Higher Education
Quality Council
of Ontario

An agency of the Government of Ontario

Learning Outcomes Assessment and Program Improvement at Queen's University

Natalie Simper, Brian Frank, Jill Scott and Jake Kaupp

Published by

The Higher Education Quality Council of Ontario

1 Yonge Street, Suite 2402
Toronto, ON Canada, M5E 1E5

Phone:	(416) 212-3893
Fax:	(416) 212-3899
Web:	www.heqco.ca
E-mail:	info@heqco.ca

Cite this publication in the following format:

Simper, N., Frank, B., Scott, J. & Kaupp, J. (2018). *Learning Outcomes Assessment and Program Improvement at Queen's University.* Toronto: Higher Education Quality Council of Ontario.

The opinions expressed in this research document are those of the authors and do not necessarily represent the views or official policies of the Higher Education Quality Council of Ontario or other agencies or organizations that may have provided support, financial or otherwise, for this project.

Executive Summary

This report describes a four-year longitudinal study that investigated a range of assessment methods for evaluating learning outcomes associated with critical thinking, problem solving, written communication and lifelong learning. The sample was drawn from the Faculties of Arts and Science, and Engineering and Applied Science. The measures included surveys, interviews, two standardized tests (the Collegiate Learning Assessment Plus and the Critical Thinking Assessment Test) and program-wide rubrics from the Association of American Colleges and Universities used to score student work samples independently of course grading. Researchers worked with course instructors to align teaching, learning and assessment, and to investigate and evaluate the utility of the instruments used.

The results of the study quantified longitudinal achievement of student outcomes on three instruments, with incremental growth in skills demonstrated across the studied undergraduate programs. The high-level outcomes were:

- Students' skills in critical thinking, problem solving and communication increased over the four years of their degree. The effects were detectable using the standardized tests (CLA+ d = .44, and CAT d = .65), but more evident using the Valid Assessment of Learning in Undergraduate Education (VALUE) rubrics. The Critical Thinking first-year median was Benchmark 1 while the second- and third-year median was Milestone 2, rising to Milestone 3 in fourth year.

- Queen's students demonstrated a higher level of skill in critical thinking than comparable students at most peer institutions participating in the CLA+ or CAT. For example, the Queen's fourth-year sample performed at the 87th percentile of CLA+ participating institutions.

- Student motivation was a significant concern for standardized tests. Results from student focus groups suggested that for students to put effort into testing, instructors need to value the test, the content needs to be relevant, careful consideration should be made to scheduling and the results should be made available to students.

- Motivation is not a concern when scoring academic work using program-wide rubrics, but alignment of course assignments to rubric dimensions is critical.

- The relative cost of implementing the VALUE rubric marking was approximately C$20 less per student than implementing the CLA+ or CAT tests.

- Qualitative and quantitative feedback facilitated through departmental reports and debriefs prompted improvements to courses.

- Work needs to continue to increase the adoption of effective practices in assessment.

Table of Contents

List of Tables

List of Figures

Definition of Terms

Critical thinking

To avoid ongoing contention as to what exactly constitutes critical thinking, the VALUE rubric definition was adopted early in the project: "Critical thinking is a habit of mind characterized by the comprehensive exploration of issues, ideas, artifacts, and events before accepting or formulating an opinion or conclusion" (AAC&U, 2014). It was operationalized using the following five dimensions: explanation of issues, evidence, influence of context and assumptions, student's position, conclusions and related outcomes.

Complex cognitive skills

This is a term borrowed from the study of cognition and working memory. It refers to the acquisition, retention and application of complex knowledge and skills "such as those involved in both moment to moment decisions and in more long-term strategies" (Logie, Baddeley, Mané, Donchin & Sheptak, 1989, p. 54). Complex cognitive skills (CCS) comprise a number of interrelated constituent skills and require considerable time and effort to acquire mastery in (Van Merriënboer, 1997). The term CCS is used very broadly in this report to collectively encompass skills described in Deller, Brumwell and MacFarlane (2015) as basic cognitive skills (communication), higher-order cognitive skills (critical thinking and problem solving) and transferable skills (lifelong learning).

Learning outcomes

Learning outcomes are measurable statements of student knowledge and abilities, described "as existing at the intersection of concepts (what students know and understand) and competencies (what students are able to do)" (Roksa, Arum, & Cook, 2016, p. 17).

Lifelong learning

Lifelong learning is a term that is widely used but does not have a universal definition. In overarching terms, it describes continuous, self-directed learning and involves motivation to acquire knowledge and skills in an ongoing manner. The term is used in this report to describe an individual's attitudes and behaviours toward learning, specifically "dispositions (how an individual feels) and learning strategies (skills, processes and meta-cognition)" (Simper, Kaupp, Frank, & Scott, 2015, p. 1160).

Transferable learning orientations (TLO)

This refers to a reflective survey developed at Queen's University as a point-in-time measure of lifelong learning. It comprises dimensions of goal orientation, learning belief, self-efficacy, knowledge transfer and organization.

Problem solving

The term problem solving is used here to describe the resolution of messy, complex problems, dealing with "a large number of barriers that coexist simultaneously (and the desire to) overcome barriers between a given state and a desired goal" (Sternberg & Frensch, 2014, p. xi). The project adopted the

VALUE rubrics for assessing student assignments, where problem solving was operationalized using the following six dimensions: define the problem, identify strategies, propose solutions, evaluate potential solutions, implement solutions and evaluate outcomes.

Value-add The difference between performance in first and final year, used to estimate the contribution of an educational institution toward student outcomes.

VALUE rubrics Valid Assessment of Learning in Undergraduate Education (VALUE) rubrics, developed by the Association of American Colleges and Universities (AAC&U).

Written communication Was evaluated in the English language in any text format applicable to the discipline (for example, short answer/extended response, essay, report or reflection.) The project adopted the VALUE rubrics for assessing student assignments, where written communication was operationalized using the following five dimensions: context and purpose, content development, genre and conventions, sources and evidence, and syntax and mechanics.

Program Codes

BA	Bachelor of Arts (includes BA Honours)
BCMPH	Bachelor of Computing (Honours)
BED	Bachelor of Education (includes Concurrent Education)
BNSC	Bachelor of Nursing Science
BSC	Bachelor of Science (includes BSc Honours)
BSCE	Bachelor of Science Engineering (program was renamed from Bachelor of Applied Science partway through the project)

Introduction

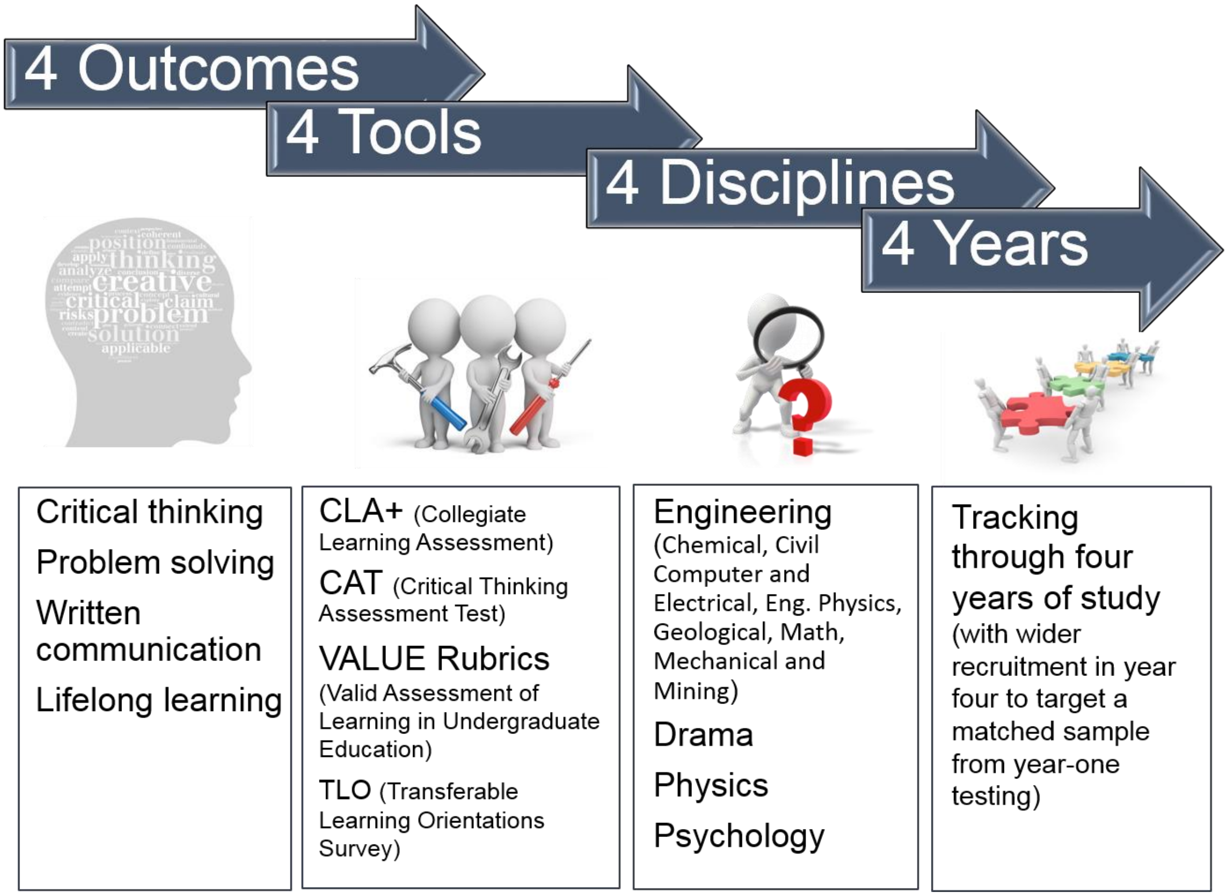
Complex cognitive skills like critical thinking, communication, problem solving and lifelong learning are fundamental elements of an undergraduate education. They are central to broad frameworks in higher education such as the Essential Learning Outcomes from the Association of American Colleges and Universities (AAC&U), the Degree Qualifications Profile from the Lumina Foundation and Partnership for 21st Century Skills (Johnson, 2009), and are critical to the practice of professional fields such as engineering (Kaupp, Frank, & Chen, 2014). Forty-six percent of Canadian university students rank thinking logically and analytically among the top three most important skills to which their university degree should be contributing, more than double that of the next highest ranked areas: content knowledge and time management (Canadian University Survey Consortium, 2015). However, even though complex cognitive skills are considered an essential element of higher education, they “are often considered to be among the most difficult outcomes to define, teach and assess” (Deller, Brumwell, & MacFarlane, 2015, p. 13).

In 2013, Queen’s University researchers began a longitudinal exploratory study to investigate the development and assessment of complex cognitive skills as part of the Learning Outcomes Assessment Consortium funded by the Higher Education Quality Council of Ontario (HEQCO). Researchers tracked skills development in disciplines spanning engineering, science, social science and humanities. The method involved testing students using standardized measures and scoring student work from select courses within specific departments. The four-year study used four approaches to assessing cognitive skills in multiple departments over four years (see Figure 1):

- Standardized instruments and surveys
- Program-wide rubrics used to score student work samples
- Student and instructor interviews
- Data linkage to demographic variables and student grades

The study documented the costs, time commitment, participation rates, motivation and correlations between measures, and evaluated their utility and reliability. In the first year of the study a double, cross-sectional assessment was used to pilot the tools with first- and fourth-year students to compare performance. There were four very broad research questions guiding the investigation and many more specific questions that arose from the process. Methods were selected as appropriate to the question and the underlying purpose of investigation. The research design is summarized in Table 1.

Figure 1: Project Overview



4 Outcomes	4 Tools	4 Disciplines	4 Years
Critical thinking Problem solving Written communication Lifelong learning	CLA+ (Collegiate Learning Assessment) CAT (Critical Thinking Assessment Test) VALUE Rubrics (Valid Assessment of Learning in Undergraduate Education) TLO (Transferable Learning Orientations Survey)	Engineering (Chemical, Civil Computer and Electrical, Eng. Physics, Geological, Math, Mechanical and Mining) Drama Physics Psychology	Tracking through four years of study (with wider recruitment in year four to target a matched sample from year-one testing)

Table 1: Research Questions, Methodology and Purpose of Investigation

Research Question	Method: Sample	Data source	Purpose
A. How much do students' complex cognitive skills change between the first and fourth year of undergraduate studies?	Quantitative: Representative sample of undergrad students	• CLA+ assessment • CAT test • VALUE Rubric assessment	To quantify student achievement between first and fourth year: Institutional value-add
B. How does the development of complex cognitive skills and lifelong learning vary between programs and individuals, and what is the relationship of standardized measures to course grades?	Quantitative: Undergrads from selected courses	• CLA+ assessment • CAT test • VALUE rubric assessment • Transferable Learning Orientation (TLO) survey • Demographic variables • Cumulative grade-point average	To understand more about standardized tests in terms of reliability and validity and relevance to different disciplines
C. Can data from instruments be used to support skills development in courses?	Qualitative: Undergraduate students; Course instructors	• Individual and group • Interviews	To encourage faculty to develop and assess complex cognitive skills in their courses and programs
D. How feasible is the use of these assessments in a Canadian university?	Quantitative: Document study Qualitative: Course instructors	• Financial and research documentation • Debrief/interviews	To understand more about standardized tests in terms of implementation and to inform institutional investment of assessment in the coming years

Instrumentation

Collegiate Learning Assessment Plus (CLA+)

The Collegiate Learning Assessment Plus (CLA+) was developed by the Council for Aid to Education (CAE) (Council for Aid to Education, n.d.). It is a 90-minute, web-based instrument that students access through a secure browser. It is made up of a performance task (60-minute maximum) and a series of selected response questions (60-minute maximum). The performance task presents a real-world situation in which students assume a role to address the problem, identify a solution, or provide conclusions and recommendations resulting from careful analysis of the evidence provided. The performance task is used to measure critical thinking, problem solving and written communication. The student responses in the performance task are scored by an automated system using a validated rubric. The selected response questions are document-based, multiple-choice or short-answer questions, and are used to measure scientific and quantitative reasoning, critical reading and evaluation, and critiquing an argument.

Critical Thinking Assessment Test (CAT)

The CAT was developed by Tennessee Tech with support from the National Science Foundation (NSF). It is a 60-minute, paper-based test. Similar to the CLA+, the CAT engages students in real-world problems. The test was designed to assess critical thinking, creative-thinking skills, as well as non-routine problem solving and effective communication (CAT n.d.). The CAT provider requires a minimum of two members from participating institutions to be trained in the scoring protocols. The trained institutional representatives then use a detailed marking key to lead the CAT scoring conducted at the institution. CAT personnel score a random sample of the tests to substantiate scoring reliabilities.

