Higher Education Quality Council of Ontario

An agency of the Government of Ontario

Scaling Up: Value-added Cognitive Assessment Redesign Network for the Development of Higher-order Thinking

Natalie Simper, Brian Frank, Jill Scott and Wanda Beyer

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., Beyer, W. (2019). *Scaling Up: Value-added Cognitive Assessment Redesign Network for the Development of Higher-order Thinking.* Toronto: Higher Education Quality Council of Ontario.

Table of Contents

List of Tables

List of Figures

Executive Summary

The Cognitive Assessment Redesign (CAR) project is an institution-wide, network-based study focusing on the development and assessment of cognitive skills (critical thinking, problem solving and to a lesser extent creative thinking) in undergraduate education. The purpose of this research was to examine the relationship between course assessment, standardized rubrics and a standardized test; determine the value-add between first- and final-year cognitive skills achievement; and investigate the effectiveness of a network approach to build capacity of instructors. The project involved incentivised recruitment of first- and fourth-year course instructors interested in purposefully aligning skill development through the design of course assessments tailored to cognitive skill acquisition. Student learning was evaluated using the following measures: course-assessment rubrics, standardized rubrics (Association of American Colleges & Universities VALUE rubrics), and a standardized test (Educational Testing System's HEIghten). Participating instructors completed pre- and post-implementation surveys, and comprehensive narrative reports were compiled to qualitatively track the change process.

The high-level outcomes of the project included the following:

- Analysis of student assessment data found significant correlations between course-based marks and VALUE rubric scores, and data was used to inform next steps for a portion of the courses involved. Low correlations between the HEIghten scores and course-based assessment suggest that the course assessments were evaluating more than critical thinking alone.
- There was a significant improvement (half a standard deviation) in critical thinking skills between first- and final-year cross-sectional student data as evidenced by HEIghten test scores. The result was reflected by gains in two levels across the median level of critical thinking performance on the VALUE rubric, with similar gains on the problem solving rubric.
- The CAR project supported 24 instructors from 15 departments (see Table 1) and allowed them to carefully align assessment of cognitive skills, effectively using quality metrics, where 40% of the instructors used rubrics for the first time in the course.
- Qualitative analysis of the narrative course reports illustrated growth and promising results on many of capacity-building indicators. One potential improvement is expanding adoption of formal processes for consistent scoring of course assignments.
- Data from instructor surveys demonstrated the inherent challenges in implementing culture change in assessment practices but also the degree of capacity building achieved. It is recommended that these metrics be tracked in the future to determine the long-term effects of the project.

Introduction

Effective practices for changing assessment protocols have been shown to enable meaningful improvements in teaching and learning in higher education (Simper, Frank, Scott, & Kaupp, 2018). The Cognitive Assessment Redesign (CAR) project was designed using the findings from a previous LOAC (Learning Outcomes Assessment Consortium) study, (Simper, Frank, Scott & Kaupp, 2018). The Queen's LOAC I project was a four-year longitudinal study that investigated a standardized and program rubric-based assessment for the evaluation of learning outcomes associated with critical thinking, problem solving, written communication and lifelong learning. In the LOAC I project, results of two standardized tests (the Collegiate Learning Assessment Plus and the Critical Thinking Assessment Test) indicated that students' skills in critical thinking, problem solving, written communication and lifelong learning increased over the four years of their degree but these tests were expensive to implement and susceptible to student motivational issues. The standardized VALUE rubrics (Valid Assessment of Learning in Undergraduate Education) proved to be more cost effective and provided a useful method for informing instructors of improvements to teaching and learning, though not without issues related to between-task variance. As a result of findings from the LOAC I project, the CAR-project research focuses on the development and assessment of cognitive skills by implementing the VALUE rubrics and orchestrating a network approach to assessment redevelopment. The CAR project enables aggregation of reliable and valid data (Mathers, Finney, & Hathcoat, 2018) by building a network of instructors from a variety of disciplines and faculties to participate in redesigning courses, a process that included aligning learning outcomes, developing authentic tasks and tailoring assignment-specific rubrics (Kezar, 2013). The project was built upon an engagement model for academic development with embedded support to develop authentic tasks and ensure reliable assessment; aligned with goals for student learning; while being sustained through a community of practice. A brief theoretical foundation for these principles follows.

Assessment Reform

For more than 20 years, higher education governing bodies have worked alongside universities and colleges to develop policies and standards aimed at driving quality assessment practices. In 2008, the Organization for Economic Co-operation and Development (OECD) launched a global investigation, focusing on the feasibility of assessment with a growing recognition of the importance of learning outcomes (Tremblay, Lalancette, & Roseveare, 2012), resulting in the Assessment of Higher Education Learning Outcomes (AHELO) project (Tremblay, 2013).

Participating members from across the world carried out assessment methods with mixed results. In the Bologna Tuning Process in Europe (González & Wagenaar, 2003), over 40 countries worked together to improve the quality of education by identifying learning outcomes and other processes to promote

transparency, mobility and employability. The efforts of "Tuning USA" led to the development of The Degree Qualifications Profile (DQP), a competency framework designed to align curriculum and pedagogy (Jankowski, Hutchings, Ewell, Kinzie, & Kuh, 2013).

More recently, the Association of American Colleges and Universities (AAC&U) coordinated a large-scale Valid Assessment of Learning in Undergraduate Education (VALUE) rubric assessment project in the United States (Drezek McConnell & Rhodes, 2017). There are many reasons why change toward recommendations has been slow, but suffice to say that change in higher education is difficult to achieve (Henderson, Beach, & Finkelstein, 2011).

Engaging Faculty

Boud and Dochy (2010) outline seven propositions for assessment reform, recommending that assessment for learning be "placed at the center of subject and program design, [be] a focus for staff and institutional development" (p. 3). The AAC&U recommends faculty be directly involved in assessing the quality of student learning (Rhodes, 2011). Successful change requires sustained participation, collaboration and support for participating faculty members (Bernstein & Greenhoot, 2014), but a survey by the National Institute of Learning Outcomes Assessment (NILOA) found that college and university deans suggest that "engaging more faculty is the major challenge to advancing assessment" (Kinzie, 2010, p. 14). One strategy for engagement is providing mini grants, where funds can be a stimulus for reform activities and "can have powerful symbolic implications…and individuals leading these efforts have a lever to obtain additional institutional resources" (Miller, Fairweather, Slakey, Smith, & King, 2017, p. 40). Recruiting faculty instructors into the research enables an embedded approach to the collection of learning evidence at the course assignment level.

Embedded Support Network

Faculty engagement is a central tenet in Chasteen and Code's (2018) comprehensive guide for pedagogical change, which highlights the role of embedded experts. Hannah and Lester's (2009) model of organizational leadership suggests a multi-level strategy for change management; strategies implemented to support a knowledge network consisting of building individual capacity (micro), implementing a network that facilitates change (meso) and, to a lesser extent, through systems of institutional sanctioning (macro level). When we consider meso level, academic microcultures can "provide a collegial supportive engagement with new teachers" (Roxa & Martensson, 2015, p. 195). Embedded expertise has been demonstrated to be both successful in large-scale systemic initiatives (Chasteen & Code, 2018; Wieman, 2007) as well as a sustainable system for quality improvement (Wieman, Deslauriers, & Gilley, 2013). The alignment of course assignments with desired assessment criteria often requires the development of tasks that take time and expertise to develop and further requires a high level of commitment from those who teach the students (Purmton &

Alexander, 2013). Research suggests that with the infusion of dedicated time and expertise, the necessary critical mass can be achieved to create a new norm (Corbo, Reinholz, Dancy, Deetz, & Finkelstein, 2014).

Authentic Tasks, Meaningful Learning

The principles of constructive alignment suggest that course assignments need to be carefully designed to align with the intended outcomes (Biggs, 2014): "Just as we understand that what is taught is not the same as what is learned, we also know that the goal is to ensure that students have gained knowledge and skills particular to an individual course" (Chun, 2010, p. 23). Further, as paraphrased from Wiggins and McTighe (2005), goals for long-term learning should include experiences for meaning-making and enable students to transfer skills and abilities from one setting to another. Authentic tasks "present students with a complex, real-world challenge in which the scenario, role, process and product are all authentic; they must then demonstrate that they have the skills and knowledge to complete the task" (Chun, 2010, p. 24). In other words, our aim is for students to meaningfully apply concepts to solve problems, use evidence to recommend solutions, or creatively develop new approaches or methods.

The nature of the problem and the disciplinary field, however, need to be considered. As Jonassen, (1997) argues, authentic assessment design choices will depend on a well-structured problem.

Specifically, authentic tasks may resemble analysis and research projects, design projects, investigations or structured inquiries (Ashford-Rowe, Herrington & Brown, 2013). Differences in such assessment types in different disciplinary fields can have implications for consistency in evaluating student learning.

Reliability

Course-based assessments are generally considered to hold face validity (Tremblay, 2013), but differences in assignment type when not properly aligned can account for up to 77% of error variance (Hathcoat & Penn, 2012), impacting the reliability of the results. Our prior LOAC I research (Frank, Simper, & Kaupp, 2016) found misalignment between the course assignment and the assessment criteria stemming from constructing the assessment without a consistent format for operationalizing the constructs of critical thinking or problem solving. If students are not guided to provide responses that demonstrate critical thinking and problem solving, then the validity of assessing those constructs is put into question. It is generally considered that the greater attention given to teaching and assessing these skills, the better the students are able to perform. The AAC&U report states that a "curricular focus on developing critical thinking skills in students through their major programs, which faculty claim is a priority, is reflected in the higher levels of performance among students in upper division course work in the majors" (Drezek McConnell & Rhodes, 2017, p. 4). Reliability in assessment is necessary to estimate learning gains through a program.

