Multi-method longitudinal assessment of transferrable intellectual 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.

Dr. James A. Kaupp, Queen’s University

Educational Researcher and Adjunct Professor (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, educational data visualization & reporting and authentic performance-based assessment.

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. Jill Scott, Queen’s University, Kingston (Canada)

Jill Scott is Vice-Provost (Teaching and Learning) and Professor, Department of Languages, Literatures and Cultures. She is leading a number of assessment initiatives across the university, including assessment of transferable intellectual skills, assessment of active learning spaces, and program and curriculum assessment.
Multi-method longitudinal assessment of transferable intellectual learning outcomes

Abstract

This paper describes the first phase of a four-year longitudinal study of transferable intellectual learning outcomes at a research-intensive university in Canada. These skills, including critical thinking, problem solving, communication and lifelong learning, are the subject of discussion in higher education generally (e.g. the American Association of Colleges and Universities (AAC&U) Essential Learning Outcomes) and engineering in the form of accreditation requirements. The project is using multiple methods to assess the development of these skills in engineering, humanities, physical science, and social science sectors. The first year of the study which involved a double cross-sectional assessment of first and fourth year undergraduate students using standardized tests, meta-rubrics used to score course artefacts, and a learning orientation instrument. Interim results are showing significant development of transferable intellectual skills from first to fourth year, with significant differences between instruments in cost, alignment, and participation rates.

Introduction

Transferable intellectual skills (TIS) like critical thinking, communication, problem solving, and lifelong learning are fundamental elements of modern undergraduate engineering programs. They are critical to the practice of engineering, and support the large number of graduates who work in non-engineering roles. The Washington Accord, to which many national engineering accreditation bodies are signatories, incorporates these in graduate attributes problem analysis, communication, and lifelong learning. Broader frameworks in higher education, including the Essential Learning Outcomes from the Association of American Colleges and Universities, the Degree Qualifications Profile from the Lumina foundation, and Partnership for 21st century skills, include elements of TIS.

These elements are common elements of learning outcomes frameworks, including AAC&U Essential Learning Outcomes and VALUE rubrics, Lumina’s Degree Qualifications Profile, Higher Education Quality Council of Ontario (HEQCO) Tuning competences, graduate attributes in the Washington Accord adopted by ABET and the Canadian Engineering Accreditation Board, and employer groups surveys.

Studying the development of TIS is critical to improving the competency of university graduates and (hopefully!) to demonstrate competence of graduates. There has been significant interest in measuring the student competencies, including the SSHR-CLA longitudinal project, Wabash
This paper reports on an institutional multi-method longitudinal study involving around 2000 students, including approximately 650 engineering students. The study aims to examine the development of TIS and correlations between instrument scores, and compare resource requirements of the methods. It also aims to develop strategies for sustainable assessment achieved within standard course contexts and develop internal processes for implementation, management, and assessment of university-wide learning outcomes that recognize and enhance disciplinary expectations.

Methods

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 research study includes students 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.

This four-year study is following a cohort through an undergraduate program using four approaches to assessing TIS as illustrated in Figure 1, including:

- standardized instruments
- meta-rubrics used to score artefacts created by students for academic purposes
- in-course assessment of TIS
- group problem-solving sessions independent of academic work

The study is documenting the costs, time commitment, participation rates, and correlations between these approaches, and evaluating the value and reliability of the measures. In the first year of the study a double-cross sectional assessment was used to pilot the tools with first and fourth year students and compare performance. This paper describes the results from the standardized instruments and meta-rubrics used in the first year of the study. These include the Collegiate Learning Assessment Plus (CLA+) Test, 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). Each of these is described in the sections below.
The Collegiate Learning Assessment (CLA+) was developed by the Council for Aid to Education (CAE), and is an updated version of the original collegiate learning assessment. This version addresses and improves on many critiques and concerns of the original. The CLA+ takes a holistic, criterion sampling approach to measuring transferable learning outcomes using both a performance task and a series of selected response questions. The organization of the CLA+ and the specific skills associated with each is illustrated in Figure 2. The CLA+ is a web-based instrument with students writing the test in a secure browser. The test takes 90 minutes to complete, with 60 minutes allotted to the performance task and 30 minutes allotted to the selected response questions.

The core of the CLA+, the performance task, is used to measure critical thinking, problem solving and written communication and remains the key focus of the CLA+. The performance tasks present a real-world situation in which students assume an appropriate role to address the problem, identify a solution, or provide conclusions and recommendations resulting from careful analysis of the provided evidence. The student responses in the performance task are scored by an automated system using a validated rubric.
The selected response questions are document-based multiple choice or short answer questions, and are used to measure scientific and quantitative reasoning, critical reading and evaluation, and critiquing an argument. These items are scored in an objective fashion, with students selecting their response to the question or problem resulting from careful analysis of the provided supporting documents. At current, there is not a rubric associated with the selected response questions, only a generated score.