Valid Assessment of Learning in Undergraduate Education (VALUE Rubrics)

The VALUE rubrics were developed by the Association of American Colleges and Universities to provide a valid assessment of learning in undergraduate education (AAC&U, 2014). These rubrics are broad, discipline-neutral descriptions of selected essential learning outcomes of undergraduate education from the Liberal Education America's Promise (LEAP) initiative. Panels of experts identified common themes and developed performance criteria for each rubric. The efforts of the experts were focused on positive demonstration of outcomes, describing performance criteria aimed at being used to assess summative displays of student learning. There are four levels of performance criteria, from the Benchmark level of students entering university to the Capstone level of students who have just completed their undergraduate experience. There are 16 VALUE rubrics in total, of which four were used in the study. The Critical Thinking, Problem Solving and Written Communication VALUE rubrics were used in their published form, and the Lifelong Learning rubric was adapted as part of a survey (see section below on the Transferable Learning Orientation survey).

Reliability and Validity of CLA+, CAT and VALUE Rubrics

Each of the above instruments has been rigorously evaluated for reliability and validity. By selecting these instruments, the Queen's Learning Outcomes Project leveraged previous empirical evidence in support of the validity of the assessment constructs and reliability of the test or rubric structure. For example, in a study conducted by the CLA, students responded positively that the CLA+ is an effective measure of critical thinking and problem solving (86.2%), reading comprehension (88%) and writing (75.8%). Benjamin, Klein, Steedle, Zahner & Elliot, (2012), and Klein et al. (2009) reported that the CLA+ was well correlated with two other measures of critical thinking: the Collegiate Assessment of Academic Proficiency, and the Measure of Academic Proficiency and Progress (CLA Performance Task $r = 0.73–0.83$, CLA Critique an Argument task $r = 0.73–0.94$). The CAT has been demonstrated to correlate with the American College Testing readiness assessment (ACT) $r = .56$ $p < .01$; SAT $r = .57$ $p < .01$; grade point average $r = .35$ $p < .0$; and the California Critical Thinking Skills Test (CCTST) $r = .64$ $p < .01$ (Stein et al., 2006). Stein & Haynes (2011) also found that the CAT instrument was sensitive to detecting gains in students' critical thinking across a variety of disciplines.

Following the LEAP initiative in 2005 (AAC&U, n.d.), the AAC&U worked consistently to establish the validity and reliability of the VALUE rubrics. A diverse interdisciplinary panel of experts was used to gauge the face and content validity of the rubrics. In each rubric, common themes were identified and panels of experts developed performance criteria to assess summative displays of student learning. These experts agreed that the rubrics were an effective and suitable measure of the underlying constructs (Finley, 2011). Ongoing work supporting the validity and reliability of the VALUE rubrics is presented in Rhodes (2011) and Rhodes & Finley (2013). The assessment constructs for each are summarized in Table 2.

Table 2: Overview of Assessment Constructs

Instrument	**Assessment construct**			
	Critical thinking	**Problem solving**	**Written communication**	**Other**
Collegiate Learning Assessment (CLA+)	• Critical reading and evaluation • Critique an argument	• Analysis and problem solving	• Writing mechanics • Writing effectiveness	• Scientific and quantitative reasoning
Critical Thinking Assessment Test (CAT)	• Evaluation and interpretation of information	• Problem solving	• Effective communication	• Creative thinking
Valid Assessment of Learning in Undergraduate Education (VALUE) Rubrics	• Explanation of issues • Evidence • Influence of context and assumptions • Student's position • Conclusions and outcomes	• Define problem • Identify strategies • Solution/ hypothesis • Evaluate solution • Implement solution • Evaluate outcomes	• Context and purpose • Content development • Genre and conventions • Sources of evidence • Syntax and mechanics	

Transferable Learning Orientations (TLO) Survey

One of the goals of the research was to develop a method for evaluating student dispositions and behaviours aligned with lifelong learning. This was not as straightforward as tracking the demonstration of the other outcomes being investigated. One of the biggest challenges was to come to agreement as to how we defined lifelong learning. The latent nature of the constructs involved meant that we were looking for a suitable self-reporting measure. The most promising instrument on the market was the Effective Lifelong Learning Inventory (ELLI) (Crick, Broadfoot, & Claxton, 2004). The research was already committed to significant costs involved in the use of the CLA+ and the CAT, and the fee for using the ELLI was not within the financial scope of the project. Researchers instead undertook a process of implementing a free inventory, the Motivated Strategies for Learning Questionnaire (MSLQ) (Pintrich, Smith, Garcia, & Mckeachie, 1993). Our investigations of the initial MSLQ pilot found some issues and inconsistencies. In our quantitative analysis, we struggled to find consistency in responses, the factor structure was problematic and short average response times suggested that students had not engaged fully with the meta-cognitive nature of the questions.

This result prompted our development of the Transferable Learning Orientation (TLO) survey. For full details on the development of the TLO see Simper et al. (2015). The process involved refining the MSLQ scales, adapting the Lifelong Learning VALUE rubric, and incorporating a qualitative response for each dimension aimed at increasing meta-cognitive engagement with the instrument. The TLO is a triangulated measure using quantitative pairing scale items together with a holistic rubric self-rating and an open-ended response used to validate the final rating. See Table 3 for an overview of the TLO dimensions.

Table 3: TLO Dimension Overview

		Goal Orientation (Motivated)	**Learning Belief (Flexible Learner)**	**Self-Efficacy (Confident)**	**Transfer Knowledge (Makes Connections)**	**Organization (Learns Independently)**
Target level	**4**	Explores the topic in depth, intrinsically motivated	Mastery approach with full control over own learning	Confident in own knowledge/skills necessary to excel	Connects knowledge in innovative ways, deep learner	Adaptive organizational techniques
	3	Motivated to explore topic in some depth	Believes that effort will lead to success	Appropriate level of knowledge/skills to meet goals	Makes references to previous learning	Systematic organization
⇑						
Low level	**2**	Explores topic, but primarily extrinsically motivated	Believes in some control of success or failure	Adequate level of knowledge/skills	Makes some references to previous learning	Organizes work processes
	1	Surface level engagement	Believes in fixed ability	Tentative about own level of knowledge/skills	Memorizes information, surface learner	Ad-hoc organization

The tool was in development for much of the longitudinal study and our work in this area is continuing, therefore no results have been included in this report.

Group Interviews

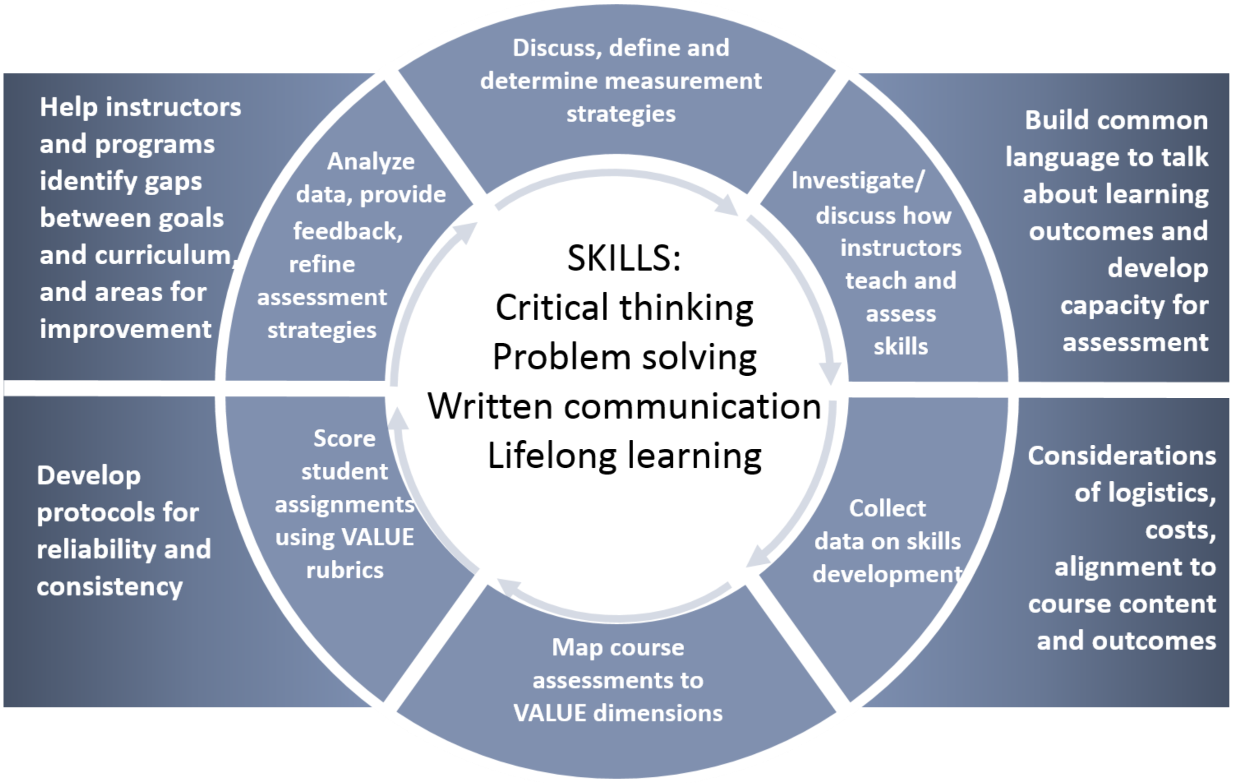
Participating instructors provided feedback on the project outcomes. Some of the feedback was delivered informally and a selection of instructors provided more formal feedback on the utility of each of the instruments through recorded debrief sessions that were transcribed for analysis. In those cases, course-specific reports were provided to instructors. The reports explained the achievement demonstrated by the students within the course and how that sample compared to student achievement in the longitudinal context. Additionally, student feedback was elicited for suggestions on presentation formats for student-learning outcomes reports. Students provided information about academic and co-curricular experiences that help them develop critical thinking, problem solving, communication and lifelong learning skills.

Method

The research was facilitated through the participation of course instructors. Each year of the project involved the six steps, all corresponding to a larger research goal. Figure 2 displays the various research steps and the corresponding goals. The first step was to define the assessment constructs and identify the tools that best suited assessment needs. Each year we investigated the efficacy of the tools and refined the tool or implementation process as necessary. The next step was to work with instructors to explain the assessment constructs and instruments being used to determine which of the course assignments might be suitable for inclusion in the VALUE rubric scoring. A typical prompt for instructors was: "In which of your course assignment(s) do students demonstrate cognitive skills?"

Through consultation and discussion, the team worked toward a common language describing cognitive skills. At this step, many of the instructors decided to tailor an assignment to better elicit demonstrations of cognitive skills. There was recruitment, testing and course data collection, then mapping of student skills demonstration to the assessment criteria to determine alignment. The investigation of costs and logistics was also a goal of the research. There was a significant investment made in scoring CAT tests and course work samples. This required developing or working to set protocols, training and ensuring longitudinal consistency of scorers.

Figure 2: Process Framework


Help instructors and programs identify gaps between goals and curriculum, and areas for improvement
Discuss, define and determine measurement strategies
Build common language to talk about learning outcomes and develop capacity for assessment
Analyze data, provide feedback, refine assessment strategies
SKILLS:
Critical thinking
Problem solving
Written communication
Lifelong learning
Investigate/ discuss how instructors teach and assess skills
Develop protocols for reliability and consistency
Score student assignments using VALUE rubrics
Collect data on skills development
Considerations of logistics, costs, alignment to course content and outcomes
Map course assessments to VALUE dimensions

VALUE Marking Protocols

The VALUE rubric scoring was conducted following the process described in an instructional video featuring Ashley Finley of the AAC&U,[1] and summarized here:

Structure of rubrics

Although it is tempting to read the rubric first, it is important to review the front page to familiarize ourselves with the framing language and the purpose of the rubric. The glossary of the rubric offers the best opportunity for modifying it and for adding or modifying terms to clarify the performance descriptors on the back page. It lists the criteria for the learning outcomes and stresses the multidimensionality of what it takes to demonstrate any of these outcomes; essentially, it is a breakdown of the components of the skills. The performance benchmarks are not intended to be time-specific but developmental and ongoing. The

[1] Ashley Finley is the vice president of Academic Affairs and Dean of the Dominican Experience and Senior Fellow, AAC&U, Dominican University of California. The video is available at https://youtu.be/josqNZpoZnkm

Capstone benchmark is intentionally placed on the left so that it is read first. If the marker reads the benchmark first, it may artificially constrain the assessment of the student; the rubric is not meant to be read based on what is expected of students in a particular year, but the level of skill they demonstrate.

Using rubrics to assess student work

It is suggested that the best way to understand the rubric is to use it through calibration sessions; that is, working through a sample of student work using a rubric. Participants can discuss the language and expectations utilized in the rubric. In our experience, this process is best started by understanding the artifact, without consideration of the rubric, to familiarize ourselves with the content. Then we review the framing language and the performance descriptors. Finally, we systematically work our way through the criteria and assign a performance level based on evidence in the artifact. Discussing with other markers the levels assigned, the evidence used and interpretation of language in the performance descriptors is key to gaining proficiency with using the rubrics.

Using rubrics at institutional level to improve evaluation and assignments

Faculty can use the rubrics for targeted assignments rather than for structuring the entire course. Faculty have several questions to consider when structuring these assignments, such as how do students demonstrate these learning outcomes in the assignment and how does this demonstration build upon or challenge existing levels of competence. Weak areas within the criteria should be identified to target specific skills, as should areas of strength to maximize work that is already occurring. The rubrics are inherently interdisciplinary; for example, writing skills are not specific to the English department nor are quantitative skills to the mathematics department. In order for the institutions to encourage this use of the rubrics, faculty must be engaged and given the opportunity to establish ownership over the rubrics. How this can be achieved must be determined at an institutional level, as schools must customize their approaches to the school culture.

The VALUE rubric was conducted using the following protocol:

a) Building a common understanding:
- Read through the assignment instructions and sample responses to build an understanding of the nature and context of the course assignment.
- Identify what the students were directed to demonstrate (this might, for example, require reading a research paper to which the students were responding).
- Operationalize the "issues," "contextual factors" and "assumptions" relevant to the student responses.

b) Rating a work sample:
- Collectively work through a single student response (not one included in the research sample) to identify evidence for each of the dimensions to be rated. Research Assistants (RAs) then discuss what level the evidence suggests the response is demonstrating.

- Individually rate five to 10 work samples at a time, compiling an annotated list to back up the decision for each of the criteria.
- Assign and record a performance level (for each dimension) for the work samples.

c) Calibration:

- The two markers use their annotations to discuss any differences between levels assigned.
- In some cases, this process results in one or the other of the markers adjusting their level on a dimension. The rating process is based on individual interpretation, so differences in level determinations were occasionally observed. These changes are recoded and reported as post-calibration agreement.
- Repeat the rating and calibration process for the remainder of the work samples. Generally, the greater the number of assignments that are rated, the fewer differences there are in ratings.

For financial and logistical reasons, many of the work samples were marked by trained undergraduate students. Disciplinary experts were employed to calibrate where necessary with the undergraduate markers. Longitudinal consistency of marking was supported by employing the same markers for various course artifacts both longitudinally and across disciplines. There were 18 VALUE rubric scorers (denoted by letters A-R in Table 4) over the four-year duration of the project.