Many authors have articulated principles of effective assessment design (for example, Popham, 1999; Wiggins & McTighe, 2005), but without a deliberate assessment plan, broad statements of outcomes are difficult to substantiate. An assessment plan needs to take many factors into account; for example, each learning context is different, which poses difficulties for a unified approach. Criterion-based rubrics are being increasingly used to evaluate student learning (Dawson, 2017). There are, however, flaws to the design and use of rubrics, such as rubrics that comprise checklists rather than quality indicators (Popham, 1997), or questionable processes for establishing inter-rater reliability (Jonsson & Svingby, 2007). A meta-analysis conducted by Jonsson & Svingby (2007) suggests that "reliable scoring of performance assessments can be enhanced by the use of rubrics. In relation to reliability issues, rubrics should be analytic, topic-specific, and complemented with exemplars and/or rater training" (p. 141). Rater training ensures that raters understand how to implement the rubric. The process includes a discussion of the terminology, practice scoring using work samples and an opportunity for participants to explain their rationale for assigning a given mark.

Some argue for a standardized testing approach to assessment rather than a rubric-based approach to quantify skill development (Council for Aid to Education, 2016; Stein & Haynes, 2011). However, these tests are still viewed by many students and instructors as disconnected from the disciplinary context, administered as low-stakes assessment, and not viewed as related to domain knowledge, and are therefore considered to be of questionable benefit (Madaus & Clarke, 2001). Standardized tests that have little or no consequence for the test-taker are susceptible to motivational and logistical issues (Liu, Bridgeman & Adler, 2012; Simper, Frank, Kaupp, Mulligan & Scott, 2018). If students are not motivated to perform, the results will not accurately capture their learning (Banta & Palomba, 2014). On the other hand, assessments that are considered high-stakes by either students (for admissions) or the institutions (for funding) can deteriorate the learning environment. This deterioration occurs because of instructional focus shifting from competencies that are not tested, or excluding low-scoring students from testing thereby inflating student performance (Koretz, 2008).

Communities of Practice

Wenger (2000, 2011) describes communities of practice as those that involve sustained integration between self-selected members who share common interests, for the purpose of sharing knowledge and collective problem solving. Mutual trust exists in such groups, and many consider them "the ideal social structure for 'stewarding' knowledge" (Wenger, McDermott & Snyder, 2002, p. 12). Hutchings, Huber and Ciccone, (2006) suggest a similar strategy for collective improvement of teaching in higher education is the teaching commons, "a conceptual space in which communities of educators committed to pedagogical inquiry and innovation come together to exchange ideas about teaching and learning and use them to meet the challenges of preparing students for personal, professional, and civic life" (p. 26).

Research Questions

The design of the Cognitive Assessment Redesign (CAR) project leveraged effective practices as outlined above, built upon embedded support, focused on engaging faculty to develop meaningful learning assessments for students, and evaluated through multiple methods. The CAR project also aligned directly with Queen's University Academic Plan, building processes and capacity for assessing learning outcomes. The specific research questions driving the processes and analytical methods were:

1. How effective is the cognitive redesign model for supporting instructors in the development of course-based cognitive skills assessments?
2. What is the relationship between the Valid Assessment of Learning in Undergraduate Education (VALUE) rubric scores, course assessments and HEIghten outcomes?
3. What is the value-add between first- and final-year cognitive skills achievement across Queen's University?
4. What are the indicators that the project is building instructor capacities?
5. How effective was the project in capacity-building and propagation of the initiative?

Methods

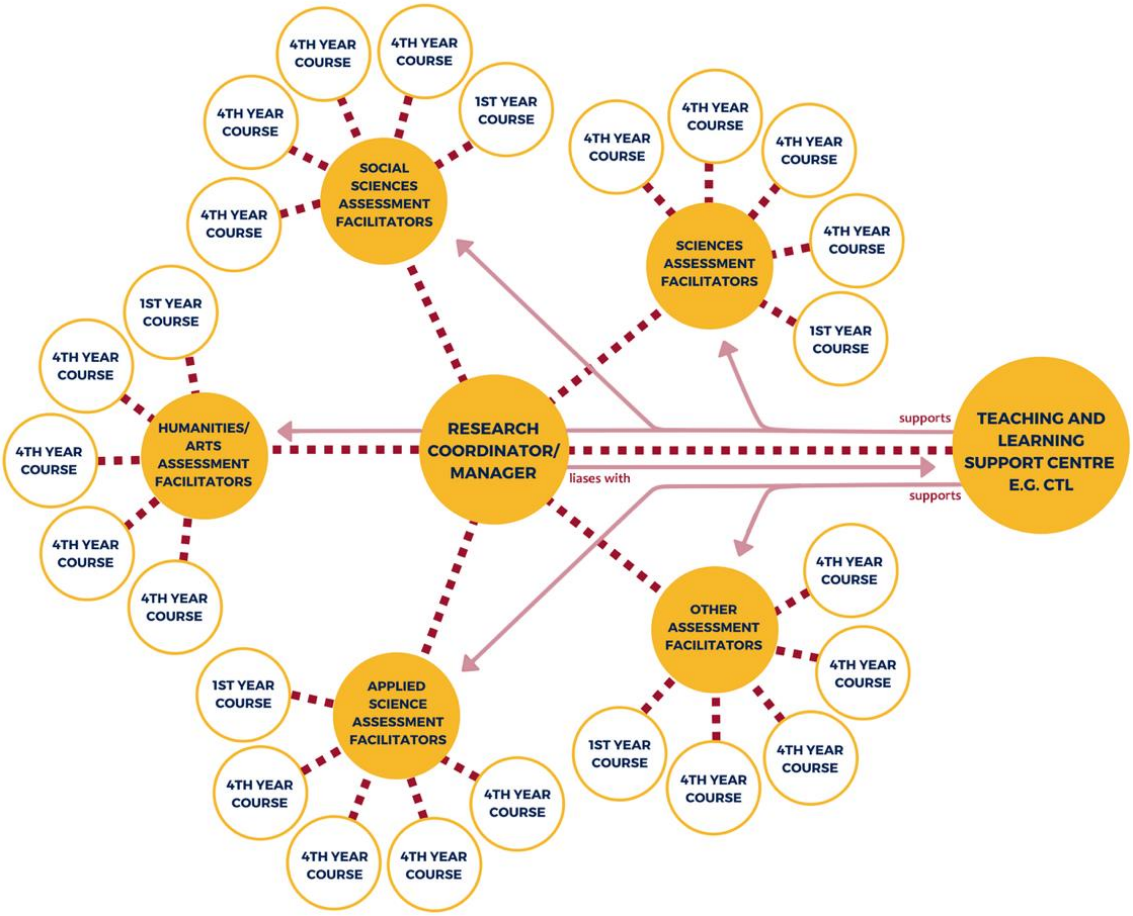
The CAR project was designed to build capacity in constructing assessments that aligned with dimensions of critical thinking, problem solving or creative thinking. This project was implemented in four phases:

- **Pre-implementation**: Open conversations and consultation conducted with educational professionals in departments. (January–April 2017)
- **Project Initiation**: A website was constructed to provide overview of the project and call for proposals. (June–July 2017)
- **Launching the Project**: The creation of the network started to form including promoting the project to department heads, orientating assessment facilitators and holding initial consultations with instructors. (August–September 2017)
- **Implementation**: Assessment facilitators collaborated with participating faculty, first- and fourth-year students recruited, standardized test administered, VALUE marking and data analysis conducted. (September 2017–July 2018)

The network structure (see Figure 1) is anchored by the research manager, who worked closely with stakeholders, sensitive to the needs of senior leadership, while respecting the disciplinary norms and practices. The research manager anchored the support personnel (assessment facilitators), who held

disciplinary expertise, to provide assessment support to the participating instructors. The five assessment facilitators, employed one day per week, assisted faculty members with developing assessments that aligned with overarching metrics. During their weekly meetings, they received training in developing authentic tasks and assessment processes and procedures. Three of the assessment facilitators were working concurrently as research associates, one as an instructional designer and one was a teaching adjunct. In their existing roles, they had built trust and were respected in their disciplines. Further, they had intellectual and human capital with which to navigate and help recruit faculty instructors (Hanson, 2001). Each assessment facilitator worked with one large first-year, and multiple fourth- (or final-year) course instructors, in matters of teaching and assessment.

Figure 1: Network Structure to Support and Sustain Cognitive Skill Assessment



Participants

Instructors

The LOAC principal investigators liaised with associate deans and department heads to engage their support in recruiting instructors for the project. A web page was also designed that described the project and advertised calls for expression of interest. Instructor participation was formalized through the submission of a short expression of interest (EOI). The three reflective question prompts in the EOI were (1) How do you currently encourage the development of cognitive skills in your course? (2) How do you currently assess cognitive skills? (3) How might you improve those assessments? There were 24 participating instructors from five disciplinary groupings, with representation from departments of biomedical science, biology, drama, engineering (chemical, civil, geological and multi-disciplinary), English, geography, health science, medicine, nursing, psychology and sociology. Each participating instructor received a small monetary stipend to use for course enhancements such as guest speakers, course specific research or student presentation sessions.