One of the key shortcomings of the original CLA was the resultant data could only be used at an institution level, with the test providing poor measures of reliability at the student level. The CAE addressed this by adding the selected response questions to improve the precision and reliability of the student level results. Construct and face validity of the CLA+ was measured, and construct validity was inherited from the original CLA, which was well correlated with two other measures of critical thinking: the Collegiate Assessment of Academic Proficiency, and the Measure of Academic Proficiency and Progress (CLA Performance Task $r=0.73-0.83$, CLA Critique an Argument $r=0.73-0.94$). Face validity was established through a separate pilot study, conducted by the CAE which had students self-report on what skills they believed the test measured. Students responded positively that the CLA+ is an effective measure of critical thinking and problem solving (86.2%), reading comprehension (88%) and writing (75.8%). Using a stratified alpha measure for reliability the two forms of the CLA+ have measures of 0.85 and 0.87 respectively.

Instrument: VALUE rubrics

The VALUE Rubrics were developed by the American Association of Colleges and Universities to provide a valid assessment of learning in undergraduate education. These rubrics are broad,
discipline-neutral descriptions of selected essential learning outcomes of undergraduate education from the Liberal Education America’s Promise (LEAP) initiative. In each rubric, common themes were identified for each outcome and performance criteria were developed by panels of experts.

The efforts of the experts were focused on:

- Performance criteria focuses on positive demonstration of outcomes rather than what was lacking
- Performance criteria can be used to assess to non-traditional modes of artefacts demonstrating student learning
- Performance criteria are developed to assess summative displays of student learning rather than developmental or formative displays
- Performance criteria are phrased in a manner as to be easily understood by non-experts

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. While the performance criteria and levels represent a consensus of experts and can be used in their original form, the rubrics are purposely designed for modification to foster alignment between course, program or institutional outcomes and to reflect the specific context in which they are used.

In this project the original VALUE rubrics pertaining to critical thinking, problem solving and written communication were modified on a case-by-case preserve alignment between course, program and standardized measures of each specific outcome as well as reflect the disciplinary expectations in which they are applied. Initially, significant work was put forth in using the VALUE rubrics to describe general indicators for each specific outcome. These indicators were invaluable in communicating with course instructors as they allowed them to easily understand what we were attempting to measure, if it was developed/assessed in their course, and a suitable student artefact to assess with the VALUE rubrics. The indicators, for each general outcome, are shown below in Figure 3.

There has been considerable effort in establishing the validity and reliability of the VALUE rubrics. A diverse interdisciplinary panel of experts was used to gauge face and content validity of the rubrics, which agreed the rubrics were an effective and suitable measure of the underlying constructs. Reliability of the rubrics has also been investigated, resulting in a varying range of scores. However, establishing reliability is a very procedure dependent process and the studies that reported low reliability conducted their reliability training in a decentralized manner. This method is not conducive to establishing agreement between raters by impeding the discussion and debate that is associated with a typical norming and calibration process. Those with higher
reliability conducted their reliability training in a local environment that promotes interaction, discussion and debate between raters, leading to increased agreement and enhanced measures of reliability. 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.

Figure 3: Indicators scored using VALUE rubrics.

Instrument: Transferable Learning Orientations survey

The development of the Transferable Learning Orientation (TLO) survey resulted from piloting the Motivated Strategies for Learning Questionnaire. The pilot incorporated ten scales, with 58 items in total, self-reported on a 5-point likert type scale with questions such as, “In a class like this, I prefer course material that really challenges me so I can learn new things”, with responses ranging from “Not at all like me” to “Very true of me”. The scales included Intrinsic Goal Orientation, Extrinsic Goal Orientation, Control of Learning Beliefs, Self-Efficacy for Learning & Performance, and learning strategies scales with Rehearsal, Elaboration, Organization, Metacognitive Self-Regulation, Time/Study Environmental Management, and Help Seeking. In line with the desired measurement constructs for lifelong learning, the scales of Test Anxiety, Task Value, Peer Learning, and Effort Regulation were excluded. Also excluded was Critical Thinking because alternate instruments in the larger research project measured this dimension.
Analysis of findings from our MSLQ pilot found lower reliabilities than expected, and that students demonstrated minimal engagement, spending an average of only nine seconds per item to complete the survey.