Table 4: Mapping of VALUE Rubric Marking

		Engineering				**Drama**		**Physics**		**Psychology**			
		Year				Year		Year		Year			
		1	2	3	4	1	2	1	3	1	2	3	4
Researchers (subject experts)	A	*	*		*								
	B				*						*	*	*
	C			*									
	D				*								
					*								
Graduate students (subject experts)	E				*								
	F			*					*				
	G				*								
	H							*					
Undergraduate Students (calibrated with subject experts)	I					*	*			*	*	*	*
	J	*						*		*			
	K	*				*							
	L	*					*	*			*		
	M	*	*										

	Engineering				Drama		Physics		Psychology			
	Year				Year		Year		Year			
	1	2	3	4	1	2	1	3	1	2	3	4
N	*	*										
O			*					*				
P			*								*	
Q				*								*
R				*								*

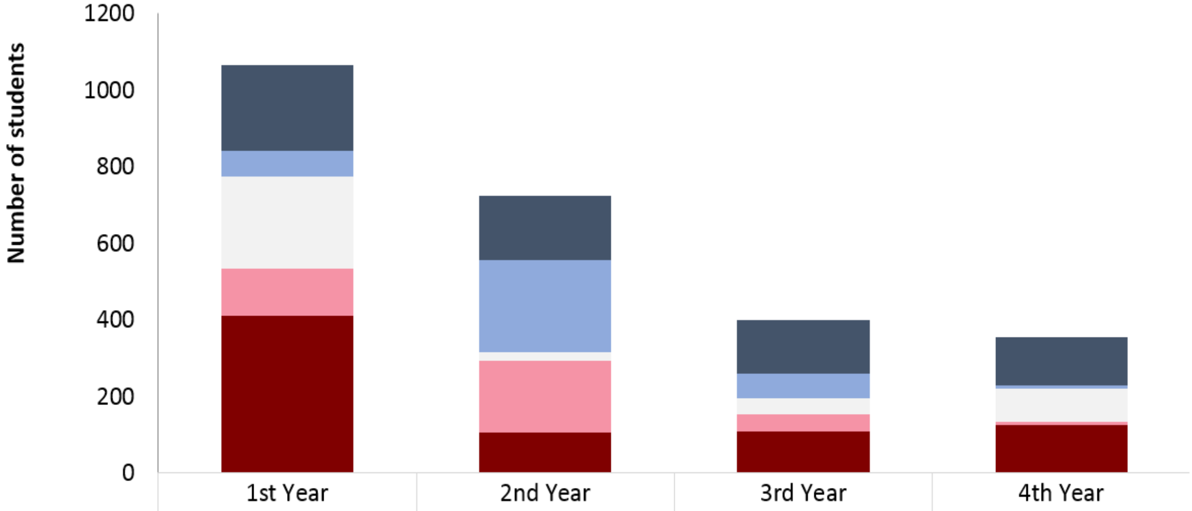
Once data collection and scoring had been completed each year, analysis was undertaken and reports generated for each of the participating courses. In many cases, this step prompted further investigation of anomalies. Where possible, the course reports were delivered in an instructor debrief session with the goal of facilitating usable feedback for course improvement. Upon invitation, reports were presented to undergraduate committees in specific departments. Finally, research was disseminated annually at conferences, symposiums and webinars.

Sample

Undergraduate students were recruited by research associates from specific courses in the programs including Psychology, Physics, and Drama in the Faculty of Arts and Science, and from Chemical Engineering, Civil Engineering, Computer and Electrical Engineering, Engineering Physics, Geological Engineering, Math Engineering, Mechanical Engineering, and Mining Engineering in the Faculty of Engineering and Applied Science. Consistent with our ethical guidelines, individual student consent was sought in each year of the project. In the initial years of the project, more students consented to the research than we had capacity to mark course assignments. The total consenting sample of students who participated the study in first year was 2,697; second year, 785; third year, 599; and in fourth year there were 419 consenting students. Figure 3 displays only the consenting students who took the test, or whose summative course assignment was rated on the VALUE rubrics. There were also many students who consented and began the test, but for a range of reasons did not complete it.

Every effort was made to evaluate student assignments from the consenting students who completed one of the standardized tests, but this was not always possible. When the assignment had a group component, ethical use of data meant that each of the members of the group were required to be consenting (this was particularly problematic in first year). In addition, the fourth-year testing was conducted separately from the courses included in the study, so few of the students who we were recruiting were enrolled in the fourth-year courses that had assignments to sample from. Researchers marked multiple assignments in some courses to investigate learning gains within a course. For this report, researchers selected only the summative assignment that the students completed toward the end of their course for inclusion in analysis.

Figure 3: Project Sample for Each Instrument



	1st Year	2nd Year	3rd Year	4th Year
■ CAT only	222	167	141	126
■ CAT & VALUE	69	239	65	7
■ VALUE only	240	24	41	88
■ CLA & VALUE	121	186	46	9
■ CLA only	412	107	108	125
Total	1150	726	444	295

Implementing the CLA+ and CAT

Where possible, CLA+ and CAT testing was conducted in participating courses as part of regular course activities. Due to class schedules and course commitments, this was not always possible. In those cases, testing was conducted out of class with either food or a financial incentive offered. For the in-class testing, students who were not scheduled in a lab were asked to bring their own laptop or required to move to an alternate environment. Students in the first-year cohort were assigned to either the CLA+ or the CAT. Every effort was made to recruit the same individuals to the same test over the four years, but because of the ethical requirement for annual consent and the different course pathways for students, the samples differed across the four years.

CLA+ Results

The CLA+ test providers allocate achievement standards of "below basic," "basic," "proficient," "advanced" and, since 2016, an "accomplished" level.[2] The cut points for each of the standards are calculated using the CLA+ total score. The total score is a numerical composite of the performance task score and the selected response score. Only students who complete the whole test are provided with their achievement standard. During test proctoring, it was observed that some students spent very little time completing the test. Since there were no stakes attached to the test, the results were considered invalid if students spent less than 10 minutes on the performance task component (60 minutes allocated), or if they declared in the exit survey that they had put "no effort" into the test. Following these criteria, excluded from the analysis were six first-year students, 12 second-year students, three third-year students and one fourth-year student. The excluded students represent 2% of the test population. Figure 4 displays the percentage of students at each level, with sample sizes for the number of consenting students who completed the test each year. The first-year CLA+ total score mean was 1,155.3 (SD 117.8), and the fourth-year mean was 1,211.5 (SD 116.9).

Data for US comparisons was drawn from the Council for Aid to Education National Report (2016). There were 80 participating institutions from the United States, with a first-year mean average of 1,053 (SD 74) and a fourth-year average of 1,126 (SD 74). Comparisons were calculated using Cohen's d (d = M_2-M_1/pooled SD), with results suggesting that Queen's students begin one standard deviation higher than the US average (d = 1.07), and finish just under a standard deviation higher (d = .90). The Queen's first-year mean was at the 89th percentile (compared with the 2013–14 CLA+ participating institutions), and the fourth-year mean was at the 87th percentile (compared with the 2016–17 CLA+ participating institutions).[3]

Effort in the test

The CLA+ exit survey asks students to report the effort separately for the performance task, and selected response sections. The performance task effort was the focus of the analysis, because it forms the largest part of the test and was more labour intensive than the selected response section. Average effort scores for the CLA+ performance task dropped over the first three years (first-year mean effort was 3.06, second-year mean was 2.74 and third-year mean was 2.69). These results were calculated following the exclusion of the 22 students whose results were considered invalid (as mentioned in the section above). Table 5 plots the effort to performance task score, displaying the regression to the mean. The R^2 = .10 for first- and second-year effort suggests that 10% of score variance was accounted for by effort in the initial years of the project. These results were reviewed through the course of the project and following recommendations from our student group interviews (see section below), we provided a C$25 incentive per test-taker and a C$750 prize draw. Table 5 displays the fourth-year CLA+ effort mean of 3.23, but by fourth year the effect of effort on test score was not significant.

[2] Research around standard-setting was undertaken by the RAND Corporation. Details are available here: https://www.rand.org/pubs/technical_reports/TR663.html

[3] Source for comparative CLA+ data derived from the CAE. Full reports are available here: http://cae.org/flagship-assessments-cla-cwra/cla-and-cwra-national-report-archive

Figure 4: Percentage of CLA+ Sample at Each Standard Level

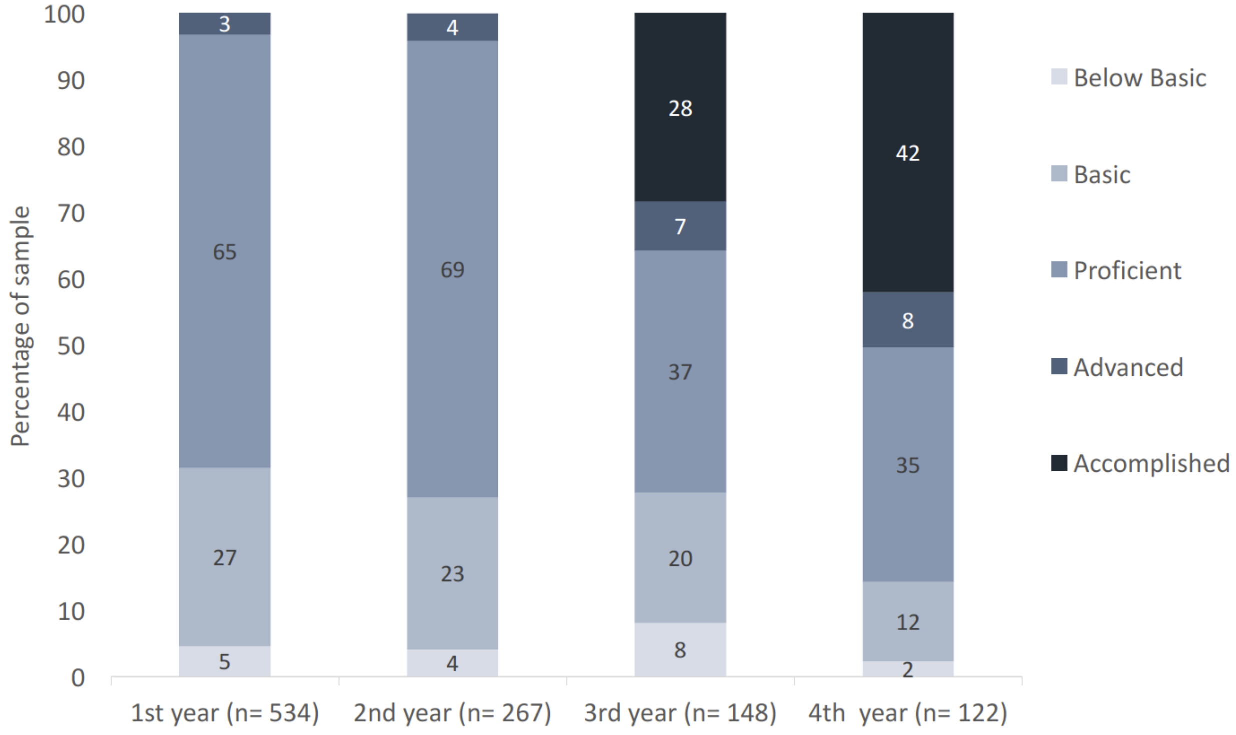


Table 5: Relationship between CLA+ Performance Task and Score

	1st year	2nd year	3rd year	4th year
Performance task effort mean	3.06	2.74	2.69	3.23
Performance task effort R^2	$.102^{**}$	$.099^{**}$	.020	.013
Performance task score variance accounted for by effort	10.2%	9.9%	2.0%	1.3%

$^{**}p<.001$

Longitudinal change

The descriptive statistics for the entire pool first- and fourth-year students by program are shown in Table 6.

Table 6: Descriptives for Whole CLA+ Sample, First and Fourth Year

Whole sample: Program	Year	N	Sex		First-year Grade point average		Performance task effort	
			Male	Female	Mean	SD	Mean	SD
BA	1	52	17	35	3.3	0.8	3.5	1.0
BA	4	25	3	22	2.8	1.1	3.7	0.9
BCMPH	1	14	13	1	2.6	0.8	3.6	0.9
BCMPH	4	4	3	1	2.8	1.1	3.8	1.0
BCOM	1	1	1	0	3.0	NA	2.0	NA
BED	1	1	1	0	2.0	NA	4.0	NA
BED	4	2	0	2	1.3	NA	4.0	0.0
BFAH	1	1	0	1	3.7	NA	3.0	NA
BNSC	1	38	3	35	3.3	0.4	3.6	0.9
BNSC	4	1	1	0	3.8	NA	4.0	NA
BPHEH	1	2	0	2	3.4	0.4	3.0	1.4
BSC	1	107	50	57	3.1	0.8	3.9	0.8
BSC	4	19	7	12	2.9	0.9	3.6	0.9
BSCE	1	210	174	36	3.1	0.8	3.8	0.7
BSCE	4	10	3	7	3.4	0.7	4.1	0.7
CIB	4	1	0	1	NaN	NA	3.0	NA

The fourth-year sample included in the analysis was 122 students. Of those, 73 were the same individuals who tested in the first year of the study (see descriptives in Table 7). There were significant gains in the CLA+ total score among this population (t = -4.42, df = 72, p<0.0001) with an effect size of d = 0.52, but there are also significant differences in how these students reported their effort on both the performance task sections in first and fourth year (V = 501, p-value = 0.01), though not in the selected response task (not shown). Differences in performance task effort means in the final column of Table 6 suggest that the fourth-years put more effort in than first-years in every program. This, however, was not the case for the 73 students who tested both in first and fourth year. It may have been that these students were exerting less effort on average than they did in first year, or that having been tested multiple times by fourth year, they underreported their effort.

T-tests were analyzed to investigate longitudinal change for students within individual programs. The difference between first and fourth year for the 28 engineering students (t = -4.2, p<0.001) and 15 arts students (t = -3.3, p<0.01) showed a significant improvement. The sample size here is small, so the finding has limited generalizability. As a result, we will look at propensity score analysis.

Table 7: Descriptive Statistics for CLA+ Total Score

Year	N	M	F	Mean	SD	PT effort mean	PT effort SD
1	73	26	47	1170.60	122.10	3.84	0.73
4				1232.74	110.42	3.59	0.72

Propensity score analysis

The underlying assumption of reporting the value-add difference between performance in first and fourth year on the CLA+, is that the first-year and fourth-year groups are comparable. The results, however, are sample-dependent and there are concerns related to student motivation, the amount of time spent on the writing task, and “that institutions may try to game the test by selecting high achievement senior year students” (Douglass, Thomson, & Zhao, 2012, p. 320). To examine the CLA+ results more closely and address concerns of sample bias, effort and engagement, and differences in academic potential, analytical matching of first- and fourth-year samples was conducted. The first step was to analyze differences in test effort and investigate the significance of individual characteristics, after which a process of propensity matching was undertaken, matching fourth-year students to students with similar characteristics from the first-year test sample.