Students

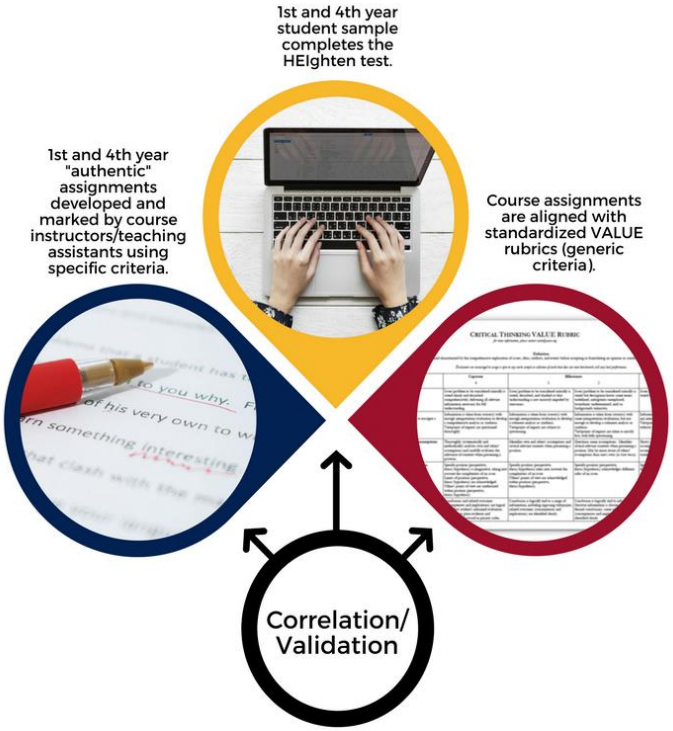
Students enrolled in participating first- and fourth-year courses were recruited during class time by the designated assessment facilitator or research manager. In some cases (e.g., online courses), students were recruited through the learning management system by the assessment facilitator or research manager. Recruiting students served two purposes, firstly for consent to participate to meet ethical guidelines, and secondly so that they were aware of the efforts of their course instructor and Queen's University to improve and assess cognitive skills. A cross-sectional design was employed. However, the sample size varied from course to course. Some of the upper-year courses had low enrolment (n=12), so the maximum consenting sample was limited by the number of enrolled students. In addition, there were students in a range of program years enrolled in multiple participating courses; as a result the sample was selected by their year in the program (first or final year). This overlap of participants limited the available number of student responses. Where possible, the sample was drawn from those who had completed the HEIghten test. For large enrolment courses, with a majority of students consenting, a process of stratifying was employed to ensure representativeness. Consenting students were selected to represent those who scored high, middle and low on the HEIghten test, included a gender balance in accordance with the breakdown in the program. English as an additional language students were included wherever possible to ensure adequate representation on the HEIghten test since part of the test includes written responses. The course assignment title and sample size are listed in Table 1.

Measures for Assessing Student Learning

The three assessment tools used to evaluate student learning are shown in Figure 2. They are:

- Standardized test (HEIghten Educational Testing System)
- Course-based assessment (rubrics and marking keys)
- Standardized rubrics (AAC&U VALUE rubrics)

Figure 2: Student Learning Measures



Standardized Test

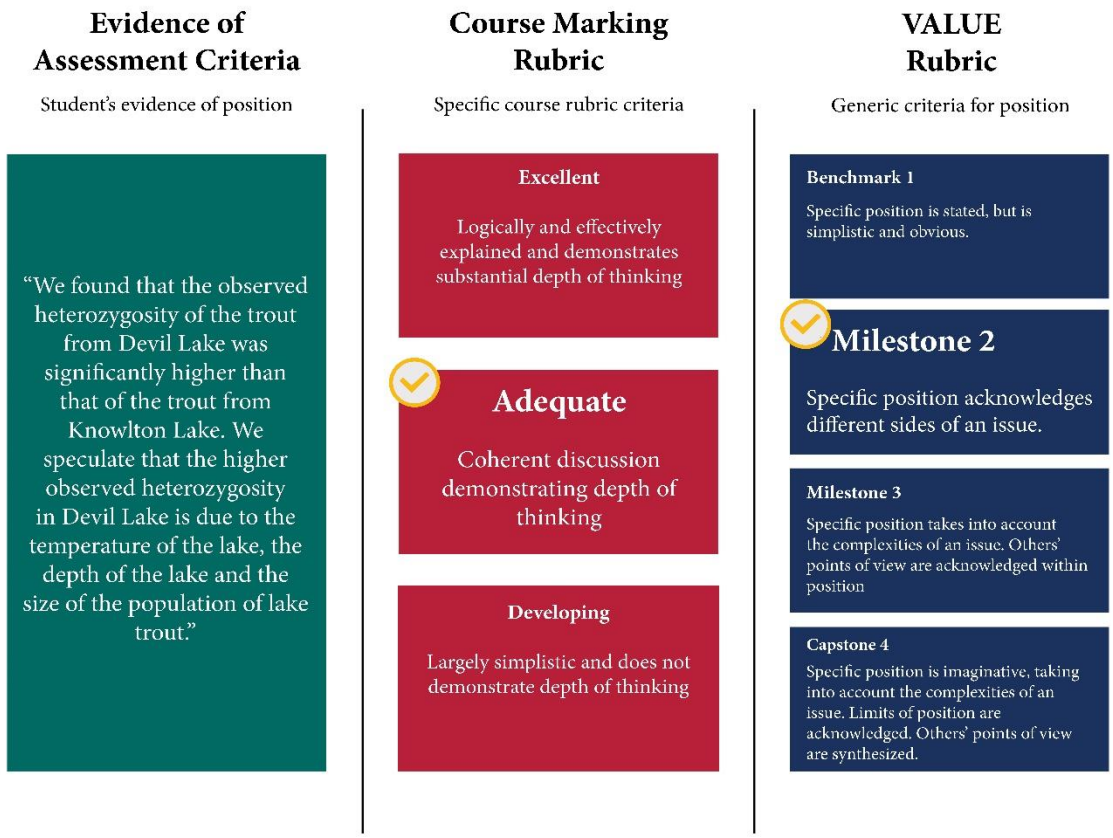
Members of the research team investigated and implemented the Collegiate Learning Assessment Plus, the Critical Thinking Assessment Test and HEIghten as part of a previous LOAC study. They found that the HEIghten was the most useful, cost effective and feasible to meet the goals of the CAR project (Simper, Frank, Kaupp, Mulligan & Scott, 2018). Additionally, the HEIghten test triangulated the data as this standardized assessment captured different aspects of critical thinking than the VALUE rubrics. The HEIghten test was developed by Educational Testing Service (ETS). It is a 26-item online critical thinking test that includes an exit survey with questions about demographics, engagement and effort. HEIghten is designed to evaluate analytical skills (analyzing argument structure; evaluating argument structure; and evaluating

evidence and its use), and synthetic skills (developing valid or sound arguments; selecting information that would constitute or contribute to such arguments for a given position; drawing or recognizing conclusions, extrapolating implications, or recognizing or generating explanations for phenomena that are described). It is administered in a one-hour test session, either proctored in-person through the test portal or un-proctored through email invitation. The results are available immediately; the test has an option to display the student's result at the conclusion of the test. It displays their score for the assessment constructs, as they compare to the international average. The results are also available by course report download or data file (.csv) as soon as the test session is closed. HEIghten testing was embedded in 10 courses and an invitation went out to all the participating students in the fourth-year courses that did not embed testing. Incentives for participating in the test ranged from no incentive, a course requirement, prize draw or a percentage of the course grade attributed to professionalism in attending the test (see Table 3 for descriptive statistics).

Course-based Assessment

The assessment facilitator and course instructor worked through an iterative process to design or redesign a course assignment. Student learning was evaluated using a type of quality framework, mostly in the form of a rubric. However, there were a few cases where the course-based scoring was undertaken holistically using grade descriptors. The objective of this work was to align the course assignment with an assessment rubric that aligns with the outcomes from the selected VALUE rubric (see Figure 3). The VALUE rubrics were used to operationalize the constructs of critical thinking, problem solving and creative thinking. The language in the course VALUE rubrics informed the development of the course rubrics, and provided a discussion point for the assessment facilitator with the instructor. Once the course-based assessment and rubric was developed, instructors reviewed the assignment and rubric with the students similarly to the way they would for any other course assignment. Figure 3 displays a representation of the alignment between the course and VALUE rubrics, in relation to the evidence exhibited in a student artefact.

Figure 3: Course and VALUE Rubric Alignment



Valid Assessment of Learning in Undergraduate Education (VALUE) rubrics

There are 16 overarching VALUE rubrics developed by the AAC&U (Rhodes & Finley, 2013). The front page of each rubric contains descriptions and definitions of terms. The criteria are on the back, with the left-hand column listing the dimensions, and the top row showing the achievement levels (Capstone 4, Milestones 2 and 3 and Benchmark 1). There are quality indicators for each level and dimension. The intended outcomes from these three rubrics are summarized in Figure 4.

