New Orleans JAZZED about Engineering Education

Paper ID #16507

How We Know They're Learning: Comparing Approaches to Longitudinal Assessment of Transferable Learning Outcomes

Dr. Brian M. Frank, Queen's University

Brian Frank is the DuPont Canada Chair in Engineering Education Research and Development, and the Director of Program Development in the Faculty of Engineering and Applied Science at Queen's University where he works on engineering curriculum development, program assessment, and developing educational technology. He is also an associate professor in Electrical and Computer Engineering.

Ms. Natalie Simper, Queen's University

Natalie Simper coordinates a Queen's research project investigating the development and measurement of general learning outcomes. Natalie comes from an Australian Senior-Secondary/ Post-Secondary teaching background, with experience at the State-wide level in curriculum development, large-scale assessment, and evaluation and assessment of outcomes based education.

Dr. James A. Kaupp, Queen's University

Assessment and Quality Assurance Coordinator (Msc '06, PhD '12) at Queen's University, Kingston, Ontario, Canada in the Faculty of Engineering and Applied Science. Educational research interests include engineering education development, cultural change in higher education, higher-order thinking development and assessment, outcomes-based data-informed continuous improvement, information visualization & analysis and authentic performance-based assessment.

How we know they're learning: Comparing approaches to longitudinal assessment of transferable learning outcomes

Abstract

This research paper describes interim results from a 4-year longitudinal study of how engineering students develop critical thinking, problem solving, and communication skills. The sample includes approximately 400 students in a mid-sized research intensive Canadian institution. The students were assessed using multiple approaches, including standardized tests, in-course activities, surveys, and course artefacts scored by a trained team using program-wide rubrics. Outcomes demonstrated in student course artefacts externally scored by VALUE rubric assessment increased over the two years. Scores on standardized tests generally trend upward with the Critical thinking Assessment Test (CAT) but are mixed on the Collegiate Learning Assessment (CLA+), most likely due to motivational and alignment issues. Student motivation is a significant issue in the project. The paper compares the assessment methods, and finds that using externally scored course artefacts is both less expensive and preferred by course instructors for course and program improvement over standardized tests.

Introduction

The general transferable intellectual skills of critical thinking, problem solving, communication and lifelong learning are key engineering accreditation requirements, and fundamental elements of undergraduate education, but "are often considered to be among the most difficult outcomes to define, teach and assess" [1]. There has been significant interest in measuring fundamental transferable intellectual skills (TIS) like critical thinking, communication, problem solving, and lifelong learning [2–7]. Past work has found average performance gains in broad TIS like critical thinking and written communication, for example, with effect sizes around d=0.5 standard deviations (SD) over a four year program[3,6,7], with some finding differences between majors[3] and others finding no strong evidence for this[6].

TIS are found in some form in the Washington Accord [8], the Essential Learning Outcomes and VALUE rubrics[9] from the Association of American Colleges and Universities, the Degree Qualifications Profile from the Lumina foundation[10], among others[11]. Program improvement requires reliable data about students' performance, necessitating an efficient and effective approach to assessment in undergraduate programs.

In 2013 the researchers established a longitudinal exploratory study of TIS development at Queen's University, a Canadian research-intensive institution with approximately 25,000

students. The institution uses a two-semester system that runs from September-December and January-April. The research study is assessing critical thinking, problem solving, written communications and lifelong learning in disciplines spanning engineering, science, social science, and humanities. In the first year of the study over 2000 first and fourth year students from the Faculty of Arts and Science (Psychology, Drama and Physics), and from the Faculty of Engineering and Applied Science (Chemical Engineering, Civil Engineering, Geological Engineering, and Mechanical Engineering) consented to participate in the project. Currently over 400 engineering students have consented to be part of the project over the past two years.

The project is using multiple assessment methods, and analyzing the approaches by factors including cost, utility, alignment with both project and instructor goals. This project differs from past work in that the goal is to compare the suitability of various approaches for long-term program improvement, where the assessment must yield results that can inform course and curriculum change.