There were some differences between the first- and fourth-year samples. There was a greater proportion of female students in the fourth-year sample, significant differences between the first- and fourth-year performance task effort and in the mean first-year GPA (FYGPA). A general linear model was used to extract the most significant predictors of the CLA+ total score, which emerged as FYGPA, performance task and selected response effort, time spent, and first language. Students who completed the CLA+ in both first and fourth years spent more time on average working on the test in fourth year (42 minutes) than in first year (37 minutes), but reported a lower effort in fourth year. It is possible that students’ perceptions of effort was lower in fourth year, rather than actual effort. Due to this uncertainty, effort was not used in propensity matching. Instead, degree program, first year GPA, performance task time and first language were used. This was run using the *MatchIt* library in R (Ho, Imai, King, & Stuart, 2007) using *optimal* matching. This pool of records was merged with the students who wrote the CLA+ in both first and fourth year, leading to a final pool with characteristics shown below. Table 8 displays descriptives of the combined repeating test students (n = 73) and propensity matched samples (n = 49). The sum of the first-year sample is 122, matched to 122 fourth-year students.

Table 8: Descriptives for the Pool of Repeating and Matched CLA+ Samples

Matched sample: Program	Year	N	First language		FYGPA	
			English	Other	Mean	SD
BA	1	25	19	6	3.02	0.80
BA	4	29	23	6	2.84	0.97
BCMPH	1	4	4	0	2.90	0.95
BCMPH	4	4	3	1	3.15	1.18
BED	1	6	5	1	3.29	0.65
BED	4	7	5	2	2.92	1.03
BNSC	1	6	3	3	3.12	0.43
BNSC	4	7	3	4	3.22	0.48
BSC	1	41	32	9	2.97	0.98
BSC	4	37	31	6	3.08	0.94
BSCE	1	40	36	4	3.33	0.59
BSCE	4	38	35	3	3.37	0.65

Where sample sizes allow, the performance task and selected response have been plotted in Figures 5 and 6. The highest fourth-year means were demonstrated by the Bachelor of Science (BSc) students, but the largest gains were demonstrated by the Bachelor of Arts (BA) students. Both of these groups include students in honours programs.

Table 9 summarizes t-tests for the whole group and three degrees with at least 20 records in each year, and Cohen's d for effect size. T-test comparing first to fourth year: t = -3.4312, df = 244, p < 0.001, d = 0.44. As previously mentioned, the CLA total is a scaled weighted score calculated using the sub-scores from the performance task and the selected response. The breakdowns for the CLA+ for the matched samples have been included as Appendix 1.

Figure 5: Score Distributions of Matched Students on CLA+ Performance Task by Degree Program

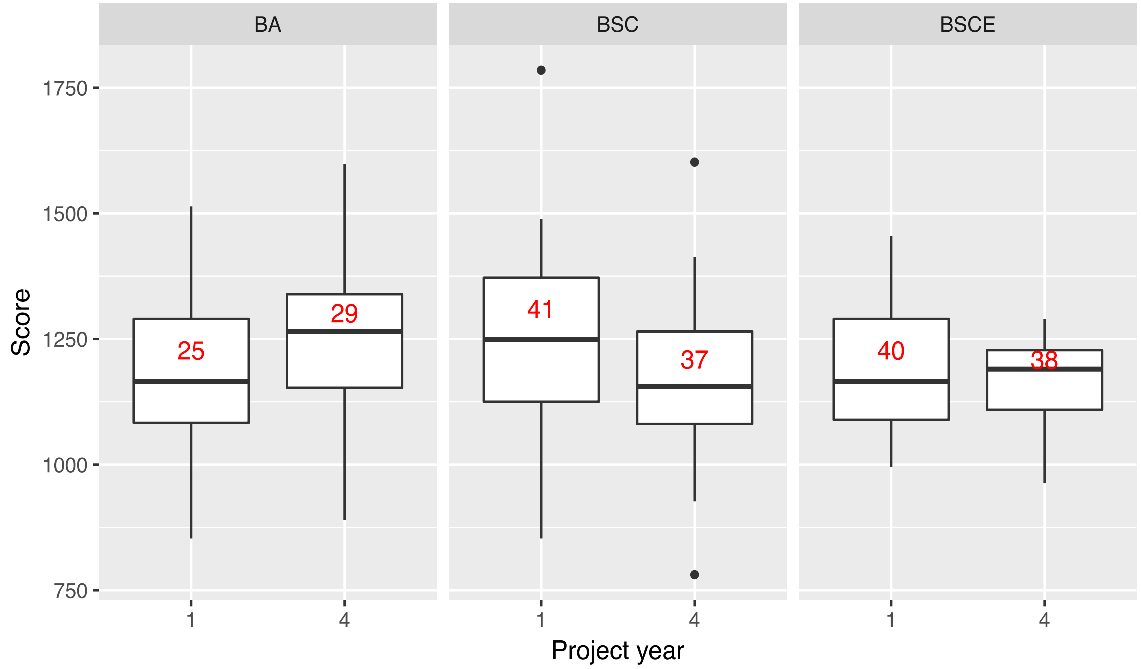


Figure 6: Score Distributions of Matched Students on CLA+ Selected Response by Degree Program

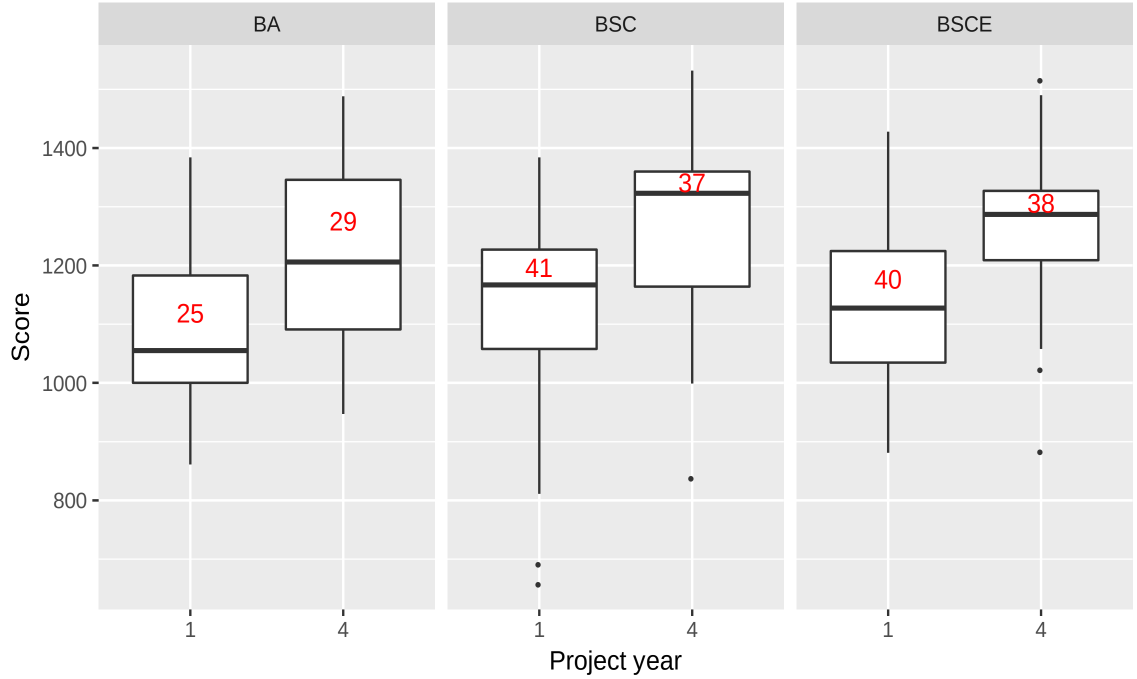


Table 9: Difference between Performance on the CLA+ with Cohen’s d Effect Sizes.

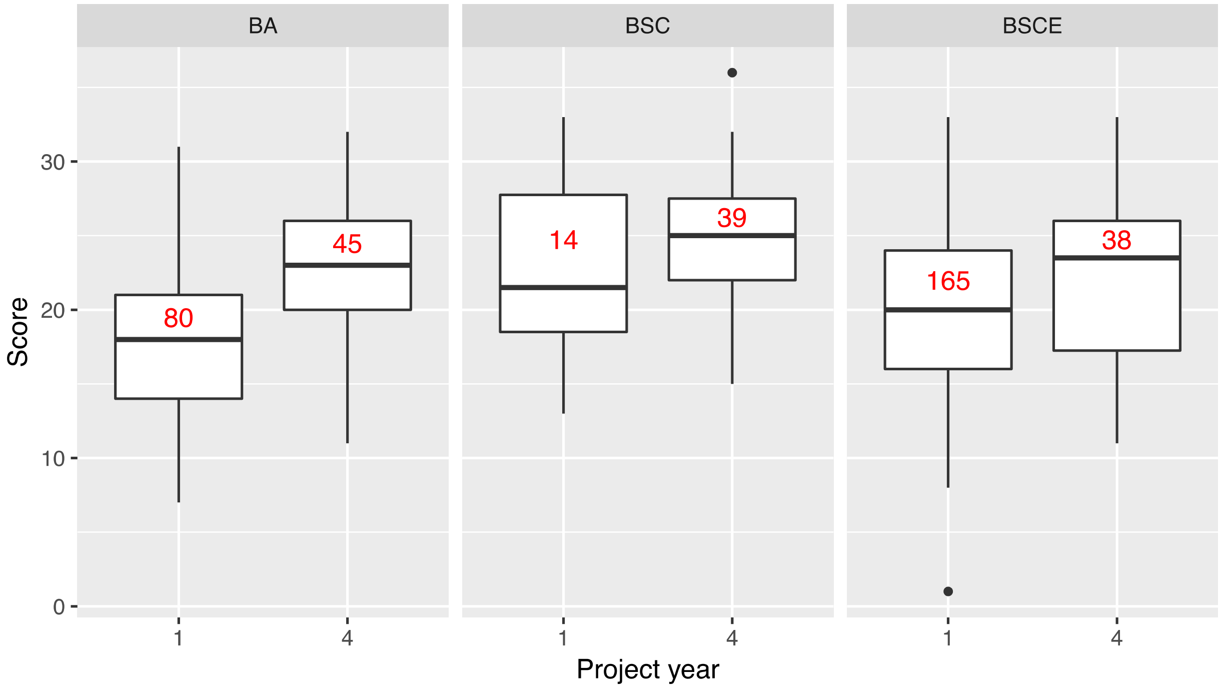
Group	t	df	p	D
Whole	3.57	242	<0.001	0.46
BA	2.63	52.0	0.011	0.71
BSc	1.15	75.4	0.25	0.26
BSc (Eng.)	2.70	74.5	0.008	0.61

CAT Results

There are 15 questions on the CAT, and 1,091 valid test results over the four years of the project. Analysis of Cronbach’s alpha found moderate internal consistency (a = .644). The first-year mean was 19.5 (SD = 5.5), (n = 291) and the fourth-year mean was 23.0 (SD = 4.7), (n = 133). Average means (but no standard deviations) were reported by the CAT provider for the lower-and upper-division groupings in the institutional reports. Differences were estimated using the Queen’s standard deviation as the denominator. Data from the whole Queen’s samples suggests that the Queen’s first-year mean was .84 of a standard deviation, better than the US lower division average; the Queen’s fourth-year mean was .76 of a standard deviation, better than the US upper division average. The distribution of CAT scores for the tested population, grouped by program, is displayed in Figure 7.

As was the case with the CLA+ sample, the CAT samples were not the same individuals across the four years. The fourth-year sample included 123 students (not the same individuals who completed the CLA+ in fourth year). As was the case with the CLA+ results, the highest fourth-year means were demonstrated by Bachelor of Science (BSc) students, but the largest gains were demonstrated by Bachelor of Arts (BA) students. Both of these groups include students in honours programs.

Figure 7: Box Plot CAT Scores Distributions by Program



Longitudinal change

Thirty students in the fourth-year sample completed the CAT in first year. The same process pairing and propensity matching process used for CLA+ analysis was applied to the CAT data, resulting in a pool of repeating and paired students shown below in Table 10. The corresponding t-test and Cohen's d results are shown in Table 11, and box-plots displaying distributions for programs with matched samples greater than 10 students are displayed in Figure 8. The effect size for year of study on CAT score mean was d = .65 (d = M2-M1/pooled SD).

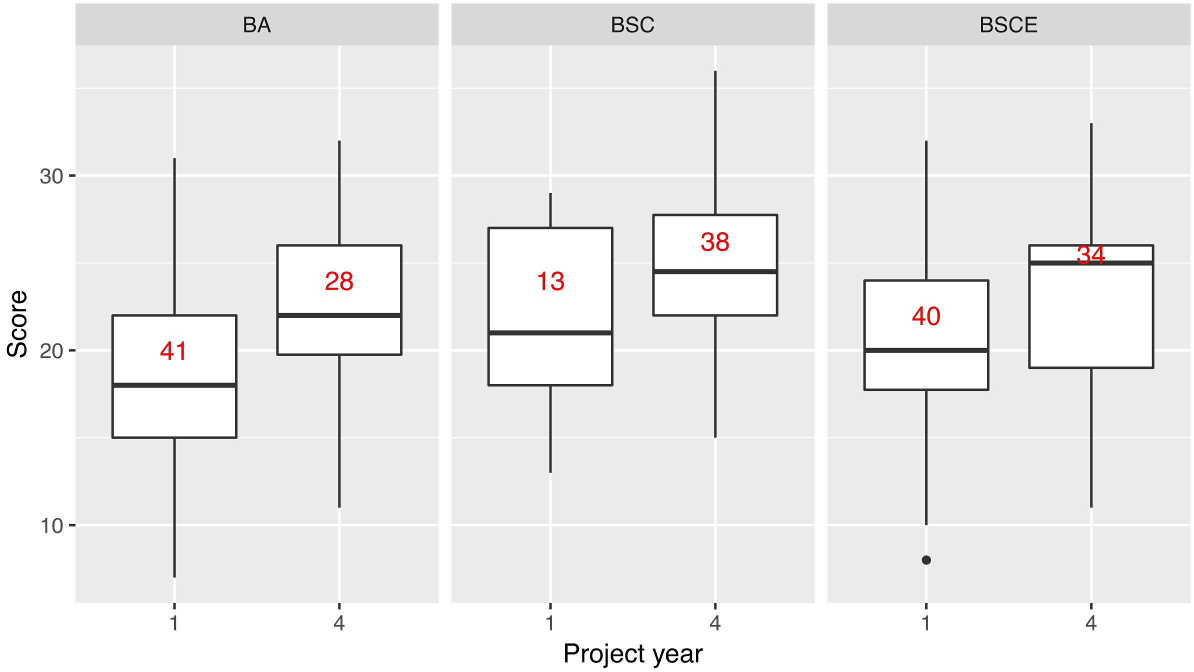
Table 10: Descriptive Statistics for the Pool of Repeating and Matched CAT Samples

Degree	Year	N	Sex		Language		FYGPA	
			Male	Female	English	Other	Mean	SD
BA	1	45	7	38	41	4	2.99	0.73
BA	4	32	3	29	30	2	3.19	0.84
BCMPH	1	2	2	0	2	0	3.80	0.71
BCMPH	4	3	2	1	3	0	2.74	0.64
BED	1	11	2	9	11	0	3.29	0.54
BED	4	4	1	3	2	2	3.09	0.34
BNSC	1	2	0	2	2	0	3.55	0.03
BNSC	4	2	0	2	2	0	3.57	0.00
BSC	1	17	8	9	17	0	3.13	0.79
BSC	4	42	7	35	39	3	3.12	0.74
BSCE	1	58	30	28	44	14	3.15	0.59
BSCE	4	52	25	27	45	7	3.13	0.61

Table 11: Difference between First- and Fourth-year Performance on the CAT with Cohen's d Effect Sizes.