Figure 4: Cognitive Skill Outcomes Operationalized Using the VALUE Rubric Criteria

Cognitive skills are expressed differently depending on the norms, practices, contexts and common assumptions within a discipline. These differences in opinion about assessment constructs can become a barrier to the work. For example, much time could be spent debating how critical thinking in the humanities is different to critical thinking in the sciences. However, such debate diminishes the goal of assessing the construct. Therefore assuming a pre-defined framework like the VALUE rubric is beneficial for operationalizing cognitive skills.

Collecting and Scoring Student Responses

Where possible, the consenting students' assignments were collected through the learning management system, but for courses where assignments were submitted in hard copy, we collected and scanned the documents. Each student response in the sample was independently scored by two people. One of these was the assessment facilitator involved in the development of the course assignment, the other a disciplinary expert. A consistent set of procedures was followed for scoring using the VALUE rubrics. The procedure involved building a common understanding by reading through the assignment instructions and an example response and identifying how the dimension criteria has been demonstrated and coming to agreement about the level of the work. The scorers then individually rated five to 10 work samples at a time, compiling an annotated list to support the decision for each of the criteria, and assigned a performance level for each dimension. The final step is called calibration, where the two scorers use their annotations to discuss any differences between levels assigned and, if possible, come to agreement on the final level. The pre- and post-calibration levels were recorded to track scoring reliability. Queen's University has been conducting VALUE rubric scoring for the past five years As such, they have access to experienced scorers and have established processes for inducting new scorers. The pre-calibration reliabilities between the two markers were high (between 68%– 78%), and a process of calibration ensured that the final scores had an agreement rate of 99%.

Table 1: VALUE rubric sample

Faculty	Course	Assessment redesign	Sample
Social Sciences	PSYC 100	Three-phase critical thinking lab	41
	PSYC 453	Critical response and research proposal	8
	PSYC 450	Multi-part critical thinking response	16
	GPHY 401	Food Systems Analysis Project	10
	SOCY 424	Power, inequalities and social justice- poster/presentations	15
Humanities	ENGL 100	Final essay: Analyzing a poem	50
	ENGL442	Reflection and critical essay	10
	ENGL 487	Close reading response	24
	ENGL 489	Using discussion to inform secondary sources paper	26
	DRAM 439	Performance process reflection and peer feedback	9
Applied Sciences	APSC 100	Mars Colony feasibility report	191
	MECH 495	Workstation design	56
	CIVL 500	Research thesis and poster session	16
	GEOE 447	Research report	10
	CHEE 470	Environmental hazards report	64
Sciences	BIOL 103	Critical thinking lab and research report	42
	BIOL 402	Research Design and Lab report	3
Health Sciences	NURS 101	Critical reflection: Indigenous health	42
	NURS 401	Critical reflection: health related issues	36
	HLTH 102	Evaluation of sources of evidence	44
	ANAT 100	Case study: Applying anatomical knowledge	14
	BMED 173	Series of critical thinking blogs (Animal experimentation)	24
	MEDS 112	WIKI Med edit and referencing project	17
	OT 871	Novice-expert patient consultation reflections	20
	TOTAL		788

Results

VALUE Results

The first- and fourth-year median scores on the critical thinking dimensions are displayed in Table 2. As indicated, the first-year median score for all dimensions was Benchmark 1 (n= 391), the final-year median scores for "explanation of issues" was Capstone 4 and the other dimensions were Milestone 3 (n= 268). The percentage of students scored at each level of the critical-thinking dimensions is displayed in Figure 5. Course-level outcomes are displayed as bar graphs, attached in Appendix 1.

The first-year median scores for the problem solving dimensions of "define problem," "Identify strategies" and "propose solution" were Milestone 2. The first-year median score for "evaluate solution" was Benchmark 1, "evaluate outcomes" was lower than Benchmark 1 (notated with a zero level) (n= 203). The final-year median scores for all dimensions was Milestone 3 (not all of the assignments were scored on every dimension n=168 to n=178). The percentage of students scored at each level of the problem-solving dimensions is displayed in Figure 6.

Table 2: Sample Sizes for VALUE Outcomes

		First year	Median	Final year	Median
Critical thinking	Exploration of Issues	391	1	268	4
	Use of Evidence	357	1	268	3
	Context and Assumptions	391	1	268	3
	Position/Hypothesis	391	1	212	3
	Conclusions and Outcomes	391	1	268	3
Problem solving	Define problem	203	2	178	3
	Identify strategies	203	2	174	3
	Propose solutions	203	2	177	3
	Evaluate solution	203	1	168	3
	Evaluate outcomes	191	0	169	3

Figure 5: Critical Thinking VALUE Outcomes

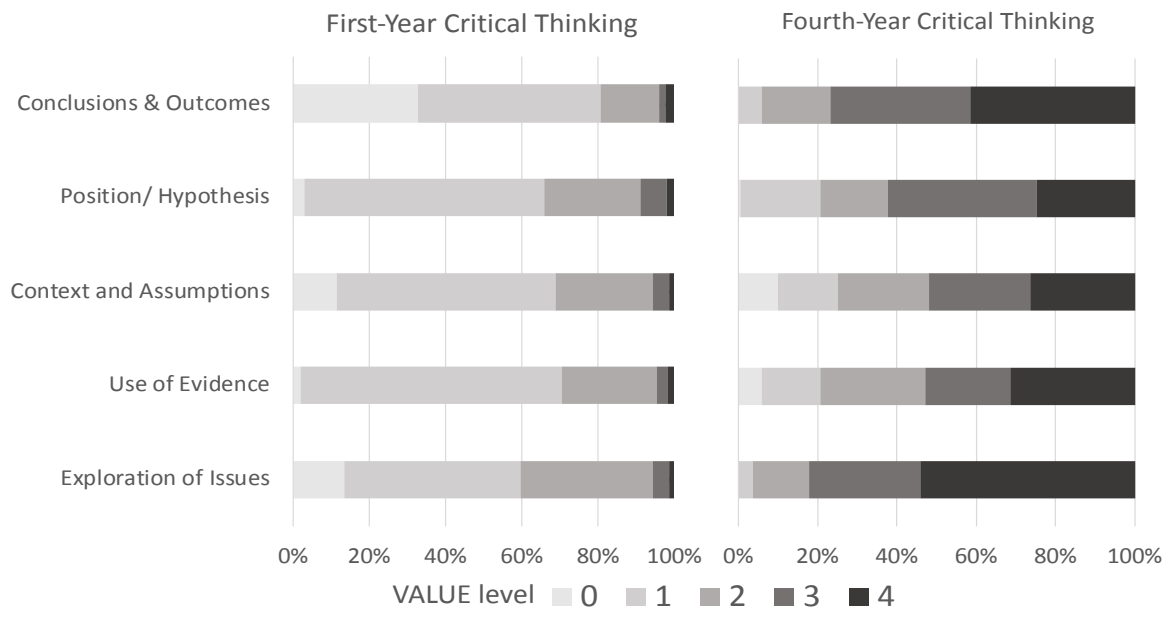
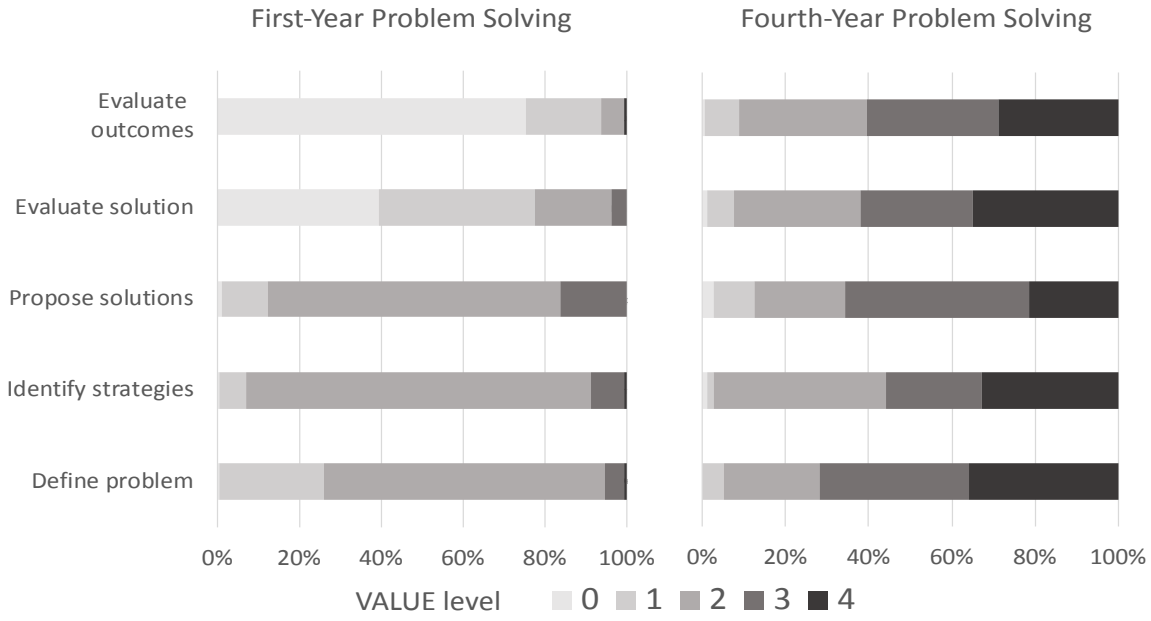


Figure 6: Problem Solving VALUE Outcomes



Note: There was an insufficient sample of assignments evaluated on "implement solution" to be representative across the institution, so they were not included in Figure 6.