Project Goals and Research questions

The frameworks and selection of tools described in this section are summarized from a previous publication that also provides more details about the motivation and context for this study [12]. The study focuses on use of authentic assessments, in which students are asked to demonstrate competencies on contextualized tasks that emulate real-world situations. This provides alignment with program outcomes, and consequential validity.[13,14]

This four-year study is following a cohort through an undergraduate program using four approaches to assessing critical thinking, problem solving, written communications, and lifelong learning, as illustrated in Figure 1, and include:

- standardized instruments
- general rubrics used to score artefacts created by students for academic purposes across multiple years and programs
- qualitative evaluation using student/instructor focus groups and interviews
- data linkage to registrar data, course grades and measured course learning outcomes

The project is investigating the utility of these assessment instruments to understand development of TI and to encouraging faculty to develop and assess transferable skills in their courses and programs, with the goal of building a foundation for a wider rollout across faculties and programs in the coming years. The study is documenting the costs, time commitment, participation rates, and correlations between these approaches, and evaluating the value and reliability of the measures.

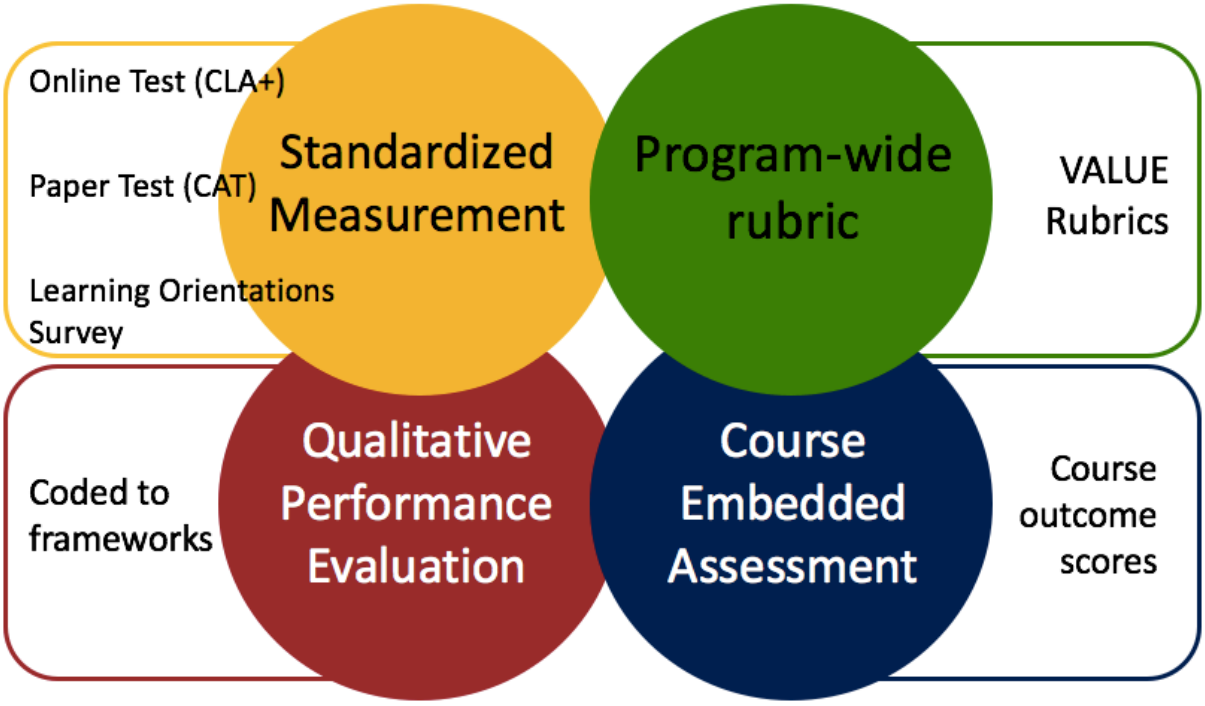

Figure 1 - Four categories of data used in this study.

The primary research questions in this study are:

1. How do TIS develop over time in the cohort under study?
2. How do the assessment approaches compare on measures including efficacy and efficiency?

Assessment tools

As shown in Figure 1 above, the tools in this study include the Collegiate Learning Assessment Plus (CLA+) Test, the Critical Thinking Assessment Test (CAT), and the AAC&U VALUE rubrics, and a new triangulated measure of Transferable Learning Orientations (TLO) based on the VALUE rubric for lifelong learning and the Motivated Strategies for Learning Questionnaire (MSLQ). The first three are described in the sections below; preliminary work results from the TLO have been published separately[15].