Group	t	df	p	D
Whole	5.37	266.8	<0.001	0.65
BA	3.54	71.7	<0.001	0.80
BSc	1.74	25.7	0.093	0.54
BSc (Eng.)	3.14	105.2	0.002	0.60

Figure 8: Box Plots of Repeating and Matched First- and Fourth-year CAT Score Means



VALUE Rubric Results

Nine summative assignment types were identified as eliciting the skills of critical thinking, problem solving and written communication. Student artifacts from course assignments were collected and rated using the specified protocol, with two raters scoring each assignment and coming to a consensus (see section on VALUE ratings above for more detail). Generally speaking, the more samples of a particular assignment that was marked, the greater the rater agreement. But the complexity of the sample was also a factor. Some of the assignments entailed a two-page response and others a 40-page report. Included in this report are the summative assessments for courses in the participating departments. Table 12 displays the various assignment types and corresponding sample sizes.

Table 12: Assignment Type and VALUE Scoring Sample Sizes Per Year

Year group	Department	Assignment type	n
1st Year	APSC	Design Lab	39
	DRAM	Essay	90
	PSYC	Exam Question	106
	PHYS	Design Report	207
	TOTAL		442
2nd Year	PSYC	Performance Task	142
	APSC	Design Report	56
	PSYC	Design Report	46
	CIVL	Design Report	26
	ELEC	Design Report	19
	ENPH	Design Report	67
	MECH	Design Report	21
	PHYS	Design Report	10
	TOTAL		387
3rd Year	PHYS	Lab Report	9
	PSYC	Design Report	29
	CIVL	Project Proposal	36
	DRAM	Research Proposal	92
	ELEC	Design Project	32
	TOTAL		198
4th Year	MECH	Design	25
	ELEC	Proposal & exam	15
	PSYC	Thesis	15
	CIVL	Thesis	9
	TOTAL		64

The pre-calibration (independent scoring) agreement for first-year assignments was 64%, in second and third year it was 68% and in fourth year it was 57%. Following calibration, the agreement for first-year assignments was 99%, second year, 93%, third year, 96%, and in fourth year, 100%. A graphical representation of the pre-and post-calibration percentage of rater agreement by rubric has been included as Appendix 2. The greater initial spread in fourth-year scores required longer conversations, but ended up in higher final agreement. Cohen's Kappa was calculated for each of the assignment types at the individual dimension level. There was one unusually low reliability of K = .5 for the evidence dimension of the research proposal, but for the remainder, K > .7. For any dimensions under contention (where raters disagreed), the score was rounded down to the lower level under the theory that the level of the dimension had not been demonstrated in its entirety.

Validity and internal consistency

Correlations between VALUE dimensions and sessional GPA were calculated as a measure of convergent validity. Significant correlations between the critical thinking, problem solving and written communication as evaluated on the VALUE rubrics, and students' academic achievement (see Table 13) suggest that there was a relationship. However, the low coefficients also suggest that the GPA captures more than these complex cognitive skills. A general descriptor table was composed such that researchers were able to discuss dimensions of the Critical Thinking, Problem Solving and Written Communication VALUE rubrics with instructors without getting sidetracked by the detail of specific criteria contained in each of the rubrics. The descriptors are included in Table 13.

Figures 8, 9 and 10 display the score distributions in percentages of sampled population for each of the rubrics over each year of the project. Further investigation of central tendencies found that the Critical Thinking first-year median was Benchmark 1, second- and third-year median was Milestone 2 and fourth-year median was Milestone 3. For Problem Solving and Written Communication, the medians were Milestone 2 for first and second year and Milestone 3 for third and fourth year. The changes from first to fourth year were significant for all rubrics, using Wilcoxon rank sum test with continuity correction.

Table 13: Internal Consistency and Validity of VALUE Scores

VALUE rubric	Dimension	Descriptor	n	Correlation-Sessional GPA	a
Critical Thinking	Explanation of issues	Issue/problem considered; relevant information necessary for understanding	1174	.260**	.886
	Use of evidence	Selecting and using information to investigate a point of view or conclusion	1096	.263**	
	Context and assumptions	Own and others' assumptions and evaluates the relevance of contexts	1054	.279**	
	Students position	Subjective/ objective perspective-thesis/hypothesis	1004	.281**	
	Conclusions and outcomes	Evaluates consequences and implications	1080	.290**	
Problem Solving	Define problem	Contextual problem statement	975	.272**	.804
	Solution hypotheses	Multiple approaches	917	.284**	

VALUE rubric	Dimension	Descriptor	n	Correlation-Sessional GPA	a
	Identify strategies	Comprehension, sensitive to contextual issues	857	.260**	
	Implement solution	Considers history of problem, reviews logic/reasoning, examines feasibility of solution and weighs impacts of solution	575	.253**	
	Evaluate outcomes	Addresses multiple contextual factors (implementation)	624	.265**	
	Evaluate solution	Reviews results relative to the problem defined with considerations of need for further work	984	.287**	
Written Communi-cation	Context and purpose	Audience, purpose and the circumstances surrounding the writing task(s)	1160	.341**	.856
	Content development	Uses appropriate and relevant content	1151	.366**	
	Sources of evidence	Formal and informal rules inherent in the expectations for writing in particular forms and/or academic fields	1075	.198**	
	Genre and conventions	Demonstrates use of credible, relevant sources to develop ideas that are appropriate for the discipline and genre of the writing	1113	.285**	
	Syntax and mechanics	Uses language that communicates meaning to readers with clarity and fluency	1165	.299**	

** Correlation is significant at the 0.01 level (2-tailed).

Note: Correlation calculated using Spearman’s Rho

Figure 9: Distributions of Scores by Year on Each Level of the Critical Thinking Rubric

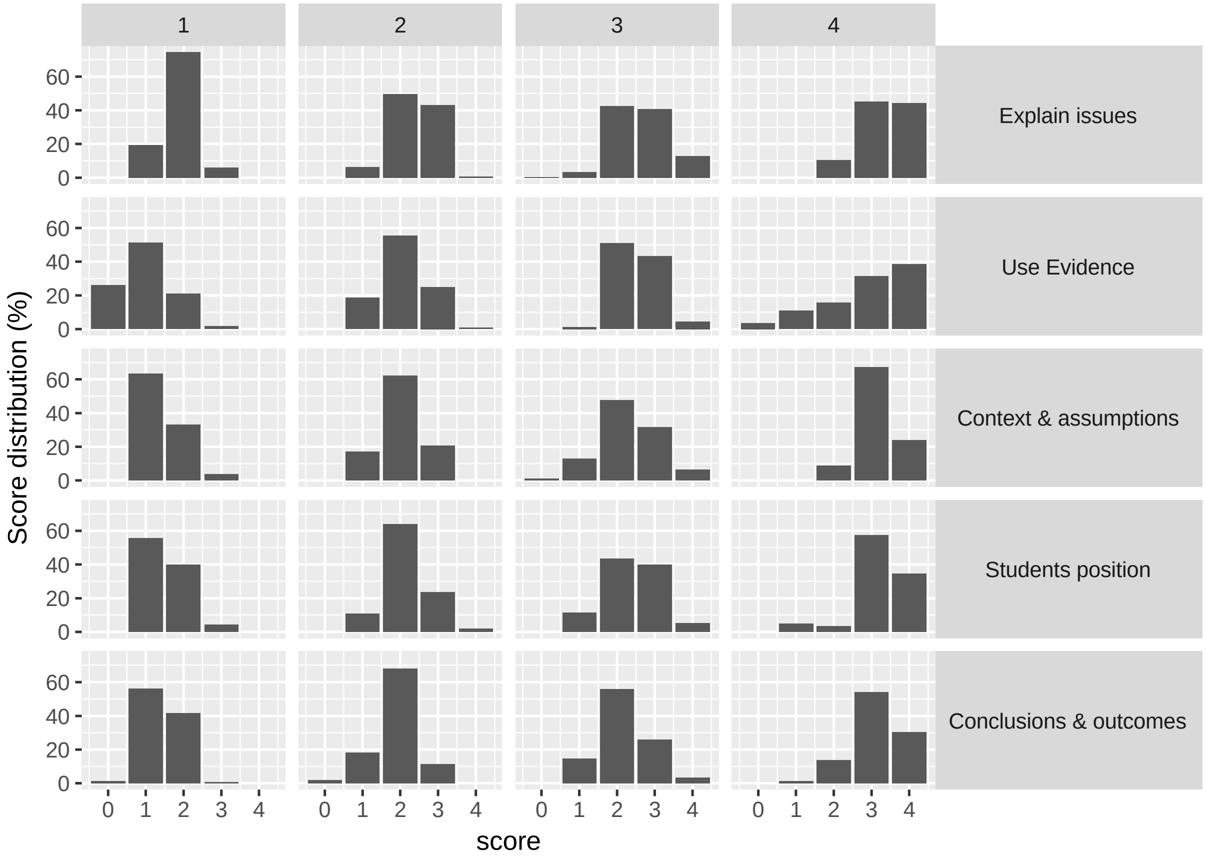


Figure 10: Distributions of Scores by Year on Each Level of the Problem Solving Rubric

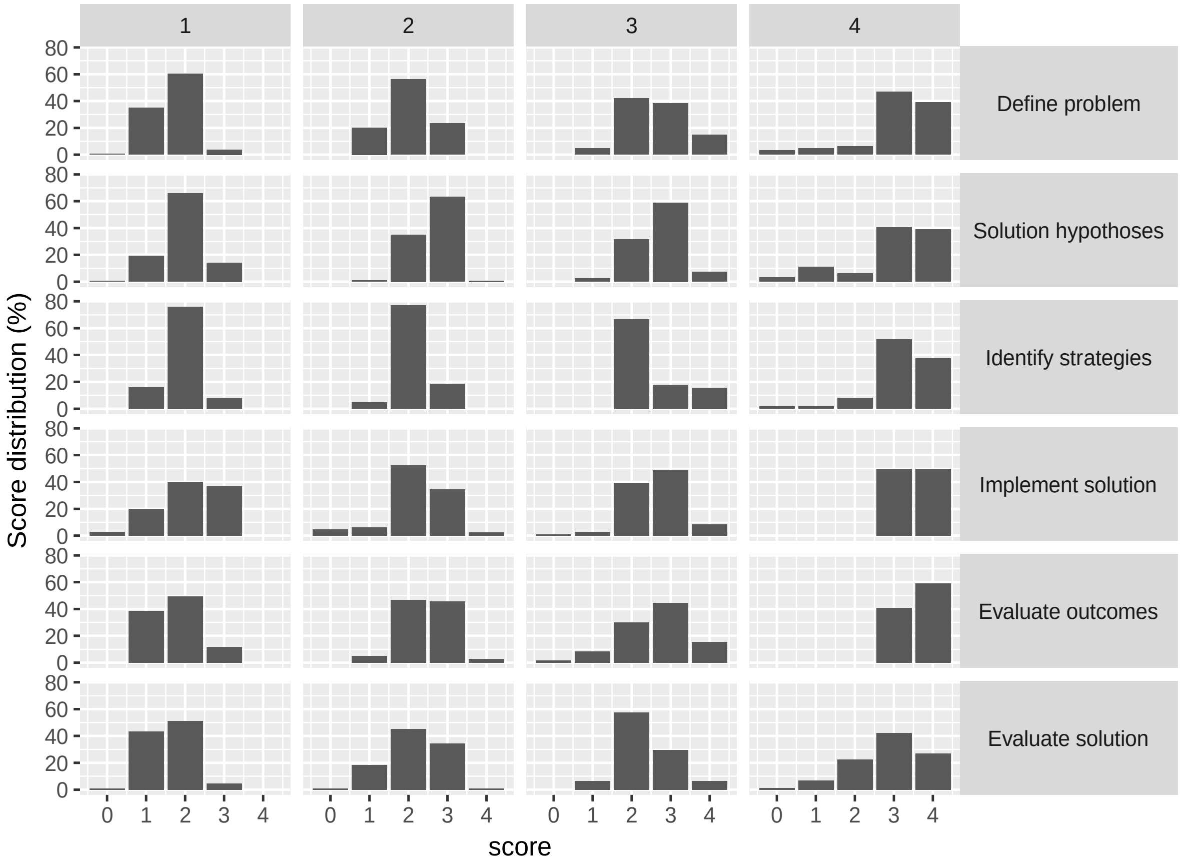
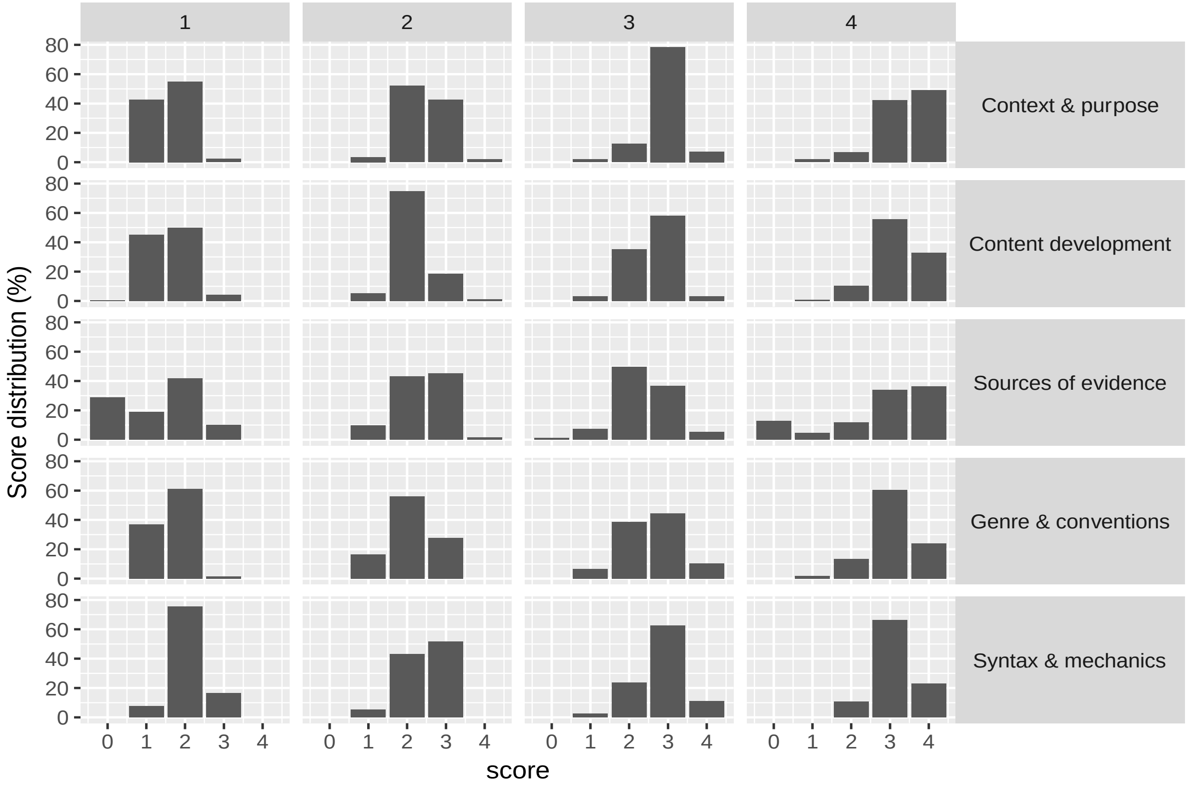