HEIghten Results

The descriptive statistics for the HEIghten test sample are listed in Table 3. Please note that the individual course results have been anonymized for participating instructors. The greatest percentage of participation and test completion was observed through on-site proctored testing, conducted as a course requirement, or with a percentage of the course grade attributed to professionalism in attending the test. The HEIghten test result is not considered valid if the students completed less than 75% of the test. Only completed test data was included for analysis.

The HEIghten proctor portal allows reports to be run, providing comparative scores and achievement cut-points. An example of a HEIghten report that was run for all students who tested in first-year courses is included as Appendix 3. The test system also allows data to be downloaded immediately following test completion. Separate analysis was conducted based on the samples of students in their first and final year of their programs, who completed at least 75% of the test, and reported that they put at least some effort into the test. Analysis of variance found significant difference between first-year (n=1523) and final-year (n=149) performance on the HEIghten test $F(1,1683)=30.82$ $p<.001$ (see Figure 7). The Cohen's effect size was $d=.497$ (half a standard deviation improvement from first to final year). Program level breakdowns are displayed as box plots, attached in Appendix 2.

Student effort was self-reported on a 5-point scale. Effort was a significant factor for test score outcomes, investigation of linear regression found the first-year effort accounted for 8% of total score variance ($R2=.08$). There was a similar result from final-year students, with effort accounting for 7% of total score variance ($R2=.07$). Figure 8 displays the mean score for each of the effort scale levels. Notably, even though there was a significant difference between first- and final-year performance, first-year students who put their best effort in, outscored those in final-year who put little effort in the test.

Table 3: HEIghten Descriptive Statistics

Course	Test context and incentive	Total test sample	Complete test	Total consent	% consent and complete	Gender			> 75% complete	Total score	
						% Female	% Male	% Other		Mean	SD
# 1	Un-proctored-0.5%	289	174	73	18%	69%	28%	3%	72%	165.4	7.21
# 2	Proctored-2%	915	882	709	74%	29%	70%	1%	95%	168	5.74
# 3	Un-proctored-5%	934	898	791	82%	74%	25%	1%	96%	164.4	7.19
# 15	Un-proctored-course requirement	87	71	83	78%	90%	9%	1%	83%	159.4	6.58
# 18	Proctored-course requirement	136	134	124	90%	94%	6%	<1%	98%	165.4	6.4
# 21	Proctored-5%	58	56	57	95%	91%	9%	<1%	96%	162.4	6.92
First-year TOTAL		2419	2215	1837	76%	56%	41%	4%	90%	164.2	6.67
# 7	Proctored-5% and food	18	17	17	94%	44%	51%	5%	94%	172.9	3.59
# 16	Proctored-food	56	54	45	80%	47%	53%	0%	98%	170.1	5.92
# 19	Proctored-course requirement	78	77	58	74%	90%	9%	1%	98%	166.9	5.36
4th Yr. Grouped	Un-proctored-prize draw	38	30	38	79%	58%	42%	1%	96%	165.9	4.06
Fourth-year TOTAL		190	178	158	94%	66%	33%	1%	93%	169.0	5.22
# 20 (Graduate course)	Proctored-course requirement	70	59	52	74%	90%	6%	4%	79%	168.1	6.11

Figure 7: Improvement in HEIghten Performance between First and Final Year

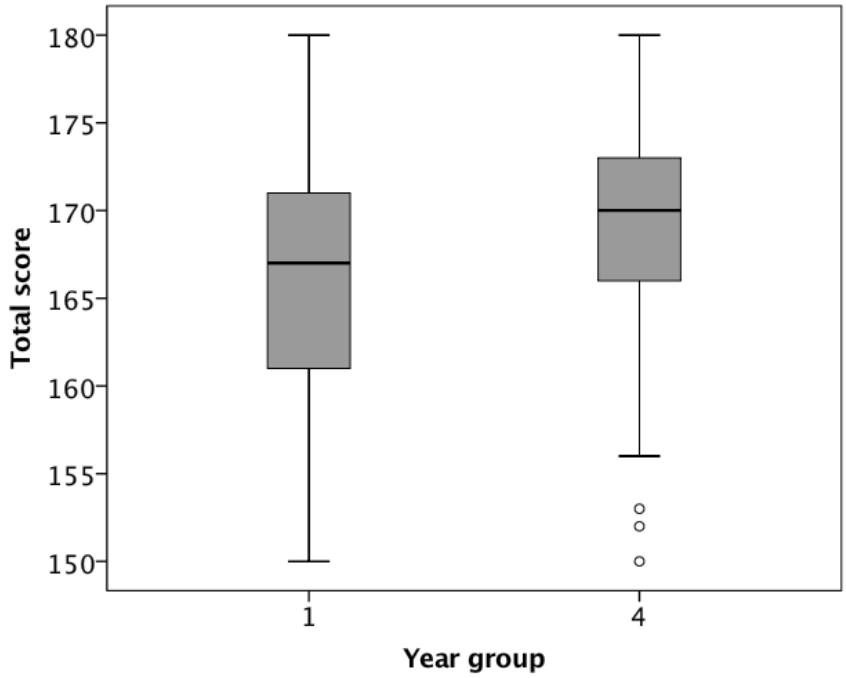
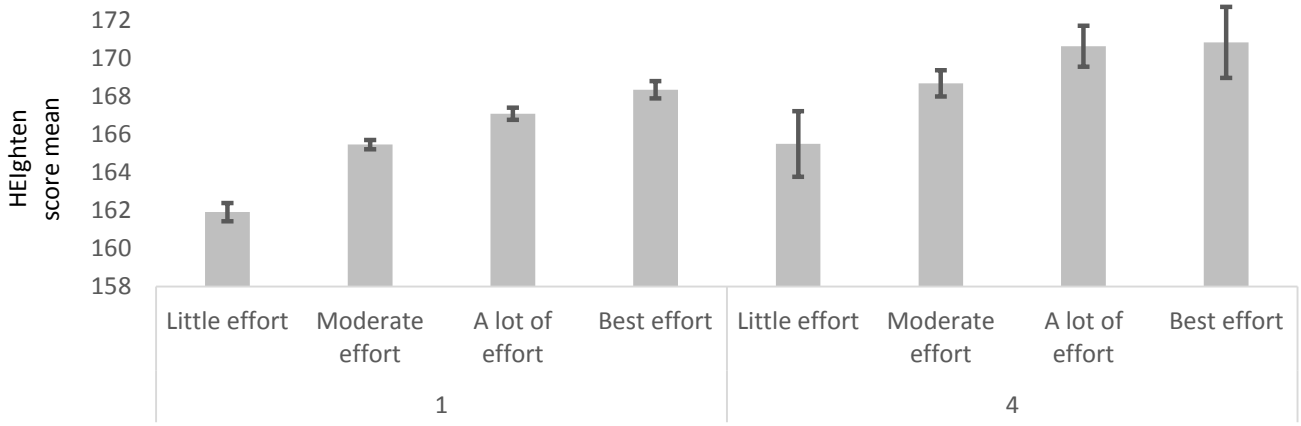


Figure 8: HEIghten Score by Effort Level per Year Group (displaying mean error)



Correlations between Measures

Given that the VALUE data was normally distributed with no outliners and there was a mix of ordinal and continuous data, Pearson's correlations were calculated between the VALUE rubric score per dimension, course assessment, course percentage, sessional Grade Point Average (GPA) and Cumulative GPA, as well as the HEIghten assessment (see Table 4). There were significant correlations between all of the VALUE dimension scores and their respective assignment marks. The highest correlation at the course level was between the problem-solving dimension of "implement solution" and the assignment grade ($r(70)$= .880 p<.001) and the lowest was the critical-thinking dimension of "conclusions and outcomes" ($r(340)$=.221 p<.001). As the assignment mark is a composite of the assessment dimensions, correlations were calculated based on a VALUE average of critical thinking (CT), and an average for problem solving (PS). The correlations between the course assignment mark and CT were $r(400)$= .303 p<.001, and between the course assignment mark and PS were $r(355)$= .319 p<.001. Correlations to course percentage were higher than the assignment grade, with CT they were $r(518)$= .355 p<.001, and with PS they were $r(248)$= .603 p<.001. Correlations to the sessional GPA dropped to the point that correlation to the cumulative GPA was not significant for CT. The correlations between the VALUE dimensions and HEIghten scores were similar to VALUE course assignment correlations. The strongest correlations to HEIghten were to the cumulative GPA (r=(2011)= .318 p<.001. The dimension level correlation table is attached as Appendix 4.

