. *Studies in Higher Education*, *15*(1), 101–111.

Boud, D. (1995). Assessment and learning: contradictory or complementary. In *Assessment for learning in higher education* (pp. 35–48). Kogan Page. Retrieved from http://www.teacamp.eu/moodle2/pluginfile.php/2910/mod_resource/content/1/UA/Assessment_and_learning_contradictory_or_complementary.pdf

Boud, D., & Falchikov, N. (2006). Aligning assessment with long-term learning. *Assessment & Evaluation in Higher Education*, *31*(4), 399–413.

Brown, G. E. (1992). *Testing in American Schools: Asking the Right Questions.* (pp. 1–314). Washington, D.C: Congress of the U.S. Retrieved from https://eric.ed.gov/?id=ED340770

Canadian University Survey Consortium. (2015). CUSC 2015 University Student Survey: Master Report. Retrieved September 25, 2015, from http://www.cusc-ccreu.ca/CUSC_2015_Graduating_Master%20Report_English.pdf

Chun, M. (2010). Taking teaching to (performance) task: Linking pedagogical and assessment practices. *Change: The Magazine of Higher Learning*, *42*(2), 22–29.

Coates, H. (2015). Assessment of learning outcomes. In *The European Higher Education Area* (pp. 399–413). Springer.

Coates, H. (2018). Research and Governance Architectures to Develop the Field of Learning Outcomes Assessment. In *Assessment of Learning Outcomes in Higher Education* (pp. 3–17). Cham: Springer.

Cumming, T., & Ewell, P. (2017). Introduction: History and Conceptual Basis of Assessment in Higher Education. *Publications and Research*. Retrieved from https://academicworks.cuny.edu/ny_pubs/231

Deming, D. J., & Figlio, D. (2016). Accountability in US Education: Applying Lessons from K-12 Experience to Higher Education. *Journal of Economic Perspectives*, *30*(3), 33–56. https://doi.org/10.1257/jep.30.3.33

Fadel, C. (2008). *21st Century Skills: How can you prepare students for the new Global Economy?* (pp. 1–22). Paris. Retrieved from http://www.aacc.nche.edu/Resources/aaccprograms/ate/conf2010/Documents/NSF%20ATE%20-%2021stCS%20-%20STEM%20-%20Charles%20Fadel.pdf

Frank, B. (2013). Web-based audience response system for quality feedback in first year engineering. In *ASEE 2013 Annual Conference*.

Frank, B., & Strong, D. (2010). Development of a Design Skill Assessment Tool. *Proceedings of the 2010 Canadian Engineering Education Association Conference*, *0*(0). Retrieved from http://library.queensu.ca/ojs/index.php/PCEEA/article/view/3165

Hutchings, M. (2015). *Exam Factories?: The Impact of Accountability Measures on Children and Young People*. National Union of Teachers.

Johnson, P. (2009). The 21st Century Skills Movement. *Educational Leadership*, *67*(1), 11.

Jonson, J., Guetterman, T., & Thompson, R. (2014). An Integrated Model of Influence: Use of Assessment Data in Higher Education | RPA Journal » An Integrated Model of Influence: Use of Assessment Data in Higher Education | Journal of Research & Practice in Assessment. *Research & Practice in Assessment*, *9*. Retrieved from http://www.rpajournal.com/an-integrated-model-of-influence-use-of-assessment-data-in-higher-education/

Klein, S., Liu, O., & Sconing, J. (2009). *Test Validity Study (TVS) Report*. Council for Aid to Education. Retrieved from http://cae.org/images/uploads/pdf/13_Test_Validity_Study_Report.pdf

Klein, S. P., Kuh, G., Chun, M., Hamilton, L., & Shavelson, R. (2005). An approach to measuring cognitive outcomes across higher education institutions. *Research in Higher Education*, *46*(3), 251–276.

Learning Outcomes. (n.d.). Retrieved August 20, 2018, from http://www.heqco.ca/en-ca/OurPriorities/LearningOutcomes/Pages/Home.aspx

Liu, O. L., Bridgeman, B., & Adler, R. M. (2012). Measuring Learning Outcomes in Higher Education: Motivation Matters. *Educational Researcher*, *41*(9), 352–362. https://doi.org/10.3102/0013189X12459679

Liu, O. L., Frankel, L., & Roohr, K. C. (2014). Assessing Critical Thinking in Higher Education: Current State and Directions for Next-Generation Assessment. *ETS Research Report Series*, *2014*(1), 1–23. https://doi.org/10.1002/ets2.12009

Liu, O. L., Mao, L., Frankel, L., & Xu, J. (2016). Assessing critical thinking in higher education: the HEIghten™ approach and preliminary validity evidence. *Assessment & Evaluation in Higher Education*, *41*(5), 677–694.

Miller, R. (2001). *Statewide Standardized Testing in Higher Education. Briefing Papers*. For full text: http://www. Retrieved from https://eric.ed.gov/?q=standardized+testing&ff1=subHigher+Education&id=ED468622

Pichette, J. (n.d.). The Postsecondary and Workplace Skills Project. Retrieved May 11, 2018, from http://www.heqco.ca/en-ca/OurPriorities/LearningOutcomes/Pages/the-postsecondary-and-workplace-skills-project.aspx

Rhodes, T. L., & Finley, A. P. (2013). *Using the VALUE rubrics for improvement of learning and authentic assessment*. Washington, DC: Association of American Colleges and Universities.

Salem, D., & Frank, B. (2017). The Role of Engineering Teaching and Learning Fellows in the Transformation Process of ECE Courses. In *Canadian Engineering Education Association Annual Conference*. Toronto.

Simper, N., Frank, B., Scott, J., & Kaupp, J. (2018). *Learning Outcomes Assessment and Program Improvement at Queen's University* (pp. 1–53). Ontario: Toronto: Higher Education Quality Council of Ontario. Retrieved from http://www.heqco.ca/SiteCollectionDocuments/Formatted%20Queens_LOAC_report.pdf

Stein, B., Haynes, A., & Redding, M. (2016). National Dissemination of the CAT Instrument: Lessons Learned and Implications. In *Proceedings of the AAAS/NSF Envisioning the Future of Undergraduate STEM Education: Research and Practice Symposium*. Retrieved from http://www.enfusestem.org/wp-content/uploads/2016/04/AAAS-Paper-2016-Stein_Haynes_Redding-revision.pdf

Tait, P. A. (2005). Assessment drives learning. *Journal of Pharmacy Practice and Research*, *35*(3), 211–212.
Tertiary Education Quality and Standards Agency, Pub. L. No. 73 (2017). Retrieved from https://www.legislation.gov.au/Details/C2017C00271/Html/Text
Wiggins, G. (1991). Teaching to the (Authentic) Test. In *Document Resume.* (pp. 353–358). Alexandria, VA: ERIC Digest. Retrieved from http://eric.ed.gov/?id=ed328611