Figure 11: Distributions of Scores by Year on Each Level of the Written Communication Rubric

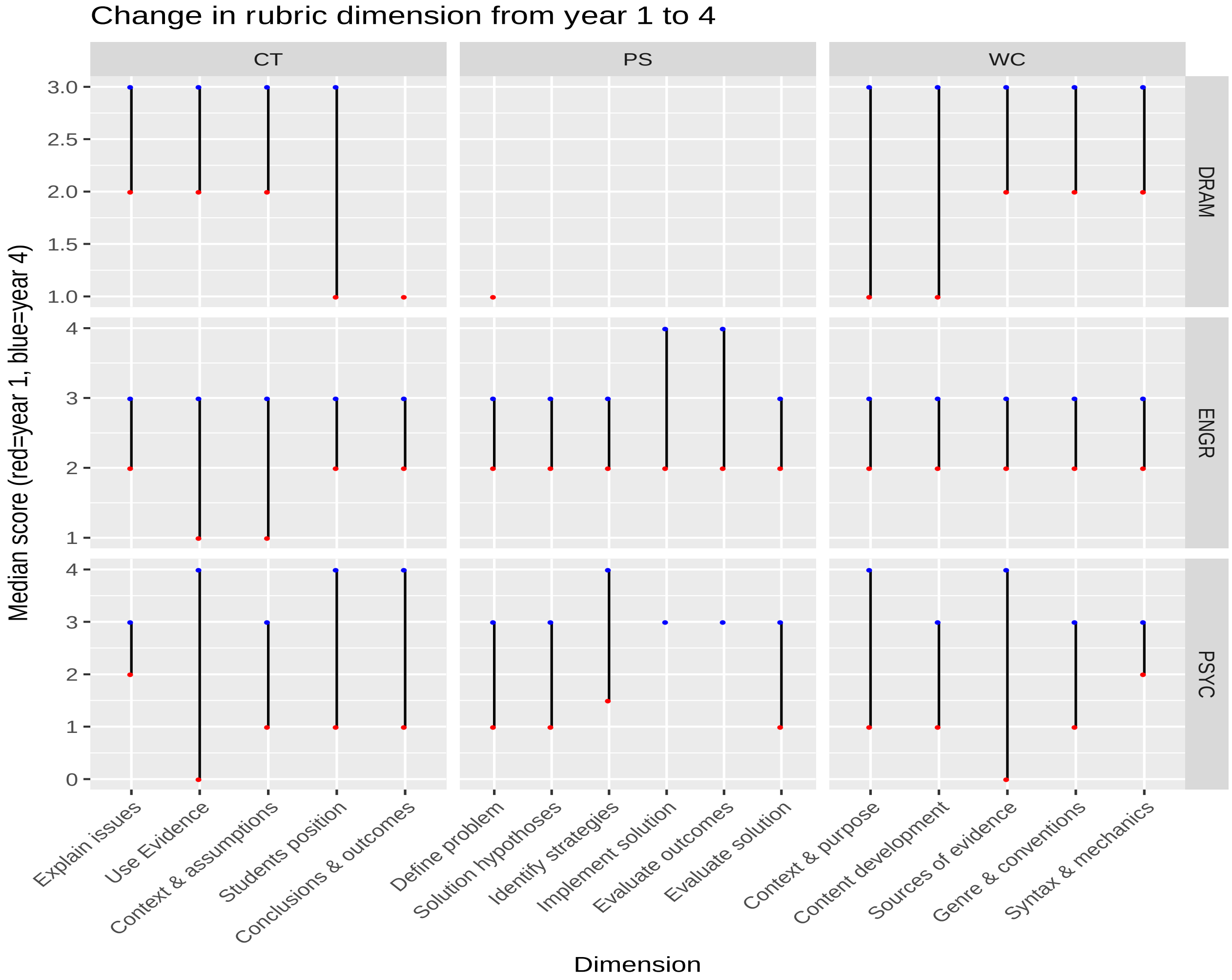


The changes from first to fourth year were significant for all rubrics, using Wilcoxon rank sum test with continuity correction. Table 14 shows the test statistic and effect size *r* by rubric. A comparison of the median score in years one and four is shown in Figure 12.

Table 14: Significance of Gains on Each Rubric from First to Fourth Year

	W	***p***	***r***
Critical Thinking	54424.5	<0.001	0.61
Problem Solving	69040.5	<0.001	0.50
Written Communication	92994.5	<0.001	0.60

Figure 12: Change in Median Score on Each Dimension



Alignment of course assignment to assessment criteria

Some of the assignments had greater alignment to the assessed dimensions than others. If the assignment did not elicit the demonstration of evidence for a particular dimension (step 1 in the scoring process), that dimension was not assessed. Generally, the greater the alignment, the higher the VALUE rubric score. The exceptions were the first-year exam question, where students' time and word-count were limited, and one of the second-year design reports, which was elicited from a technical subject and was not readily suited to assessment on the VALUE rubric dimensions. Generally speaking, the more dimensions assessed, the higher the median score on the rubric. Appendix 3 provides a colour coded graphic, displaying the relationship between the percentage of dimensions assessed plotted against the average level of achievement, grouped by year and by subject. The greater the slope of the line, the more significant the relationship.

Correlations between Instruments

Because we were working with various data types, Spearman's Rho rank order correlations were used for comparisons of the instruments and biographical data (sex and language) and grade point averages. Median values were calculated for the VALUE rubrics as they comprise ordinal data. The sex was nominally coded as male = 1, female = 2; language was coded by first language, English = 1, French = 2 and other = 3. The results displayed in Table 15 are based on cumulative data across the years of the project and were calculated using pairwise analysis. Correlations between GPA, sex and language were not included in the table to avoid sample conflation of pairwise comparisons (i.e., the data was constrained by year, paired by instrument, with valid data points included).

The VALUE rubric scores were highly related, but there was a weak relationship between the test scores and rubric scores. The highest correlation was between the CAT score and the Problem Solving rubric score (*r*(303) = .234 *p* < .01). The VALUE rubric scores were more closely related to the students' cumulative grade point average than either the CLA+ or the CAT scores. The inference from the negative correlations between sex and VALUE rubrics scores would be that males did marginally better than females on the course-based assignments. The relationship between sex and test scores, however, was not significant. Language was not significant in the VALUE rubric scoring, but there was a weak correlation suggesting that English speakers performed slightly better on the CAT and on the CLA+ selected-response questions than the French or other language students. (Note: There were eight individuals who took both the CLA+ and the CAT in the same year.)

Table 15: Correlation between Instrument Sub-scores Biographical and Grade Point Average

		VALUE rubrics			CLA+			CAT score
		CT	**PT**	**SR**	**PT**	**SR**	**T**	
VALUE rubric- median value (n)	Critical Thinking (CT)	-						
	Problem Solving (PS)	.720** (1051)	-					
	Written Communication (WC)	.767** (1183)	.715** (1050)	-				
CLA+ (n)	Performance task (PT)	.031 (366)	.002 (341)	.043 (365)	-			
	Selected response (SR)	.202** (351)	.198** (326)	.231** (350)	.182** (1131)	-		
	Score total (T)	.155**	.149**	.179**	.714**	.792**	-	

	VALUE rubrics			CLA+			CAT score
	CT	PT	SR	PT	SR	T	
	(347)	(322)	(346)	(1131)	1131		
CAT score (n)	.133**	.227**	.191**	-.829*	.902**	.659	-
	(384)	(306)	(384)	(8)	(8)	(8)	
Sessional GPA (n)	.304**	.357**	.348**	.145**	.260**	.269**	.253**
	(1184)	(1051)	(1183)	(1181)	1135	1131	1089
Sex (n)	-.087**	-.166**	-.101**	-.029	-.012	-.024	-.024
	(1184)	(1051)	(1183)	(1181)	1135	1131	1091
Language (n)	-.023	-0.019	-.028	-.012	-.075*	-.052	-.108**
	(1135)	(1002)	(1134)	(1114)	1068	1064	1033

** Correlation is significant at the 0.01 level (2-tailed).

Comparison of Costs

The associated costs of using the CLA+, CAT, and VALUE rubric approaches were based on a nominal sample of 100 students (see Table 16). They were calculated by adding the fee for the instrument, the ancillary costs (training fees and salaries) and/or the salaries for markers (undergraduates were paid C$14 per hour and graduate markers were paid C$24 per hour). Although the fee per test-taker was US$35 for the CLA+ and US$9.95 for the CAT, once the additional costs were taken into account (and using an exchange rate of C$1.2), the total costs were C$51 and C$47.54 respectively. The VALUE samples took varying amounts of time to mark, ranging from 30 minutes to three hours depending on the complexity of the assignment. Undergraduates marked the majority of first-year samples, whereas the fourth-year samples were marked primarily by researchers or graduate students (see Table 4). The average cost across all of the work marked was C$32.

Table 16: Comparative Cost of Each Instrument (Canadian Dollars)

	Training/ or technical support	Instrument fee per 100 students	Test proctoring (4 sessions - 100 students)	Marking costs	TOTAL	Cost per student
CLA+	$100.00	$4,200.00	$800.00	-	$5,100.00	$51.00
CAT	$1,000.00	$1,194.00	$560.00	$2,000.00	$4,754.00	$47.54
VALUE marking	$200.00	-	-	$3,000.00	$3,200.00	$32.00

Qualitative Components

Qualitative methods were used to investigate the perceived benefits of each of the tools. Two interview sessions were conducted, which involved four instructors from two departments who each taught a course involved in the project. Each instructor was provided with a summary report and was invited to provide comment. Where available, the debrief sessions were recorded and transcribed for tracking and evaluation. We had very positive comments from instructors regarding involvement in the project and the feedback they received:

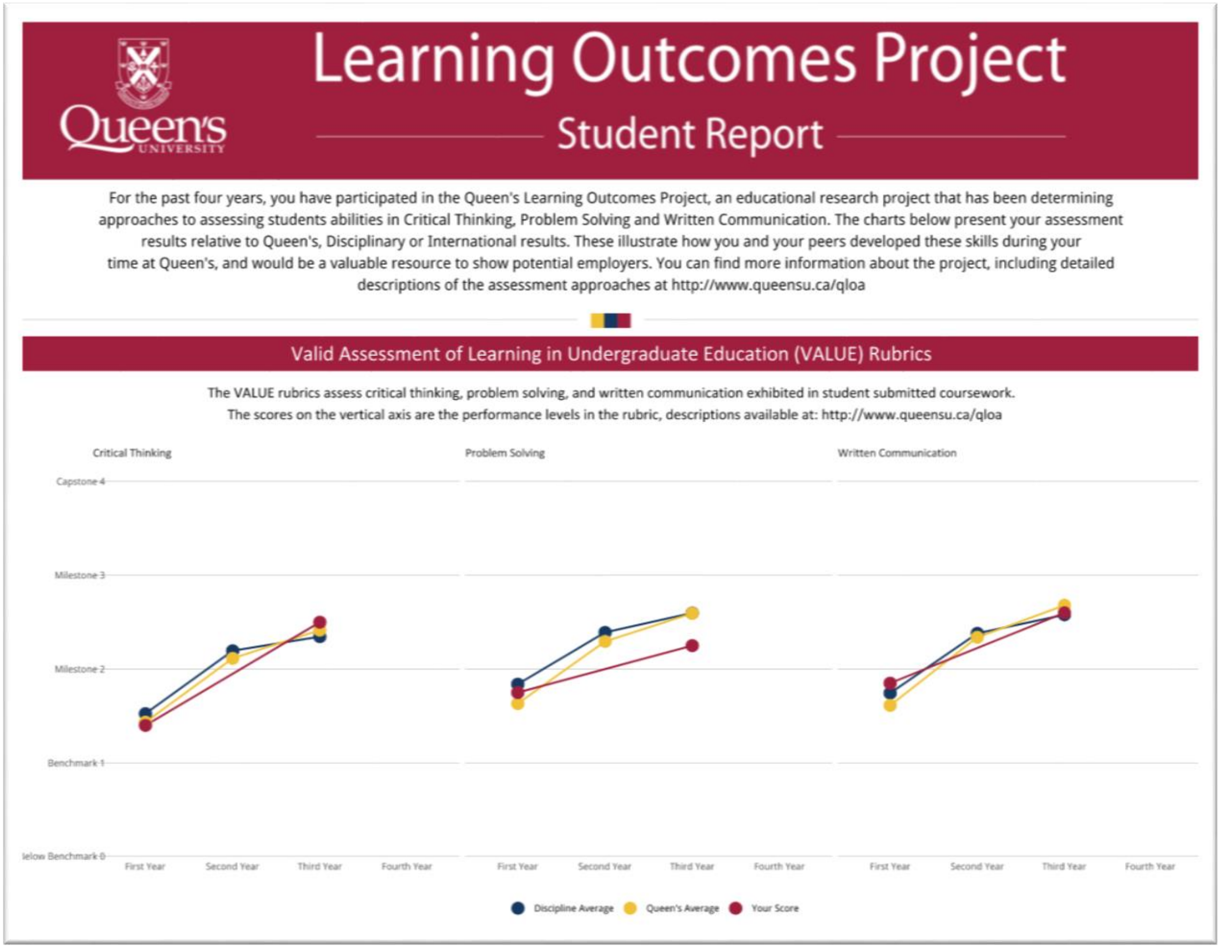
> “I feel like we got something of immense value for free, from our departmental point. Right? Like we just got a huge amount of valuable information for instructional design. That’s like gold. And it was outstandingly valuable, really exciting. What's more valuable is a separate set of eyes on this. Like, I’ve been assessing this assignment for five years; it’s taken me five years to realize what you did in one go around.” (First-year course instructor)

The provision of feedback to instructors often had the effect of prompting reflective questions. Common questions that arose were, “What are the desirable characteristics/skills for students?” “Should I give them the step-by-step plan for resolution (scaffolding)?” “What information is useful for the students?”

This project was not intended to be a study of educational interventions, but it was still important to note the impact of the project on educational improvement efforts. Through self-reflection and by providing student feedback from the course assignments, many of the instructors involved in the project chose to adapt new assessment strategies or adopt new teaching strategies to better target student cognitive-skill acquisition. A list of specific changes made to courses is provided in Appendix 4.

In addition to the instructor debriefs, a series of student group interviews were conducted each year to explore topics of interest to the investigation. By the second year of the project, students had expressed interest in accessing their individual results. During the third year of the project, a format was developed to provide students with their score on the VALUE rubrics on one page and on the reverse, results from the CAT and CLA+ were shown. The data points on the report displayed the individual score (if available), the discipline average and the institutional average. In the third year of the project, students provided feedback on the format and refinements were made. Figure 13 displays the front page of the report.