The two standardized tools, the CLA+ and CAT, were selected because they assess both critical thinking and written communication and include open-text response questions emulating the kinds of tasks expected in realistic situations. Other tools were considered, including the California Test of Critical Thinking and the Cornell Level Z, but previous experience found that multiple choice questions of critical thinking did not address the ability

to formulate an argument, the ability to evaluate effort, nor disposition to apply critical thinking to a problem.

The Collegiate Learning Assessment (CLA+) is a web-based holistic assessment of critical thinking, problem solving, and written communication instrument. The test takes 90 minutes to complete, with 60 minutes allotted to an open-text response to a realistic situation (performance task) and 30 minutes allotted to multiple-choice or short answer that measure scientific and quantitative reasoning, critical reading and evaluation, and critiquing an argument. The test comprises two sections, reporting the following sub-scores:

- Performance task (PT): analysis and problem solving, writing effectiveness, and writing mechanics
- Selected Response Questions (SRQ): scientific and quantitative reasoning, critical reading and evaluation (detecting logical flaws), and questionable assumptions to critique arguments

The test asks the students to judge the evidence base and decide if the claims supported by the evidence, and analyze the data/ evidence and decide on a course of action or propose a solution. The CLA+ total score is an average of the scaled PT score and weighted SRQ scores. Details about the validation of the CLA+ has been presented previously[16,17].

The Critical thinking Assessment Test (CAT) is designed to assess multiple areas of critical thinking and problem solving[18]. It was developed by different institutions across the United States with support from the National Science Foundation (NSF). The test is paper based and is completed in 60 minutes. Like the CLA, the CAT test questions are based on real world situations, most requiring short answer essay responses. Tests were scored on campus, using an assessment protocol by trained markers, then a random sample were scored by the test creators.

The VALUE Rubrics were developed by the American Association of Colleges and Universities to provide a valid assessment of learning in undergraduate education[7]. These rubrics are broad, discipline-neutral descriptions of selected essential learning outcomes of undergraduate education There are four levels of performance criteria, from the benchmark level of a student entering university to the capstone level of a student who has just completed their undergraduate experience. There has been considerable effort in establishing the validity and reliability of the VALUE rubrics[9,19].

The VALUE rubric marking was conducted by raters whose training addressed the specific context and content of course assignments. Raters were undergraduate students and graduate students, with faculty called on for subject area expertise when necessary. The raters were engaged longitudinally through the study and where possible markers used across disciplines to provide consistency of ratings. This stresses the importance of having a well-planned, well-

supported process to rate artefacts using the VALUE rubrics and an environment which facilitates rater discussion and interaction.

Participants and Results

Participants consented to participate in standardized tests and to have samples of their course work scored by trained graders using VALUE rubrics. The subsections below describe the results from these two approaches.

CAT and CLA+

The total first year consenting student sample of engineering students was 378, and 516 in the second year consenting sample. The students were randomly selected from each of the eight engineering departments tested, and randomly assigned in first year to either the CAT (n=151 first year, n= 343 second year), or CLA+ (n= 227 first year, n= 174 second year).

In both years, the testing was administered during a course lab time slot. Wherever possible, each second year student was provided the same test as they completed in their first year. If students did not have access to a computer for the online CLA+ test they were provided the CAT instead. The first year testing was conducted in week one of first year, the second year testing was conducted toward the end of the second year (4- 6 weeks before the end of the winter term). After the CAT was scored by trained graders a random sample of tests were rescored by the test creators; the average difference between the institutionally-scored mean and the scoring accuracy check mean was less than 3%.

Mean scores for each of the disciplines and year groups are reported in Table 1, and summarized in **Error! Reference source not found.**. There was a significant improvement between first and second year on CAT scores, with larger effects , $F(1, 494) = 7.34$ $p = .007$. The overall effect size was fairly small, 0.241, though it varied considerably by discipline. There was no observable improvement in CLA+ score between first and second year, and possible reasons for this are discussed below.