Table 4: Correlations between Assessment Measures

	Assignment grade	Course percentage	Sessional GPA	Cumulative GPA	VALUE CT Average	VALUE PS Average
Course percentage	.236**					
	358					
Sessional GPA	.130**	.633**				
	457	1437				
Cumulative GPA	.108*	.534**	.932**			
	457	1437	2606			
VALUE CT Average	.303**	.355**	.140**	0.002		
	400	518	740	740		
VALUE PS Average	.319**	.603**	.425**	.253**	.899**	
	355	284	394	394	337	
HEIghten total score	0.086	.198**	.279**	.318**	.193**	.241**
	302	1045	2011	2011	491	274

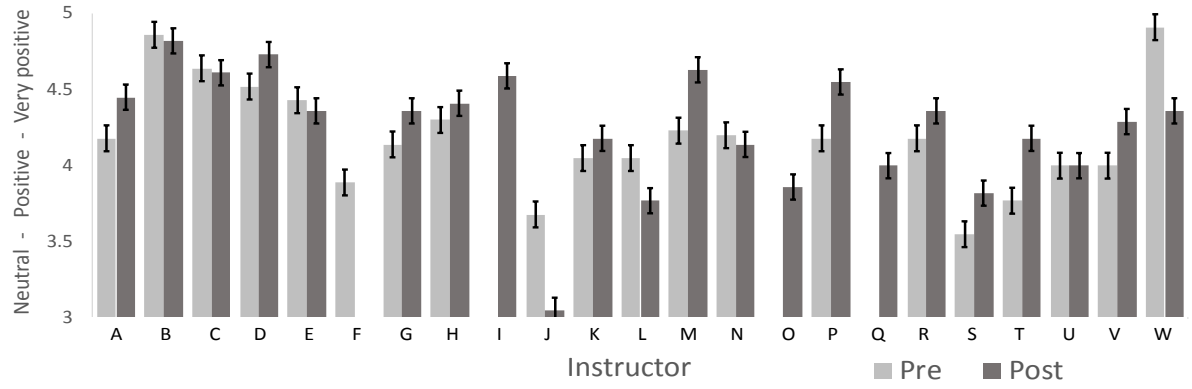
** Correlation is significant at the 0.01 level (2-tailed).

* Correlation is significant at the 0.05 level (2-tailed).

Instructor Survey

Prior to the commencement of the assessment redesign, the participating instructors completed a survey of teaching attitudes, and teaching and assessment practices. The survey included questions related to the specific assessment intervention and a section related to teaching attitudes (15 items reported on a 5-point agreement scale, and five items reported on a 5-point importance scale). The scale reliability for these 18 items was calculated using Cronbach's alpha = .86. An example of the agreement scale questions is "To teach effectively requires knowing how students learn a subject and not just knowing the subject." An example of the importance scale questions is "Understanding how students learn a particular subject." There were 23 instructors who completed the survey, 19 of whom completed both pre- and post-surveys. The scale scores for pre- and post-survey for each of the course instructors are displayed in Figure 9. The survey data suggests that of the 19 paired responses, there was a significant positive change for eight instructors, no significant change for eight, and a negative change for three.

Figure 9: Teaching Survey Quantitative Results



Course Narratives

The assessment redesign efforts were reported by the assessment facilitators using a narrative report template. The questions in the template were developed by the research group to meet the research goals and to track process, problems and qualitative outcomes. There were 25 narrative reports (one for each course), approximately five pages each. The notable key features from the narratives were summarized into a seven-page table with activities grouped in the descriptive column, and outcomes in another column (the table is attached as Appendix 5). An inductive analysis process followed reflecting on common threads, generalizing from behaviours and outcomes that indicated professional growth.

The next step was to group behaviours that were common to the majority of cases, and included as growth indicators. The frequencies of those actions from the participating course instructors are indicated in Table 5. Examples of the "discovery" indicator were things like making a substantial shift in perspective and teaching style due to conversations with the students and assessment facilitator, or the course instructor realizing the incidental learning built into the critical-thinking activity had the effect of developing persistence and professionalism skills. These discoveries were not always positive; for example, when working through a rigorous process for critical-thinking assessment, realizing that the content focus did not align with the criteria on the critical-thinking dimensions of the VALUE rubric.

Table 5: Capacity-building: Growth Indicators

	Number of instructors	% of participating instructors
Used a criterion-based rubric for the first time	10	40%
Discovery/revelation pedagogy/cognitive assessment in the discipline	5	20%
Engaged formal process for consistent (TA) scoring of course assignments	6	24%
Employed strategy promoting student effort in standardized assessment	8	32%
Instructor presented at a Conference/Symposia	6	24%
Instructor is sharing their expertise in a formal way (assessment coach)	3	12%
Work is ongoing (instructor/course participating again)	12	48%

Discussion

Model to Support Instructors in Development of Course-based Assessments

A network approach provides an opportunity to negotiate and interpret meaning to learning, teaching and assessment practices (Roxå, Mårtensson & Alveteg, 2011). The CAR enabled a support network to engage faculty as well as promote change in assessment practices. Working at the "meso level" to foster this change, assessment facilitators successfully supported 24 instructors in redesigning their assessments to specifically target critical thinking, problem solving, and in one case creative thinking. These assignments were diverse (see Table 1), and required an iterative process of planning, design and refinement. For some of the participating instructors, this process was a steep learning curve; one instructor commented that "we

tried a radically different format and had some implementation challenges" (Instructor Q). Another suggested the "a little crash course on the vocabulary of pedagogy would be great — perhaps one additional 90-minute meeting where we could learn the meanings of terms." (Instructor H). In addition to assessment development, assessment facilitators helped instructors articulate the specifics of the VALUE rubric language; for example, "to break down those broad cognitive skills into component parts and then try to notice how/when students are using them" (Instructor J). This critical dialogue unpacked the discipline-specific nuances of the VALUE rubric terminology. This level of support at the "micro" level helped instructors constructively align their course assignments to the VALUE rubrics; for example, they need to "become familiar with the VALUE rubric in advance of designing the assessment to get a sense of how cognitive skills can be evaluated, then work backwards when designing the assessment 'backward design'" (Instructor O).

Given that change in teaching and learning requires sustained participation, collaboration and support (Bernstein & Greenhoot, 2014), some instructors were not able to fully implement the assignment that they modified for the 2018 iteration of the course. For online courses, the assignments are fixed in advance as part of the online course quality assurance development process meaning few revisions are allowed after the start of term. Some instructors reported that the course assignment took a considerable amount of time to mark, especially with respect to feedback. Taking a continuous improvement approach has allowed a number of instructors to critically reflection on their involvement and adapt their assignments to address concerns. Feedback from the students about the assessment process was sought in some courses, which proved to be an important source of evidence to use for course improvement. One instructor stated:

> I will be recommending that we implement the students' recommendations! This includes providing them with examples and practice working through the process of conducting an independent research project before expecting them to be able to do it on their own…the biggest thing is that we need to define what we even mean by critical thinking to our students, and give them practice working through the critical-thinking process (with much guidance) before we assume they are going to be able to do it on their own. (Instructor M)

The course instructors also had excellent advice for others undertaking similar work, examples of the advice include:

> Think first about what you intend for the students to be able to demonstrate at the end of the course, and then build strategies for teaching those cognitive skills. Don't assume the way you've done things in the past is accurately assessing what you intend to assess in your course. (Instructor K)

> I would say to try to gain before and after snapshots of skills, to give students plenty of feedback, and to listen to students' own language and reflections and goals, to have a thoughtful rubric, and to ask for advice! (Instructor H)
>
> Start by brain storming (with trusted colleague, education developer, etc.) what the overall critical thinking goals are for the course, and how they can be met with specific skills development (again, which skills, at what level, in what order). And some of this ideally should require attention to where the course fits in the student's program of study (where this is relevant). And encourage looking at the educational literature on critical thinking in their particular discipline as well as find out what colleagues are doing. (Instructor D)
>
> Clarify which cognitive skill they would like their students to further develop and create/adopt a frame of thinking for that particular skill to help guide rubrics, purpose of the assignment, instructions to students, marking and feedback. (Instructor A)

These comments are indicative of course instructors understanding the process for assessing cognitive skills and the benefits of constructively aligning learning goals.

Relationship of Course Outcomes with VALUE and HEIghten Results

As previously stated, the definitions of cognitive skills are diverse, and each are likely influenced by disciplinary contexts as they are expressed differently depending on the norms, practices, contexts and common assumptions within a field. A concerted effort was made in the project to recognize these differences, in support of face validity, while still using a consistent framework for assessment and rubric design. During multiple weekly assessment facilitator meetings, the project manager and facilitators applied the VALUE rubrics to assessment samples from each of the various disciplines involved. These working meetings ensured understanding and application of cognitive skill constructs as related to their subject areas. Assessment facilitators liaised with instructors to redesign course assessments and co-create course rubrics to ensure the performance descriptors represented the disciplinary context and aligned with the VALUE rubrics. In most cases, this work was an iterative process that required ongoing revisions.

The correlations between measures was undertaken as a cross-validation for the course assessments. The course-based results were significantly correlated with the VALUE outcomes, but the degree varied greatly. The course assessments comprised a number of aspects that were not necessarily captured by the VALUE rubric and these aspects may have conflated results. For example, most of the course rubrics included dimensions for written communication, and some for technical skills. The areas with low correlations may be specific to one or two courses; further research will be undertaken to address this area of analysis. The correlations at the dimension level could also be investigated, but this would have required a mapping exercise to determine which of the course rubric dimensions matched to which VALUE dimensions. Further

complicating this was that many of the course rubrics combined language from multiple dimensions. Steps have been put in place to tag the course rubrics with the relevant VALUE dimension code. For example, a course rubric dimension that mapped to several VALUE dimensions would read like this: "Hazard Identification [CT1, CT2]."