Figure 13: Example of a Third-year Student Outcome Project Report



Learning Outcomes Project

Student Report

For the past four years, you have participated in the Queen's Learning Outcomes Project, an educational research project that has been determining approaches to assessing students abilities in Critical Thinking, Problem Solving and Written Communication. The charts below present your assessment results relative to Queen's, Disciplinary or International results. These illustrate how you and your peers developed these skills during your time at Queen's, and would be a valuable resource to show potential employers. You can find more information about the project, including detailed descriptions of the assessment approaches at http://www.queensu.ca/qloa

Valid Assessment of Learning in Undergraduate Education (VALUE) Rubrics

The VALUE rubrics assess critical thinking, problem solving, and written communication exhibited in student submitted coursework.
The scores on the vertical axis are the performance levels in the rubric, descriptions available at: http://www.queensu.ca/qloa

Student Perspectives

In the final year of the project, students who had expressed a willingness to participate in group interviews were recruited (there was an option for this on the consent form). There were three group interviews with a total of nine students, six women and three men, from the participating Engineering, Physics and Psychology departments. The nine students generally had positive past experiences with standardized tests. The purpose was to gather information from the students about their experiences with taking the Collegiate Learning Assessment Plus and the Critical Thinking Assessment Test. Specifically, our goals were to garner a better understanding of students' perspectives on the following:

- Their motivations for participating and putting effort into the tests
- Their perceptions of low effort and motivation for taking the tests

- Their insights on effective recruitment strategies and for increasing student motivation for taking the tests in the future

In the following sections, we describe the results of the group interviews and articulate our understanding of students' perspectives associated with each of three goals identified above.

Goal 1: To understand students' motivations for participating and putting effort into the tests

For all the students we interviewed, there was some evidence of intrinsic and extrinsic motivation for taking the tests and for putting effort into the tests. To a small degree, some students saw professors' buy-in and promotion of the tests in their classes as a motivator for participation. According to one physics student, this was a factor:

> "I think it depends on how their [the students'] prof frames it [the test]. Because for us, the professor obviously thinks this is very important, and the way that she describes it also, I think, has an effect on whether or not students feel the need to participate." (Group 1)

Other students who mentioned professors' buy-in as a source for motivation gave reasons that included how the professor presented the tests to them or whether the professor gave the impression that they thought the tests were important. More prominently, students' motivation was related to their disciplinary content knowledge and their experiences with research in their respective disciplines.

Students said that their motivation for taking the test was linked to their prior experiences, which included a lack of familiarity for some with the particular subject content covered in the test questions. For example, students indicated that test questions that focused on content knowledge and problem-solving processes from their subject area were more relatable and engaging for them. An engineering student noted,

> "I think it depends on your past experiences. I just switched into engineering last year. I was taking a lot of philosophy courses and stuff like that before. So, it [the test question(s)] kind of played into my interest." (Group 2)

Another psychology student responded, "I was able to apply what I had learned in those classes in the questions that were being asked" (Group 1). Conversely, a student who could not relate as easily to the test questions commented, "I didn't like the ones where it was medical stuff and reading that and drawing different things from that . . . but it could have been I just don't have knowledge in that subject area so I found it harder to draw connections" (Group 4). Psychology students in particular articulated that their motivation was connected to their experiences in psychology, where participation in research projects and research labs is often a requirement for their degree program. One psychology student commented, "I volunteer in a psych research lab and I know how hard it is to get participants to participate in things. After having that experience it made me . . . want to participate" (Group 1). These students said that they knew how difficult it was to get participants to do research so this was a motivator for them to get involved and put effort into the tests.

Finally, students' motivation was also linked to their enjoyment of the test questions that offered them opportunities to apply problem-solving and critical-thinking skills to novel situations. As part of the group interviews, students were asked to talk about their experiences inside and outside of their academic work where they had to think critically and apply problem-solving strategies. All students, in some manner, drew comparisons between the test questions, their academic work that required them to think critically and problem-solve, and their problem-solving and critical-thinking experiences outside of academic work. Students who felt that the tests were interesting and challenging said that this was mainly because of the open-ended nature of the questions and that there was no right answer. For example, there was a common thread that students were motivated by learning experiences that encouraged them to find information and solutions to problems in their own way. This was consistent when they talked about the test questions, their academic experiences and their non-academic experiences. In one student's experience with the tests she noted, "There's a lot of different ways to come up with the problem. And then you also get to present it in an intelligent way" (Group. 2). Similarly, a student, commenting on a class project, said, "I like having the pressure to be able to fulfill your [sic] project. It gives you the drive to find the information in your own way" (Group 2). Another student commented on a summer project that was connected to his field of Engineering:

> "They [the Faculty of Engineering] brought me on the solar team this year. I was thrown into just this new project idea and they wanted me to look into it and do a little bit of construction and stuff like that. Again, it's a different way of going about a problem when you're actually trying to put something real down, not just talking about it theoretically." (Group 2)

A student who was majoring in both Psychology and Education commented on her experiences of problem solving while on her practicum placement outside of school:

> "Outside of the [university] classroom … I'm on practicum placement and problem solving and critical thinking is a must; there's 30 students and you just have to figure it out. You're pretty much forced to problem-solve on the spot. And I think that's also an important skill. There's the experience of having to do it immediately and then also being able to do it in a more academic, less practical sense. (Group 1)

Interestingly, students did not initially make the connection between how their problem-solving and critical-thinking experiences outside of their programs actually transfer into their thinking in their programs (and vice versa). When prompted, these students were able to see this connection more clearly. We concluded that most of the students employed similar critical-thinking and problem-solving skills in their experiences outside of academic programs and in some classes, as they encountered on the tests. This helps to explain why they expressed interest and engagement in test questions that allowed them to apply novel solutions to complex problems. However, from these conversations with students, we understood that they often distinctly separate learning that happens in their academic classes from learning that occurs outside of these classes whether or not they are adopting the same or similar learning processes. It is plausible that the problem-solving and critical-thinking test questions — given in an academic setting yet focusing on content outside of their academic disciplines (at least, for many students) — act as a bridge between these two learning contexts. Put another way, the tests act as a third space between students' academic and non-academic contexts, from which they can become more aware of cognitive skills and learning processes that are inherent to both, and that we try to foster in 21st century learners in higher education today.

Goal 2: To understand students' perceptions of low effort and motivation for taking the test

The students we interviewed offered some key information about why student effort and motivation might be low on both the CLA+ and the CAT. Some of the major reasons included that students were contending with competing demands such as having to focus on other school work and the lack of time in their schedules. A common rationale was that when students were busy with other commitments related to their degrees (such as projects and labs), they were not as motivated or as likely to participate in external tests. Students also surmised that motivation was affected because of the early time of day and the busy time of the school year in which the tests were offered. One student said, "Because it [the test] was offered at 8:30 on a Monday and I'm not a morning person. I think that timing does matter quite a bit in terms of performance." (Group 2). Another person gave the following reason: "I really think it depends on what else is going on. This semester we had just had two or three weeks of just insanity and so no one would have showed up" (Group1).

Other explanations for low motivation had to do with the repetitive nature of taking the same tests over again, that the tests required a lot of writing, that the multiple-choice questions became arduous, and that they got bored and tired by the end of the test. In the case of the CLA+, students commented on the fact that the group dynamics matter a great deal: "If you're with people who don't want to do the project and don't want to put effort in, then it's a lot of pressure," said one student. Students had a more positive view of the CAT and they felt it was more representative of their individual effort and abilities. According to some students, the VALUE rubric associated with the CLA+ did not account for unequal distribution of effort and this was seen as potentially discouraging for some students.

Goal 3: To garner students' insights on effective recruitment strategies and for increasing student motivation for taking the tests in the future

Finally, we sought students' insights into how we might improve student recruitment and participation, and increase effort on the tests. All students stressed the idea that test administrators could be more explicit about the benefits of the tests to the test-takers. For instance, students favoured the idea of being able to see how they compare to other students in their departments and, more broadly, across the whole university in order to benchmark their own academic successes. One student brightly suggested that knowing these scores would help students to market themselves for future job opportunities: "If it was one [a score] they could post like a qualification, I think that would motivate a lot of people that wouldn't normally be interested [in taking the tests]" (Group 2).

There was also a strong suggestion among our student participants that we change the time when the test is offered in the school year and even introduce flexible hours for people to take the tests both inside and outside of class time. They also strongly suggested that students receive some kind of credit, bonus marks or extra percentages in the course in which they take the tests, or make the tests count as part of their actual assessed course work. According to one psychology student:

> "If you want to provide feedback and if you wanted to really make sure that you have the students engaged and performing at their best, you have to make it a course assignment. Only because if you

> do that you're going to ensure that people participate and will be putting in their best effort." (Group 3)

Students suggested other incentives to increase student motivation and effort, including offering free food and small monetary gifts. One of the students commented that while money might be an incentive, it may not generate better quality of participation and effort because the tests are not associated with their grades. Collectively, the comments suggest that the most important factors to consider when promoting motivation and effort for testing were:

- The instructor valuing the test
- Ensuring test content familiarity
- Effective timing
- Providing feedback on achievement
- Making the test score count
- Providing monetary incentive

Discussion

Development of complex cognitive skills over a four-year degree

This report focuses on the implementation of tools to evaluate student achievement of complex cognitive skills across a four-year degree. Although there was an underlying motivation to encourage skills development, the primary concern of the exploratory study was to determine a viable way of assessing the true ability of students' skills over time and across the institution. The chosen instruments comprised similar assessment constructs and the investigation of each found significant gains in student learning. While standardized tests were the most reliable approach from a psychometric perspective, the drawback was student motivation and poor alignment with the activities that instructors and students really cared about in their classes.

Queen's University's incoming students perform significantly better than incoming US students, largely due to the selective nature of Queen's. With high initial levels of critical-thinking, problem-solving and written communication skills, the challenge is to demonstrate how much they improve over a four-year degree. Researchers recognize that many factors contribute to the development of these skills, including extra- and co-curricular activities, and life skills gained as students mature. The scored assignments and tests in this research were all written. There is work to be done in investigating the assessment of these skills in oral and information communication formats.

Assessment can be compromised by a variety of factors; in the case of standardized tests, these include test-scheduling constraints, effort and incentives, and, in the case of using program rubrics to score academic work, they include differences between assignment expectations and varying alignment to assessment constructs. The in-class testing was conducted according to course schedules; as a result, test sessions were occasionally run at sub-optimal times of the day, days of the week or weeks of the term. The CLA+ test window did not open until March, which we found to be a poor time as it overlapped with the mid-term examinations. Many of the participating courses offered their one-hour lab or tutorial slot for student testing, which prevented those courses from administering the 90-minute CLA+ in class. Researchers initially recruited students to take the test outside of class time, but attendance at out-of-class testing was very low and therefore sample bias was a concern. For students who were not scheduled in a computer lab, they either had to move to a different room or use their own computers. Moving proved to be time consuming and disruptive for students, and using their own laptops presented a range of technical issues. Issues like these can impact student motivation and performance.

It is also important to consider that standardized tests do not specifically contain content directed at a particular discipline. If the test content does not interest the student, or if students are in an unfamiliar space or don't have enough time to complete the test, then motivation can be impeded. In first year, all the students who took the test had their attendance rewarded with a course percentage mark. As effort level dropped, researchers looked to alternate methods for incentivizing students. Financial incentives did yield improved effort, but the cost may be prohibitive.

Leaning gains were most detectable using the VALUE rubrics. The weakest area for the Queen's first-year students was critical thinking, with the majority of first-year students demonstrating at the Benchmark 1 level, whereas the first-year median performance of Problem Solving and Written Communication was Milestone 2. Critical-thinking outcomes also showed the greatest improvement. The majority of students were at the Milestone 3 level by fourth-year. Over the four years of their undergraduate programs, students demonstrated sizable gains on all the VALUE outcomes.

Non-parametric analysis was used for the VALUE data in this report because the results were not derived from scale data. That is to say, there is not the same improvement from Benchmark 1 to Milestone 2 on the VALUE rubrics as there is between Milestone 3 and Capstone 4. The Capstone level criteria on the VALUE rubrics is fairly aspirational. By way of analogy, think of the terrain map of a ski hill, then consider the VALUE levels as the difficulty of ski runs. Benchmark 1, for example, would be a green run; with a little bit of practice, anyone can do it. Then liken Milestone 2 to an intermediate blue run; you need some training but it's not too hard. Then consider Milestone 3 as a black diamond run, where a moderate level of expertise is necessary, but so long as the run is groomed, it is quite achievable. Finally, consider Capstone 4 as an ungroomed double black diamond. In this terrain, you really need to be an expert. However, if the students have a guide, say a supervisor in a thesis or design project to show them the way, and they employ a great deal of rigour and effort they may achieve a Capstone 4 level.

Achievement in fourth year differed depending on the amount of support the students had. This difference introduced error to the analysis. Another point of error arose from the course weighting for the assignment. It was hardly fair to compare a weekly lab report to a year-long research thesis. Our marking protocols describe the process for evaluating the assignment to determine which of the dimensions were valid to assess. Part of the reason for this was to minimize these differences. That is, we only assessed dimensions of the rubric when the assignment elicited demonstrations of the criteria. It was unusual for an assignment with low course weighting to elicit demonstrations across all of the assessment dimensions. The relationship between assessed dimensions and score suggests that there is still work to be done aligning the course assessment to the VALUE rubrics.

Skills development between programs and individuals, and the relationship to grades

Results from the CLA+ and CAT suggested that the Bachelor of Science (BSc) students performed at the highest level and that Bachelor of Arts (BA) students consistently saw the largest gains. It was more difficult to directly compare programs based on VALUE results because the assignment types varied. The students from the research thesis courses (one in an Engineering course, the other in the Psychology thesis course) achieved the majority of the Capstone 4 outcomes.

Gender was not a significant factor in the sampled population, the inference being that the instruments do not contain gender bias and that gender does not influence cognitive development. As might be expected, students with English as their first language performed marginally better on the standardized tests. These tests were offered only in English, and more than one student reported that the test would have been much easier for them if they could have taken it in Mandarin or another language. Language was not, however, a factor for the course assignments. With course assignments, there was a longer duration for completion, where effects of language were mitigated.

Scores from all of the tools correlated with sessional GPA, but the VALUE rubrics were most closely related. It could be argued that the student GPA reflects the sum of the course learning outcomes, the knowledge, behaviours and habits of mind that we consider important for students in preparing them for their future. Depending on the assessment structures, this might not be entirely true, but we can assume at least that the GPA reflects some, if not all, declarative and technical knowledge, procedural skills, teamwork or professionalism, in addition to critical thinking, problem solving and communication. For that reason, we would not expect high correlations between GPA and the critical thinking, problem solving and written communication as evaluated on the CLA+, CAT and VALUE scores. We would, however, seek significant correlations and place more weighting on the assessment instruments that most closely reflect the course outcomes. Coefficients for VALUE rubrics were all greater than .3, whereas the CAT and CLA+ coefficients were less than .3.