Value Rubric Assessment

In both years, the VALUE rubric marking of course projects was matched to the available consenting population who sat either the CAT or CLA+. The first year marking included two assignments, the first project was a model eliciting activity completed in groups with an individual component[20]. Fifty-two assignments were included in the sample, representing 152 students. The second course assignment was taken from the second semester and was a final team design report. Twenty-four assignments were used in that sample, representing 99 students working in teams of 5, on average.

The second year course assignments varied slightly depending on the discipline. All students were in the same course, but the activity and deliverable varied by discipline. In most cases, students were asked to design something specific to the discipline. For example:

- a new or improved equipment for a specific geographic location, so they could be integrated into current professional practices, or
- a grouping algorithm that would optimize a real world process, or
- a station that performed a specific task within an assembly line, or
- a rapidly and cheaply deployable tool that could be used to measure one or more environmental variables, or
- a truss bridge that could span a set distance with specific design measurements, or
- a plan for large boats to access a harbour to dock at a fictional theme park, or
- a plan to build a parking garage for the incoming theme park

In most programs students were instructed to include the following in their reports: the problem definition, background research, creative thinking/decision-making, economic and triple bottom line considerations. Students were also asked to include an assessment of their final design. The second year sample represented 178 consenting students working in teams of 3, on average.

There was a significant improvement between first and second year on average VALUE rubric scores, $F(1, 430) = 15.32$, $p<.01$. The improvement in student's critical thinking, problem solving and written communication in engineering courses was demonstrated with a medium effect size of $d= .395$. Since the artefacts were team activities, the scores are expected to be higher, on average, than would be expected if individual performance was measured [21].

Table 1 and Figure 2 compare data by engineering discipline (i.e. department) in year one and two. For the CAT and VALUE rubric scoring, the mean gain in scores varied by department from negligible to moderate, but the CLA+ scores for some disciplines (1, 2, 3 and 8) actually dropped. The greatest observed drop in CLA+ scores were in disciplines with a heavy technical focus and where the mean reported effort on the test was low. In the CLA+, students report the effort they put into the test on a scale from one to five (1= no effort, 2=little effort, 3= moderate effort, 4= a lot of effort, 5= best effort). In second year students put significantly less effort into the test $F(1,401)= 28.58$, $p< .01$). In the disciplines with negative effect sizes there were between 30% and 45% of students who put little or no effort into the test. The greatest VALUE rubric gains were observed when the following characteristics were true: (1) the course instructor engaged a real-world applied approach to the curriculum, and (2) the assignment scored using the VALUE rubric aligned well with the project outcomes. The second year assignments in disciplines 1 and 6 did not align very well with the project outcomes, whereas in the first year the assignments consistently aligned well.

Table 1 – Descriptive statistics for first and second year engineering on CLA+, CAT and VALUE rubric scores

Discipline	Year	CAT Score n	CAT Score Mean (SD)		CAT Score *d*	CLA+ Score n	CLA+ Score Mean (SD)		CLA+ Score % low effort	CLA+ Score *d*	VALUE Rubric Average n	VALUE Rubric Average Mean (SD)		VALUE Rubric Average *d*
1	1	26	19.1	(5.4)	.496	37	1170.6	(125.1)	15	-.154	46	1.85	(.35)	.059
	2	64	21.9	(5.9)		28	1150.6	(133.8)	30		33	1.87	(.33)	
2	1	21	21.0	(5.2)	.171	34	1165.4	(92.9)	11	-.229	41	2.22	(.42)	.675
	2	39	21.9	(5.3)		34	1140.7	(123.1)	45		22	2.49	(.38)	
3	1	20	21.1	(5.4)	.000	38	1188.0	(116.2)	10	-.379	39	1.94	(.46)	.870
	2	40	21.1	(5.3)		40	1135.8	(158.9)	23		24	2.34	(.46)	
4	1	11	20.6	(5.7)	.227	21	1215.1	(120.3)	14	-	16	1.98	(.29)	-
	2	36	21.7	(4.0)		-	-	-	-		4	2.25	(.21)	
5	1	20	19.1	(6.9)	.307	8	1108.9	(118.5)	15	.200	23	2.17	(.51)	.168
	2	28	21.2	(6.8)		10	1132.0	(112.2)	16		16	2.25	(.44)	
6	1	29	21.5	(5.1)	.178	57	1145.2	(87.2)	16	.116	51	1.95	(.32)	.092
	2	81	22.4	(5.0)		39	1156.3	(104.3)	35		42	1.98	(.33)	
7	1	12	18.1	(4.8)	.557	21	1180.7	(128.0)	1	.119	19	1.8	(.47)	.747
	2	32	20.8	(4.9)		12	1194.1	(97.0)	38		20	2.17	(.52)	
8	1	10	22.6	(5.9)	.082	10	1230.9	(125.4)	13	-.742	16	2.05	(.47)	.652
	2	22	23.0	(3.8)		10	1147.3	(99.9)	30		17	2.34	(.42)	
Total	1	151	20.5	(5.6)	.241	227	1174.0	(113.1)	13	-.212	251	1.99	(.42)	.395
	2	342	21.8	(5.2)		174	1148.7	(125.3)	30		178	2.16	(.44)	