Given that the participating courses identified their focus of cognitive development as part of the project, the overall course percentage has a strong relationship to the VALUE assessment; for example the correlation between the course percentage and the PS average was $r(284)= .603$ $p<.001$. Sessional GPA is derived by all the assessments the student completed in all their courses for that term, some of which had little focus on specifically assessing cognitive skills. The decreasing correlations between the VALUE results and the other course measures speak to divergent validity of the assessment. The cumulative GPA is an aggregate of all of the assessment thus far in the student's program; it could be thought of as representing overall skill and effort. It is therefore apt that the transferable skill and effort involved in the HEIghten test were moderately correlated with Cumulative GPA.

Value-add between First- and Fourth-year Cognitive Skills Achievement

Analysis of quantitative data suggests significant growth of cognitive skills between first and final year. The Cohen's effect size of $d= .497$ in critical thinking performance on the HEIghten test was reflected by gains in two levels across the median level of critical thinking performance on the VALUE rubric. It is important to recognize, however, that there are still students in their final year of undergrad scoring at the Benchmark level. Students ranged in performance level by program, but all of the programs demonstrated improvement between first and final year (see Appendix 2). Effort was a significant factor, and for reliability of data, continued attention should be paid to methods of incentivizing student effort. Methods such as the provision of course percentage marks to the completion of the test proved to be effective in this regard. This provision, however, is dependent on the commitment of the instructor.

Building Instructor Capacity

Student learning in all but two of the redesigned course assignments was evaluated using a type of quality framework such as a rubric or holistic grade descriptors. This adaptation by instructors was especially impressive, as 10 of the instructors had no experience using rubrics prior to participating in the CAR project. Generally speaking, the instructors found the (re)designed rubrics were very effective in communicating expectation to students: "I have become a real convert to rubrics, and I have already recommended the Cognitive Assessment Redesign to a friend in a different department." (Instructor H)

There are instructors in every institution who are early adopters and lead the way for others to promote change. Although there were some individuals in the group who would be described as early adopters, with 25 courses involved, the participants were not the "usual suspects." The teaching survey results indicated significant differences in teaching attitudes (see Figure 9). The survey data suggested that there was a

significant positive change for eight of the instructors. Most of those that didn't see significant change were already highly engaged in teaching, and already had very positive teaching attitudes. The two who reported reduced attitudes toward teaching voiced difficulties with the implementation of new assessment regimes. They had made significant changes to the structure of their courses, requiring a great deal of work on their part, and some of their students had complained that the assignment was too hard.

Making significant realizations or discoveries about pedagogical approaches or assessment was also indicative of growth. These transformations were usually preceded by challenges. There were various challenges for all the instructors, but they appeared to be greater for fourth-year courses where the assignments needed to be tailored to individual student content focus. Instructors experienced difficulty in applying a common rubric to the student assignments; for example, "the first priority is to revisit the assignments and update them based on the project work. A challenge that required continued attention is offering students multiple assignment options while using a universal rubric" (Instructor D). Likewise, instructors teaching large courses referred to the need to ensure consistency of marking by multiple teaching assistants (TAs). Some of the courses engaged in marking calibration with their TAs. Support from the project personnel facilitated marking calibration exercises, with positive comments such as: "I thought it was very helpful and gave me good insight for why and how I can do this for all assignments in all courses that I teach" (Instructor N). Given the positive experience with calibration exercises, a wider adoption of this consistent scoring process could easily be implemented.

Building capacity among instructors was partially influenced by the conversations between diverse subject areas that were rich with recognition and respect for similarities and differences. At one of the CAR instructor network meetings, two instructors articulated their feelings about rubrics being too lengthy or cumbersome to be practical. These instructors were able to openly discuss these challenges with other CAR instructors and as a group discussed constructive options or alternatives for alleviating the issue. As a result, both of these instructors opted to participate again in the following year of the project. While the project provided support through assessment facilitators, equally important was the initiation of a peer-support network. Three of the course instructors have transitioned into a formal mentor role. The role was labeled "assessment coach," and next steps for the research include investigating how these roles may be sustained. This role is one of a "critical friend," who is a coach, mentor or discussant, and assists in developing course materials and reviewing assessment materials. The time spent on this project as an assessment coach has been officially recognized by the faculty as counting toward administration load. By leveraging personal networks such as assessment facilitators and coaches, instructors were able to seek advice and critical feedback, resulting in the development of instructional, pedagogical, and curricular knowledge (Kreber and Cranton, 2000).

Propagation of the Initiative

There were three network meetings, one in the fall, one in the winter, and a day-long mini-symposium in the spring, anchored by invited speaker Dr. Randy Bass.[1] The purpose of the meetings was to share experiences and build a community of practice. As part of the CAR project, participating instructors were expected to present their initiative and their experiences in the project. This sharing of ideas and experiences was the primary goal of the meeting, such as those goals for communities of practice (Wenger, 1998). The presentations prompted conversation across the disciplinary groups. Prior to the meeting, instructors generally believed that the challenges they were facing were specific to their discipline. As a result of the meetings and presentations, challenges such as assumptions regarding students' abilities to incorporate perspectives beyond their discipline, however, were found to be quite universal. Conversations often reached beyond the meetings and had impact back in the departments. For example, "involvement in the project has prompted a larger discussion about the nature of assessment in the course labs. We have yet to come to a resolution, but the discussion is rich and important" (Instructor B).

A strategy used to promote instructor engagement was to include them as research co-investigators. Where requested, the research manager completed amendments to ethics to describe individual research designs, further enabling scholarly investigation by co-authoring papers as required. The initiative has provided a rich professional development opportunity for many instructors, some of whom are conducting their own scholarly investigation (references for these are listed in Appendix 4). There were seven conference presentations from project participants on work directly related to their assessment redesign. Many of the course instructors have been making ongoing improvements to their courses. Some of the instructors were already collaborating with their redesign efforts; the nursing instructors who implemented similar critical reading assignments in the first and fourth year, have developed similar course rubrics to investigate the growth in critical thinking through the nursing program. There has been a positive response to tackling challenges in the future; for example one instructor said, "we plan to provide students with more examples of the type of work we are expecting of them and also have them spend time critically evaluating other studies (for example, published and peers) to gain a better understanding of the process in advance" (Instructor M).

1 Randy Bass is the Assistant Provost, Teaching and Learning Initiatives, and Professor of English at Georgetown University. He was the "Principal's Distinguished Visitor for the Enhancement of Learning, and his keynote was titled: "Assessment Matters: Integrative Learning in a Dis-integrative Era."

One of the major research outputs at the "macro" level was the development of an institutional guide for the assessment of cognitive skills (Simper, Fostaty Young, Frank & Scott, 2018). The guide was written as a handbook to enable the project to transition from funded research to ongoing and sustainable evidence-based practice.

Conclusions

The cognitive assessment redesign has been an institution-wide, network-based project focused on the development and assessment of cognitive skills (critical thinking, creative thinking and problem solving). This project focused on two main objectives: supporting instructors as they develop course-based assessments and assess the development of students' cognitive skills through standardized measures. Through the implementation of a network, 24 instructors from 15 departments were supported in constructively aligning assessment of cognitive skills, effectively using quality metrics. Interestingly, 40% of the instructors involved, implemented rubrics for the first time to assess student artefacts in their course. Qualitative analysis of the narrative course reports was conducted to develop growth and capacity-building indicators, with promising results demonstrated on many of these metrics. Moreover, instructors describe the significance of the network including the support of assessment facilitators and peer-support networks. One area for further improvement would be the wider adoption of formal processes for consistent scoring of course assignments. Data from instructor surveys demonstrated the inherent challenges in implementing culture change in assessment practices but also the degree of capacity building achieved. It is recommended that these metrics be tracked in the future to determine the long-term effects of the project.

To evaluate student learning, three learning measures (course rubrics, VALUE rubrics and the HEIghten test) were used. Analysis of student assessment data found significant correlations between course-based marks and VALUE rubric scores and demonstrated significant gains in critical-thinking and problem-solving skills between first and final year. This data has implications at the institutional level, highlighting students' increased performance in cognitive skills and also at the department level, informing curricular decisions or course improvements. VALUE rubric scoring showed that between 20% and 40% of fourth-year students demonstrated performance at the level of Benchmark 2 or lower in dimensions related to problem solving and critical thinking, and the HEIghten test scores indicated that 13% of fourth-year students performed at the developing level, whereas 59% of fourth years performed at the proficient level and 28% at the advanced level. Assessment measures like these provide an opportunity to look for cohorts of students that fall below expected competence levels. Given the significance of cognitive skill development in undergraduate education, the next iteration of the CAR project will continue to mobilize and sustain an institution-wide assessment network aimed at developing and assessing these skills with an emphasis on the role of discipline-specific assessment facilitators and coaches.

References

Ashford-Rowe, K., Herrington, J., & Brown, C. (2013). Establishing the critical elements that determine authentic assessment. *Assessment & Evaluation in Higher Education*, *39*(2), 205–222. https://doi.org/10.1080/02602938.2013.819566

Banta, T. W., & Palomba, C. A. (2014). *Assessment essentials: Planning, implementing, and improving assessment in higher education*. John Wiley & Sons. https://books.google.ca/books?hl=en&lr=&id=rO_sBQAAQBAJ&oi=fnd&pg=PA13&dq=Assessment+Essentials:+Planning,+Implementing+and+Improving+Assessment+in+Higher+Education&ots=NiNDO4gSAq&sig=I_DgF-oXEE_dLOSzVfFX0w2pX-s

Bernstein, D., & Greenhoot, A. F. (2014). Team-designed Improvement of writing and critical thinking in large undergraduate courses. *Teaching and Learning Inquiry: The ISSOTL Journal*, *2*(1), 39–61. https://doi.org/10.1353/iss.2014.0008

Biggs, J. (2014). Constructive alignment in university teaching. *HERDSA Review of Higher Education*, *1*(1), 5–22.