It is also important to note that course assessment strategies and processes differ greatly and further work is needed to determine what component of course grading reflects particular assessment constructs. There were assignments offered for inclusion in the project that could not be marked using the Critical Thinking, Problem Solving or Written Communication VALUE rubrics because there was little or no alignment with the assessed dimensions. This was particularly so in courses that were very technical in nature, with little flexibility to adapt any of the course assessments to specifically target learning outcomes.

Cost and time-efficiency of instruments

Researchers faced difficulties in coordinating technical and logistical requirements necessary for standardized testing. Even when researchers managed to get the students to take the tests, the students didn't necessarily put in their best effort. This was problematic from a reliability standpoint; we didn't know if the students were capable of better results. In the final year of the project, monetary incentives were used in an attempt to minimize this problem, but fiscal as well as logistical considerations for assessments of this type are a concern for sustainability. The CAT tests must be marked using a rigid protocol, which requires trained markers and sufficient time, then sent to Tennessee for calibration. Data from the CLA+ and CAT tests are not available until approximately two months after testing. Delays in receiving data make it impossible to use the test score for a low-stakes course assessment, and also make it difficult to use the data for course improvement. 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. Data from the tests is confidential and both test providers have security protocols in place to ensure that the data is not inappropriately accessed. These tests provide data that can be confidently aggregated across disciplines and can also be used for comparisons to US institutional averages published by the test provider.

Working toward long-term sustainability, the research team is focusing much of its time and effort working on the ground with existing faculty to create instruments that minimize any additional workload on their part and to create instruments that are virtually invisible to the student. We have a rising level of awareness about assessment of transferable learning outcomes within the departments involved in the project, and are running information sessions and workshops with faculty to share our understanding with a broader audience. This understanding is a necessary foundation for a wider-scale rollout. Discussions have begun regarding possible inclusion of additional departments for involvement of general learning-outcomes assessment.

During the study, we found anecdotal evidence that some faculty and instructors were open and keen to implement new assessment processes, but that many were wary of change. Some faculty members felt that these intellectual skills were captured within their current assessment methods, and they remained unconvinced about the need to specifically assess these skills. On the other hand, the instructors involved in the study found tangible benefits in the provision of student achievement data, and many used the information to make improvements to their courses.

Additional outcomes and impact

Communication and professional development

- A project website was developed (http://www.queensu.ca/qloa/home)
- The project team facilitated workshops, presented webinars and presented at conferences across Canada and US
- Facets of the project were published in *The European Journal of Engineering Education*, and *Assessment and Evaluation in Higher Education* (details available here: http://www.queensu.ca/qloa/resources-and-research)
- Task development workshops have been run each year to build common understanding of expectations and structures for assessment of critical thinking and problem solving

Development of rubric building tool

The VALUE rubrics do take some time to become familiar with. The language in the descriptors is very high level and general. The project researchers used them in their published form, but course instructors found them difficult to relate directly to their course assignments. In the second year of the project, researchers used assignment annotations to adapt the VALUE rubrics to include department-specific language. The resulting rubrics were still very general, and were no more useful to the instructors for direct application to their course assignments than the published form of the VALUE rubrics.

The next step in the research was to develop a web-based rubric-builder, which asks the instructor specific leading questions about what the assignment is about and what they want their students to demonstrate. A menu-driven web application was developed as a support mechanism for educators in developing rubrics for student assessments such as analysis and research projects, design projects, investigations or structured inquiries. The rubric builder, Building Assessment Scaffolds for Intellectual Cognitive Skills (BASICS), focuses on the cognitive and intellectual skills of critical thinking, creative thinking and problem solving. The items and language in BASICS were developed through a process of collecting annotations from over 900 work samples. The annotations were grouped by level and high frequency terms were identified. The rubric scaffolds represent the common elements for each of the rubric dimensions. Construction of the rubrics is a five-step process, resulting in a rubric scaffold with action verbs consistent with the levels on the VALUE rubrics that instructors can fine-tune to their own needs.

Over the duration of this project, experts from across the world continued to investigate assessment and the value of college education. The Measuring College Learning (MCL) project synthesized existing literature and key recommendations from six faculty panel white papers. The core principles presented were:

- Faculty should be at the forefront of defining and measuring undergraduate-level learning outcomes
- Students from all backgrounds and institutions should be given a fair opportunity to demonstrate their knowledge and skills
- Any single measure of student learning should be a part of a larger holistic assessment plan
- Institutions should use assessment tools on a voluntary basis
- Measures of student learning should be rigorous and high quality and should yield data that allows for comparisons over time and across institutions (Roksa et al., 2016, pp. 7–11)

The research, analysis and results from the Queen's LOAC project were completely independent of work undertaken in the MCL project, but our empirical methods derived similar findings. The discussion that follows echoes the above principles.

Limitations

This report itemizes numerous limitations of the approach and methodology throughout this paper, and space limitations do not permit a full discussion here. However, in brief, our findings are that standardized, non-embedded tools such as the CLA+ and the CAT have significant limitations, including low student motivation due to lack of discipline-specific content, repetition of the same test over multiple years, failure of instructors to encourage student effort and the fact that these assessments do not count toward the course grade. Other logistical factors include time required to complete the tests, technical difficulties in test administration and delay in receiving the results. The CLA+ and CAT were also significantly more costly. While the VALUE rubric assessments were less problematic in terms of student motivation, logistical challenges and cost, limitations included less comparable results due to differences in assignment type and some misalignment between the assignment and the VALUE criteria.

Another limitation of the longitudinal data is the decrease in the overall number of students tested from first to fourth year, and the fact that the number of matched students tested with specific measures across all years of study is relatively low. While testing all measures in all years provides for robust longitudinal data, test fatigue could have been mitigated and student motivation and effort may have increased by conducting the assessments in first and fourth year only. Studies of this kind are also susceptible to further factors, including the lack of a control group among the general population not engaged in postsecondary education, meaning that it is difficult to factor in developmental growth that is unrelated to purposeful academic instruction in the acquisition of higher-order cognitive skills. Further study is needed on many fronts: the implementation and correlation of yet more tools that measure critical thinking; studies that maintain a control group not influenced by postsecondary education; and studies that correlate growth in higher-order transferable skills and labour-market and life outcomes several years beyond the completion of postsecondary education.

Conclusions

The purpose of the Learning Outcomes Assessment Consortium project (LOAC I) was to investigate sustainable methods of assessing critical thinking, problem solving, written communication and lifelong learning in undergraduate programs and to yield information for course and program improvement. The Collegiate Learning Assessment Plus and the Critical Thinking Assessment Test provided demonstration of learning gains and allowed for inter-institutional comparisons, but were expensive and susceptible to motivational issues. The Valid Assessment of Learning in Undergraduate Education rubrics were found to be the most cost effective and useful method for informing instructors of improvements to teaching and learning. The greatest variability for consistent use of the VALUE rubrics was in the alignment between course and VALUE dimensions. Using a network approach and the VALUE rubrics for purposeful alignment is the focus of the next iteration of the Learning Outcomes Assessment Consortium project (LOAC II).

All of the stakeholders involved in this institution-wide project (administration, heads of departments, instructors, students and researchers) found it valuable to know how well students were developing the skills of critical thinking, problem solving and written communication. Lifelong learning was of interest to some, but was more difficult to quantify and remained a lesser focus of the study. The primary challenge was assessing students in a way that provided data that could be aggregated. Standardized testing was costly and difficult to implement. Participation rates among fourth-year students were much lower than first-year rates, primarily because the first-year testing was conducted in scheduled courses, which was not possible for the fourth-year students. Along with motivational issues, these are two of the disadvantages of standardized tests compared with using VALUE rubrics to score course assignments. Interviews with students suggested that the following are key to encouraging students to put effort into the test: that instructors value the data, that the test content is relevant to students, that test times should not conflict with other commitments, and that achievement should be made available to students. Additionally, if the test score counted toward a grade, then students would put effort in. Failing that, the suggestion was to provide financial incentives.

We found the most successful strategy was to work directly with instructors to help them engage students with assignments that specifically encouraged and elicited critical thinking and problem solving. When comparing methods, feedback from instructors suggested that the VALUE rubrics could be more effective in the short term, but they also recognized that there may be benefits in the CLA+ or CAT for longer-term evaluation and comparison with other institutions. This project provided rich lessons, as researchers navigated the challenges and contextual constraints. To encapsulate its findings in a few bullet points does not do justice to its complexity, but the for the sake of quantifying results we provide them here:

- CLA+ and CAT scores demonstrated significant improvement from first to final year.
- VALUE rubric scores demonstrated large, ongoing increases in critical thinking, problem solving and communication between first-, second-, third- and fourth-year samples.
- Correlations between critical thinking and communication dimensions measured on the CLA+, CAT and VALUE rubrics were low but significant.
- The cost of implementing the VALUE rubric marking was approximately C$20 less per student than implementing the CLA+ or CAT tests. The VALUE rubrics themselves don't cost anything, so if marking was part of an institutional assessment plan, the cost would be negligible.
- The CLA+ results provide an ability to compare institutional performance with other schools. This is not possible with VALUE rubric scoring without common training and calibration procedures.

Acknowledgements

The research team gratefully acknowledge the numerous undergraduate and graduate personnel working on the Learning Outcomes Assessment Consortium project, for their commitment to VALUE rubric marking, analysis and departmental reporting. We'd like to thank and recognize the contribution from research associates Sima Zakini, Launa Gauthier and Nerissa Mulligan, who are not listed as authors here, but whose work on the project was integral to the achievement of project goals. Institutional support was also essential for the successful completion of the project goals, and the team would like to recognize Brenda Ravenscroft for her leadership during the project, in teaching and learning within the Faculty of Arts and Science; Chris Conway and staff at Institutional Research and Planning; and staff at the Queen's Research Ethics office, who discussed and processed 16 amendments to the study over the four years of the project. We would also like to thank all of the course instructors, many of whom gave up time for us to speak with their students and run testing. Finally, we are sincerely grateful for the generous support of the Higher Education Quality Council of Ontario, with particular thanks to Alexandra MacFarlane.

References

Association of American Colleges & Universities (AAC&U). (n.d.). About LEAP. https://www.aacu.org/leap

AAC&U. (2014). Programs | VALUE: Valid Assessment of Learning in Undergraduate Education. http://www.aacu.org/value/index.cfm

Benjamin, R., Klein, S., Steedle, J., Zahner, D., Elliot, S., & Patterson, J. (2012). The Case for Generic Skills and Performance Assessment in the United States and International Setting. http://cae.org/images/uploads/pdf/The_Case_for_Generic_Skills_and_Performance_Assessment.pdf

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

Critical Thinking Assessment Test (CAT). (n.d.). https://www.tntech.edu/cat/

Council for Aid to Education. (2016). *CLA+ National Results, 2015–16*. New York, NY. http://cae.org/images/uploads/pdf/CLA_National_Results_2015-16.pdf

Council for Aid to Education. (n.d.). CLA+ Overview. http://cae.org/participating-institutions/cla-institution-users-portal/cla-overview/

Crick, R. D., Broadfoot, P., & Claxton, G. (2004). Developing an Effective Lifelong Learning Inventory: The ELLI Project. *Assessment in Education: Principles, Policy & Practice*. *11*(3), 247–272. https://doi.org/10.1080/0969594042000304582

Deller, F., Brumwell, S., & MacFarlane, A. (2015). *The Language of Learning Outcomes: Definitions and Assessments*. Toronto, ON: Higher Education Quality Council of Ontario. http://www.heqco.ca/SiteCollectionDocuments/The%20Language%20of%20Learning%20Outcomes-Definitions%20and%20Assessments.pdf

Douglass, J. A., Thomson, G., & Zhao, C. M. (2012). The Learning Outcomes Race: The Value of Self-reported Gains in Large Research Universities. *Higher Education. 64*(3), 317–335. https://doi.org/10.1007/s10734-011-9496-x

Finley, A. P. (2011). How Reliable Are the VALUE Rubrics? *Peer Review*, *13*(4/1), 31.

Ho, D. E., Imai, K., King, G., & Stuart, E. A. (2007). Matching as Nonparametric Preprocessing for Reducing Model Dependence in Parametric Causal Inference. *Political Analysis. 15*(3), 199–236.

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

Kaupp, J., Frank, B., & Chen, A. (2014). *Evaluating Critical Thinking and Problem solving in Large Classes: Model Eliciting Activities for Critical Thinking Development. Toronto, Canada: Higher Education Quality Council of Ontario*. Toronto: Higher Education Quality Council of Ontario. http://www.heqco.ca/SiteCollectionDocuments/Formatted%20Queen%27s_Frank.pdf

Klein, S., Liu, O. L., Sconing, J., Bolus, R., Bridgeman, B., Kugelmass, H., Nemeth, A., Robbins, S., & Steedle, J. (2009). *Test Validity Study (TVS) Report.* The Fund for the Improvement of Postsecondary Education.

Logie, R., Baddeley, A., Mané, A., Donchin, E., & Sheptak, R. (1989). Working Memory in the Acquisition of Complex Cognitive Skills. *Acta Psychologica. 71*(1), 53–87.

Pintrich, P. R., Smith, D. A. F., Garcia, T., & Mckeachie, W. J. (1993). Reliability and Predictive Validity of the Motivated Strategies for Learning Questionnaire (Mslq). *Educational and Psychological Measurement. 53*(3), 801–813. https://doi.org/10.1177/0013164493053003024

Rhodes, T. L. (2011). Emerging Evidence on Using Rubrics. *Peer Review. 13*(4/1), 4–5.

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.

Roksa, J., Arum, R., & Cook, A. (2016). Defining and Assessing Learning in Higher Education. In R. Arum, J. Roksa, & A. Cook (Eds.) *Improving Quality in American Higher Education: Learning Outcomes and Assessments for the 21st Century*. Jossey-Bass.

Simper, N., Kaupp, J., Frank, B., & Scott, J. (2015). Development of the Transferable Learning Orientations Tool: Providing Metacognitive Opportunities and Meaningful Feedback for Students and Instructors. *Assessment & Evaluation in Higher Education*. *41*(8), 1159–1175. http://www.tandfonline.com/eprint/vYJm8CYjrD4I7ThUwDrn/full

Stein, B., & Haynes, A. (2011). Engaging Faculty in the Assessment and Improvement of Students' Critical Thinking Using the Critical Thinking Assessment Test. *Change: The Magazine of Higher Learning*, *43*(2), 44–49.

Stein, B., Haynes, A., & Redding, M. (2006). Project CAT: Assessing Critical Thinking Skills (pp. 290–299). Presented at the Proceedings of the National STEM Assessment Conference. Retrieved from http://www.openwatermedia.com/downloads/STEM(for-posting).pdf#page=294

Sternberg, R. J., & Frensch, P. A. (2014). *Complex Problem Solving: Principles and Mechanisms*. New York: Psychology Press.

Van Merriënboer, J. J. (1997). *Training Complex Cognitive Skills: A Four-component Instructional Design Model for Technical Training*. Englewood Cliffs, NJ: Educational Technology Publications.

Higher Education
Quality Council
of Ontario

An agency of the Government of Ontario