CLA+ "low effort"= reported little or no effort put into the test

Effect size *d* calculated M2-M1 divided by pooled SD.

Missing values where test type was not able to run, or sample size was too small to calculate

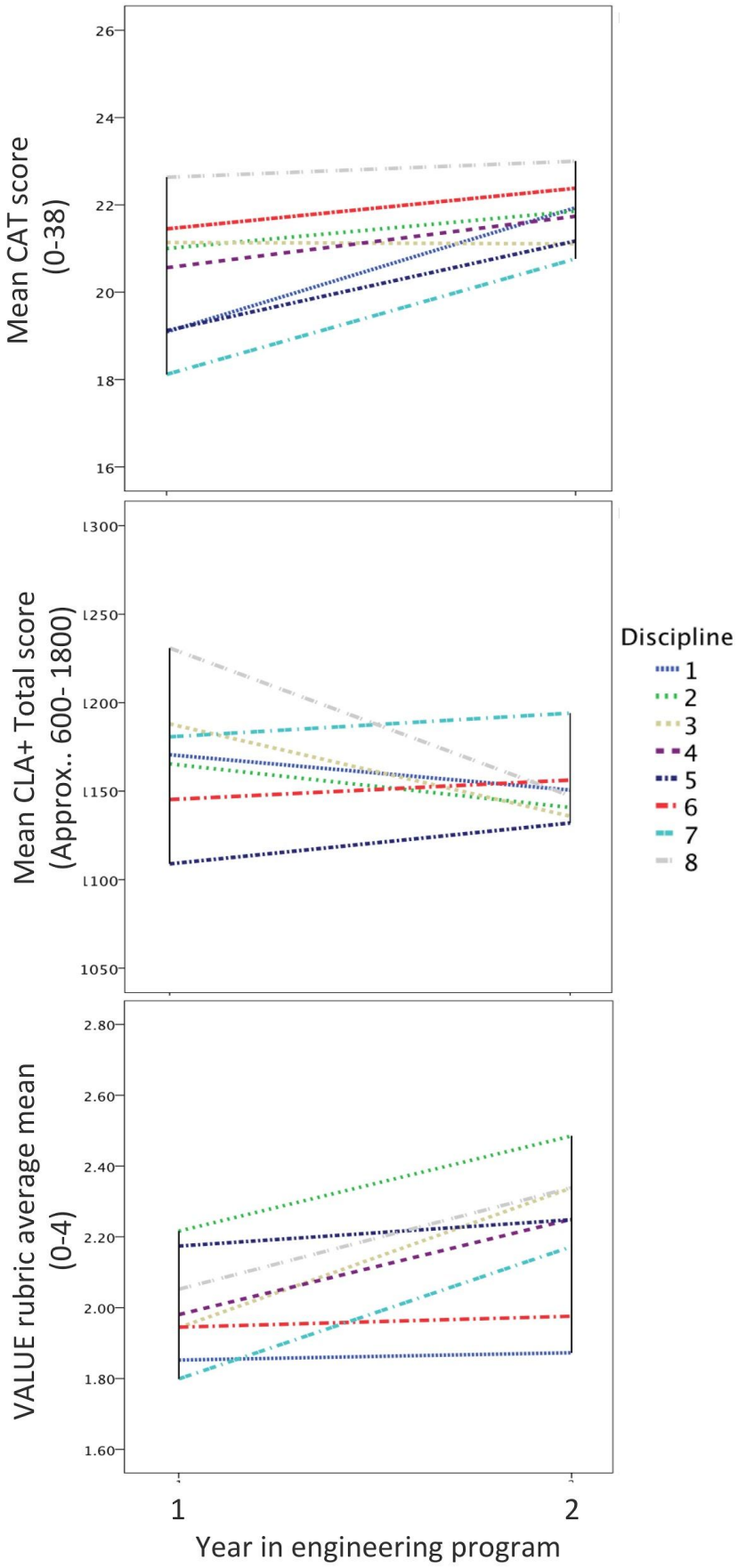

Figure 2 – Mean score on the CAT, CLA+ and Value rubric by year and discipline.

Many correlations between test scores were moderate, but an interesting one was the correlation between the CLA+ overall score and (a) student self-reported estimates of effort and (b) the time students spent on the task. Generally CLA+ scores do increase with effort and time spent.

Cost Benefit Analysis

Qualitative and quantitative methods were used to analyze the costs and perceived benefits of each of the tools. The quantitative analysis was based on the larger project sample comprising the Arts and Science departments of Drama, Physics and Psychology, together with the Engineering sample presented in the results section of this paper. For the CLA+ n= 598 over two years, and n=621 for VALUE rubric scoring of course artefacts. The associated costs were calculated by summing the fee for the instrument, the ancillary costs (training fees and salaries), and the salaries for markers for the time spent marking each sample (Undergraduates were paid $14 per hour and graduate markers paid at $35 per hour), then this sum was divided by the valid number in the sample. Although the fee per test taker for the CLA+ is $35, once the additional costs have been taken into account, and the number in the sample excluded due to incomplete data, the fee per valid n was $41.87.

The CAT test costs $8 per student, and there is an admin fee of $200. CAT marking protocols require each paper to be double marked and in the case of questions without exact agreement, a third marker to be used. This process took an average of 30 min per paper to complete. After test fees, marking, and proctoring costs, the cost per test was $28. The VALUE samples took varying amounts of time to mark, they ranged between 30 minutes and 3 hours depending on the complexity of the artefact. The first-year samples were marked by undergraduates, whereas the fourth-year samples were marked by graduate students. The average cost per valid n for the VALUE rubric marking was $14.82. Figure 3 shows a comparison of cost per consenting sample for the CLA+, CAT, and VALUE rubrics.

Faculty feedback was provided by 12 of the participating instructors from the project. Instructors were all provided with a report for their course, detailing descriptives from each test with breakdowns on assessed criteria and subscores for the tests. Following a debrief session, they provided feedback to researchers about their perceived benefit of the tools under the following groupings *Ease of logistics; Confidence in the reliability and validity; Alignment to the course;* and *Applicability to affect course improvement*. Figure 3 summarizes the results of the survey. The VALUE rubrics are both less expensive and overall better received by instructors.