Boud, D., & Dochy, F. (2010). *Assessment 2020. Seven propositions for assessment reform in higher education* (pp. 1–4).

Chasteen, S. V., & Code, W. (2018). *The science education initiative handbook. A practical guide to fostering change in university courses and faculty by embedding discipline-based education specialists within departments.* Boulder: The University of Colorado Boulder.

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

Corbo, J. C., Reinholz, D. L., Dancy, M. H., Deetz, S., & Finkelstein, N. (2014). Sustainable Change: A Model for Transforming Departmental Culture to Support STEM Education Innovation. *ArXiv Preprint ArXiv:1412.3034*. http://arxiv.org/abs/1412.3034

Council for Aid to Education. (2016). *CLA+ National Results, 2015-16* (pp. 1–21). New York, NY. http://cae.org/images/uploads/pdf/CLA_National_Results_2015-16.pdf

Dawson, P. (2017). Assessment rubrics: Towards clearer and more replicable design, research and practice. *Assessment & Evaluation in Higher Education*, *42*(3), 347–360.

Drezek McConnell, K., & Rhodes, T. L. (2017). *On Solid Ground, Value Report 2017* (pp. 1–52).

Frank, B. M., Simper, N., & Kaupp, J. A. (2016). How we know they're learning: Comparing approaches to longitudinal assessment of transferable learning outcomes. In *ASEE Annual Conference and Exposition, Conference Proceedings* (Vol. 2016-June).

González, J. & Wagenaar, R. (2003). *Tuning educational structures in Europe*. University of Deusto Final report. Phase one. Bilbao.

http://eua.be/eua/jsp/en/upload/TUNING_Announcement_Closing_Conference.1084282515011.pdf

Hannah, S. T., & Lester, P. B. (2009). A multilevel approach to building and leading learning organizations. *The Leadership Quarterly*, *20*(1), 34–48. https://doi.org/10.1016/j.leaqua.2008.11.003

Hanson, M. (2001). Institutional Theory and Educational Change. *Educational Administration Quarterly*, *37*(5), 637–661. https://doi.org/10.1177/00131610121969451

Hathcoat, J. D., & Penn, J. D. (2012). Generalizability of student writing across multiple tasks: A challenge for authentic assessment. *Research & Practice in Assessment*, *7*(2), 16–28.

Henderson, C., Beach, A., & Finkelstein, N. (2011). Facilitating change in undergraduate STEM instructional practices: An analytic review of the literature. *Journal of Research in Science Teaching*, *48*(8), 952–984. https://doi.org/10.1002/tea.20439

Hutchings, P., Huber, M. T., & Ciccone, A. (2011). *The scholarship of teaching and learning reconsidered: Institutional integration and impact* (Vol. 21). John Wiley & Sons. Retrieved from https://books.google.ca/books?hl=en&lr=&id=GPbdPMVOyAIC&oi=fnd&pg=PT6&dq=Hutching,+Huber,+and+Ciccone,+2011&ots=xiottWIwMO&sig=FtBxtufGR5dCIC4oKMmRj4Ykfno

Jankowski, N., Hutchings, P., Ewell, P., Kinzie, J., & Kuh, G. (2013). The degree qualifications profile: What it is and why we need it now. *Change: The Magazine of Higher Learning*, *45*(6), 6–15.

Jonassen, D. H. (1997). Instructional design models for well-structured and iII-structured problem-solving learning outcomes. *Educational Technology Research and Development*, *45*(1), 65–94. https://doi.org/10.1007/BF02299613

Jonsson, A., & Svingby, G. (2007). The use of scoring rubrics: Reliability, validity and educational consequences. *Educational Research Review*, *2*(2), 130–144.

Kezar, A. (2013). "Theories of change: Change agent guides." In *How colleges change understanding, leading, and enacting change* (1st Edition, pp. 20–40). Taylor & Francis Group. https://www.taylorfrancis.com/

Kreber, C., & Cranton, P. A. (2000). Exploring the scholarship of teaching. *The Journal of Higher Education, 71*(4), 476–495. doi:10.1080/00221546.2000.11778846

Kinzie, J. (2010). Perspectives from campus leaders on the current state of student learning outcomes assessment. *Assessment Update*, *22*(5), 14–15.

Koretz, D. M. (2008). *Measuring up*. Harvard University Press. https://books.google.ca/books?hl=en&lr=&id=Nw-qBPCTcYYC&oi=fnd&pg=PA1&dq=Koretz+2008+high+stakes+testing&ots=k8YvOqRftq&sig=wljSkJZCKvh-iuMQ53Gxc-WK73I

Liu, O. L., Bridgeman, B., & Adler, R. M. (2012). Measuring learning outcomes in higher education: Motivation matters. *Educational Researcher*, *41*(9), 352–362. https://doi.org/10.3102/0013189X12459679

Madaus, G. F., & Clarke, M. (2001). The adverse impact of high stakes testing on minority students: Evidence from 100 years of test data. http://eric.ed.gov/?id=ED450183

Mathers, C. E., Finney, S. J., & Hathcoat, J. D. (2018). Student learning in higher education: A longitudinal analysis and faculty discussion. *Assessment & Evaluation in Higher Education*. https://srhe.tandfonline.com/doi/abs/10.1080/02602938.2018.1443202

Miller, E. R., Fairweather, J. S., Slakey, L., Smith, T., & King, T. (2017). Catalyzing institutional transformation: Insights from the AAU STEM initiative. *Change: The Magazine of Higher Learning*, *49*(5), 36–45.

Popham, W. J. (1997). What's wrong — and what's right — with rubrics. *Educational Leadership*, *55*, 72–75.

Popham, W. J. (1999). *Classroom assessment: What teachers need to know*. ERIC.

Purmton, E. F., & Alexander, E. R. (2013). Obtaining faculty motivation and "Buy In" to a major program change: A case study in assessment. *Business Education Innovation Journal*, *5*(1), 53–57.

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.

Roxa, T., & Martensson, K. (2015). Microcultures and informal learning: A heuristic guiding analysis of conditions for informal learning in local higher education workplaces. *International Journal for Academic Development*, *20*(2), 193–205. https://doi.org/10.1080/1360144X.2015.1029929

Roxa, T., Martenson, K., Alveteg, M., (2011). Understanding and influencing teaching and

learning cultures at university: A network approach. *Higher Education, 62*(1), 99–111. doi:10.1007/s10734-010-9368-9

Simper, N., Fostaty Young, S., Frank, B., & Scott, J. (2018). Guide for Institutional Assessment of Cognitive Skills. Queen's University. http://www.queensu.ca/qloa/sites/webpublish.queensu.ca.qloawww/files/files/CAR%20handbook-Web.pdf

Simper, N., Frank, B., Kaupp, J., Mulligan, N., & Scott, J. (2018). Comparison of standardized assessment methods: Logistics, costs, incentives and use of data. *Assessment and Evaluation in Higher Education*, *44*(6), 1–16.

Simper, N., Frank, B., Scott, J., & Kaupp, J. (2018). *Learning Outcomes Assessment and Program Improvement at Queen's University* (pp. 1–55). Ontario: Toronto: Higher Education Quality Council of Ontario. http://www.heqco.ca/SiteCollectionDocuments/Formatted%20Queens_LOAC_report.pdf

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.

Tremblay, K. (2013). OECD Assessment of Higher Education Learning Outcomes (AHELO). In *Modeling and measuring competencies in higher education* (pp. 113–126). Springer. http://link.springer.com/chapter/10.1007/978-94-6091-867-4_8

Tremblay, K., Lalancette, D., & Roseveare, D. (2012). *Assessment of Higher Education Learning Outcomes feasibility study report: Volume 1 - design and implementation*. OECD.

Wenger, E. (1998). *Communities of practice: Learning, meaning, and identity*. Cambridge, U.K;New York, N.Y;: Cambridge University Press.

Wenger, E. (2000). Communities of practice and social learning systems. *Organization*, *7*(2), 225–246.

Wenger, E. (2011, October). *Communities of practice: A brief introduction*. Presented at the STEP Leadership Workshop, University of Oregon. http://scholarsbank.uoregon.edu/xmlui/handle/1794/11736

Wenger, E., McDermott, R. A., & Snyder, W. (2002). *Cultivating communities of practice: A guide to managing knowledge*. Harvard Business Press. https://books.google.ca/books?hl=en&lr=&id=m1xZuNq9RygC&oi=fnd&pg=PR9&dq=communities+of+practice+wenger&ots=ZUadhIae9_&sig=KCT1b83mzJB-dm4O0zAO26DunRg

Wieman, C. (2007). Why not try a scientific approach to science education? *Change: The Magazine of Higher Learning*, *39*(5), 9–15.

Wieman, C., Deslauriers, L., & Gilley, B. (2013). Use of research-based instructional strategies: How to avoid faculty quitting. *Physical Review Special Topics-Physics Education Research*, *9*(2), 023102.

Wiggins, G. P., & McTighe, J. (2005). *Understanding by design*. Alexandria: Association for Supervision and Curriculum Development.

Higher Education Quality Council of Ontario

An agency of the Government of Ontario