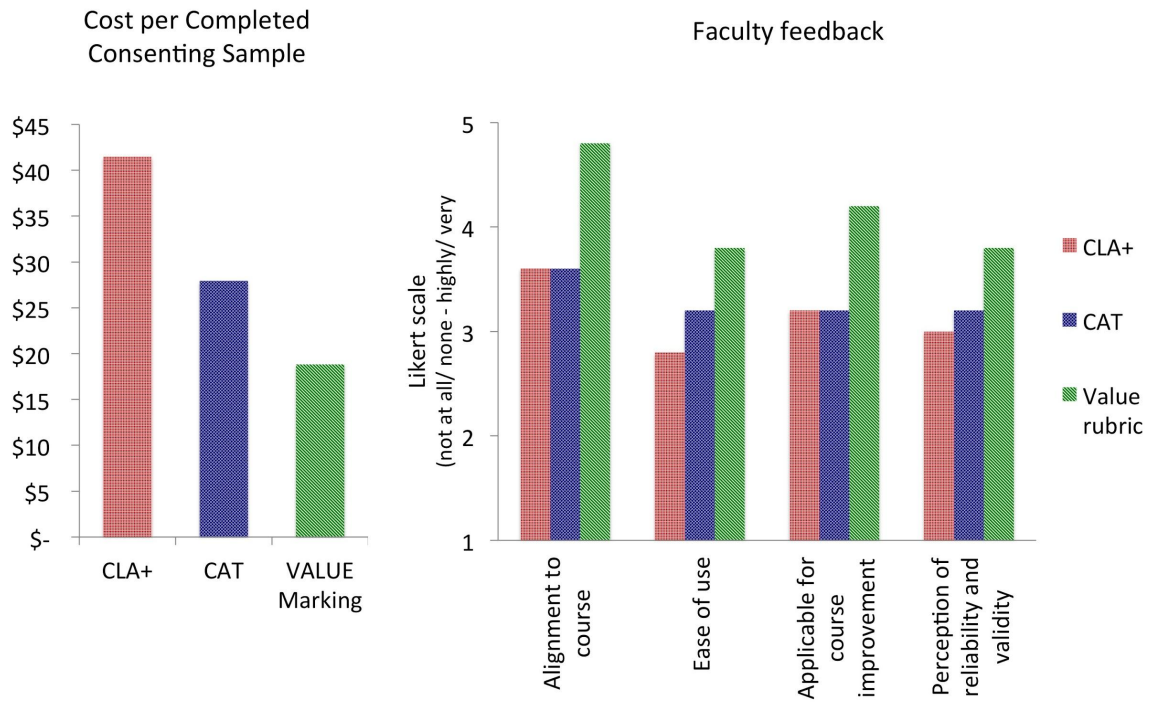

Figure 3 - Cost per completed consenting sample and instructor perceptions for each of the three assessment approaches.

Conclusions

The two research questions in this study are:

1. How do TIS develop over time in the cohort under study?
2. How do the assessment approaches compare on measures including efficacy and efficiency?

Regarding question 1, the development of transferable intellectual skills is observable using two of the three approaches described here. Overall we can observe significant changes ($p<0.01$) in performance on two of the measures, with effect sizes of $d=0.395$ using VALUE rubrics and $d=0.241$ using the Critical thinking Assessment Test. These effect sizes are in line with previous tests of critical thinking, as discussed above. No significant change is seen on CLA+ scores in the overall cohort, though there are when broken out by department. Results on the CLA+ are heavily impacted by student motivation in second year.

Regarding question 2, the VALUE rubrics are the most economically efficient when considering cost per valid response (by about a factor of two compared to the CLA+). The group of instructors who have received reports about their students' performance on the tools rank the VALUE rubric approach as being better aligned to their course, likely because the

scores they receive are assigned to academic work they created. Instructors found the standardized tests to be somewhat useful, but expensive and time consuming to run, whereas the feedback from the course-based assessments suggested they were more effective for supporting course improvement.

Instructors also perceived the VALUE rubric scoring to yield valid conclusions, more so than the other tests, likely due to the proximity of the task to the course activities. Since one of the goals of this work is to identify approaches that are sustainable inside the institution, instructor perceptions are important, though educating instructors on interpreting data is also important. Measures of test reliability for the standardized tools have been presented by the test creators, and inter-rater reliability of the VALUE rubric scores is around 90%. The factors influencing validity of conclusions drawn from the tests are:

1. For standardized tests, student effort and time spent may not reflect their ability. This is commonly seen in low-stakes standardized tests[22], and this study observed a drop in self-reported effort on the CLA+ from first year to second year. Student participation rates and effort improve when the assessment is completed within class time and encouraged by the course instructor.
2. For the VALUE rubric scored artefacts, the biggest threat to validity of interpretation is alignment between the assignment expectations and the VALUE rubrics. The course artefacts available in some courses sampled in second year were not well aligned with the VALUE rubrics, even though the courses were selected because they were the most promising option. For example, some programs do not expect their second-year students to write supported argument about an open-ended problem, to evaluate evidence, etc.

We find that there are faculty and instructors who are very open and keen to implement new assessment processes, but the majority are wary of change. Many faculty members feel that these intellectual skills are captured within their current assessment methods, and remain unconvinced about the need to specifically assess these skills. Our interim results suggest that the courses where students are not specifically directed to demonstrate critical thinking and problem solving (i.e. not elicited in the assignment), the performance (as rated using the VALUE rubrics) is significantly lower than their peers.

Fiscal and logistical consideration of assessment of this type is a concern for sustainability. Working toward long-term sustainability, the research team focuses much of their time and effort working on the ground with participating faculty to create instruments that minimize additional workload on their part and are virtually invisible to the student.

The authors hope that this study can help to inform sustainable assessment processes that can be used to improve the quality of education. It is only by knowing how well students develop in critical thinking, problem solving, and written communication, and identifying where there are weaknesses, that programs can make improvements.

References

[1] Deller, F., Brumwell, S., and MacFarlane, A., *The Language of Learning Outcomes: Definitions and Assessments*, Higher Education Quality Council of Ontario, 2015.

[2] Voogt, J., and Roblin, N. P., "A comparative analysis of international frameworks for 21st century competences: Implications for national curriculum policies," *Journal of Curriculum Studies*, vol. 44, Jun. 2012, pp. 299–321.

[3] Arum, R., Roksa, J., and Cho, E., "Improving Undergraduate Learning:' Findings and Policy Recommendations from the SSRC-CLA Longitudinal Project'.," *Social Science Research Council*, 2011.

[4] Goodman, K. M., Magolda, M. B., Seifert, T. A., and King, P. M., "Good practices for student learning: Mixed-method evidence from the Wabash National Study," *about caMPus*, vol. 16, 2011, pp. 2–9.

[5] State Higher Education Executive Officers Association, "MSC: A Multi-State Collaborative to Advance Learning Outcomes Assessment."

[6] Pascarella, E. T., and Terenzini, P. T., *How college affects students: A Third Decade of Research*, San Francisco: Jossey-Bass, 2005.

[7] Huber, C. R., and Kuncel, N. R., "Does College Teach Critical Thinking? A Meta-Analysis," *Review of Educational Research*, Sep. 2015, p. 0034654315605917.

[8] International Engineering Alliance, *Graduate Attributes and Professional Competencies*, 2013.

[9] Rhodes, T. L., and Finley, A. P., *Using the VALUE rubrics for improvement of learning and authentic assessment*, Association of American Colleges and Universities, 2013.

[10] Adelman, C., Ewell, P., Gaston, P., and Schneider, C., *The Degree Qualifications Profile*, Lumina Foundation, 2013.

[11] "Employability Skills Framework - Source Matrix."

[12] Frank, B. M., Kaupp, J. A., and Simper, M. N., "Multi-method longitudinal assessment of transferrable intellectual learning outcomes," *age*, Indianapolis, IN: 2015, p. 1.

[13] Wiggins, G., "The Case for Authentic Assessment. ERIC Digest.," Dec. 1990.

[14] Gulikers, J. T. M., Bastiaens, T. J., and Kirschner, P. A., "A five-dimensional framework for authentic assessment," *Educational Technology Research and Development*, vol. 52, Sep. 2004, pp. 67–86.

[15] Simper, N., Kaupp, J., Frank, B., and Scott, J., "Development of the Transferable Learning Orientations tool: providing metacognitive opportunities and meaningful feedback for students and instructors," *Assessment & Evaluation in Higher Education*, vol. 0, Jul. 2015, pp. 1–17.

[16] Zahner, D., *Reliability & Validity - CLA+*, Council for Aid to Education, 2013.

[17] Klein, S., Liu, O., and Sconing, J., *Test Validity Study (TVS) Report*, Council for Aid to Education, 2009.

[18] Stein, B., Haynes, A., Redding, M., Ennis, T., and Cecil, M., "Assessing Critical Thinking in STEM and Beyond," *Innovations in E-learning, Instruction Technology, Assessment, and Engineering Education*, M.I. PE, ed., Springer Netherlands, 2007, pp. 79–82.

[19] Finley, A., "Reliable Are the VALUE Rubrics?," *Peer Review*, vol. 13/14, 2012.

[20] Kaupp, J., and Frank, B., "Impact of Model Eliciting Activities on Development of Critical Thinking," *ASEE Annual General Conference*, Atlanta, GA: 2013.

[21] Cohen, E. G., "Restructuring the Classroom: Conditions for Productive Small Groups," *Review of Educational Research*, vol. 64, Mar. 1994, pp. 1–35.

[22] Attali, Y., "Effort in Low-Stakes Assessments What Does It Take to Perform as Well as in a High-Stakes Setting?," *Educational and Psychological Measurement*, Mar. 2016, p. 0013164416634789.