The TLO was developed with the goal of providing reliable informative feedback to students and faculty about the development of approaches to and skills for lifelong learning. The process involved refinement of the MSLQ scales, and adaptation of the Lifelong Learning VALUE rubric, and incorporation of open-ended responses for each dimension aimed at increasing metacognitive engagement with the instrument. The TLO is a triangulated measure using a quantitative pairing comprising four scale items for each dimension, correlated to a holistic rubric self-rating, with an open-ended response used for the purpose of validation of final rating. The structure of the TLO is overviewed in Figure 4.

![Learning Orientations/ Self Regulation](image)

**Figure 4: Structure and factors of the Transferable Learning Orientation (TLO) Survey**

**Instructor Feedback**

Part of the larger project included feedback sessions with instructors to present the departmental results. Following the presentation, the instructors were paired to provide feedback to researchers about their perceived benefit of the information they received. The discussion was guided by the
question “so now that we have presented you with our findings, can you tell us which tool you feel has the most utility for you as an instructor? Two of the research team facilitated two focus groups, comprising four instructors from two departments. Each of the instructors taught a different course as part of the project, and none of the researchers participated in the focus groups.

Results

The consenting sample of first year engineering students (n= 569), all completed the TLO. A subset of those students (n= 250) sat the CLA+ test. Fifty-two first term assignments were rated using the VALUE rubrics, representing 150 (n = 150) students, and 24 second-term assignments were rated, representing 118 (n = 118) students. The fourth year sample (n=40) comprised students from courses in Civil (n=5), Geological (n=11), Mechanical (n=2), Physics (n=1) and multi-disciplinary Engineering (n=21).

CLA+ Results

The Engineering first year mean for the CLA+ Total was 1169, and the fourth year mean was 1254. Figure 5 shows significant gains in critical thinking, problem solving and written communication demonstrated between 1st and 4th year for each of the CLA+ sub-scores. The CLA+ institution mean scatter plot (see Figure 6) is based on the 2014 report distributed by CLA+ test providers, and displays de-identified mean scores 4th year seniors. The first and fourth year mean scores were overlaid to demonstrate what the CLA+ refer to as the “value added”, and also to locate where the Queen’s Engineering cohort sit in relationship to other 4-year institutions.

Figure 5: Bar graphs displaying comparative mean scores between Engineering 1st and 4th Year for CLA+ Performance Task and Selected Response sub-scores
Student reports from first and final year design courses were sampled and scored using VALUE rubrics. The first report, labeled Assignment 1 below, was submitted at the end of the first term in first year, and entailed students working in groups of two or three to create a model of heat transfer in a CPU heat sink using a numerical analysis program (MATLAB), and adapt the model to a profitable application. The task brief provided with the assignment closely aligned with the VALUE rubrics used; as such, students were able to demonstrate their ability across the breadth of the rubric dimensions. There were group and individual components of the report. The following criteria were evaluated based on the individual analyses: Influence of Context and Assumptions, Student’s Position, Conclusion and Related Outcomes, Evaluate Outcomes. The remainders were evaluated on the basis of the group, and every individual in the group was assigned the same levels. Each assignment was independently rated by two trained markers, and took an average of 120 minutes per report; with 52 reports, including calibration time, the total combined time commitment was approximately 212 hours. Inter-rater agreement was very high, the average Kappa statistic for exact agreement between two independent raters, across 16 separate dimensions on the three rubrics was 90%. Students demonstrated between benchmark 1 and milestone 3 for all criteria across the rubrics (see Figure 7).
The second report, labeled Assignment 2 below was submitted at the end of the winter term, and focused on engineering design. The reports required a problem statements and scope definition, background information, conceptual design solutions, decision making process for choosing the final solution, the design proposal, implementation (for example building a prototype), economic analysis, evaluation based on the objectives in the problem statement, and conclusions, which included recommendation for continuing the work and possible ways to improve the outcomes. Every individual in the group received the same grade on each dimension because this was a group report with no individual contributions indicated. Each assignment took an average of 105 minutes each to mark; 24 reports, each marked twice, and including calibration time, the total combined time commitment was approximately 87 hours. Kappa statistic was even higher than for assignment 1, with 94% exact agreement between two independent raters. In these assignments, groups received support from faculty and the project manager, which aided students in producing high quality of the reports. Students again demonstrated between benchmark 1 and milestone 3 for all criteria across the rubrics.

In three of the four engineering fourth year courses, the artefact assessed was a team-based, client driven design project. Also included in the sample was select assessment dimensions rated on a problem task posed in an exam setting of a technical course. Subject area graduate students, and research associates rated the fourth year course artefacts. There was a subset (between 5-10% depending on the number in the sample) of artefacts marked by both undergrad and graduate markers to ensure vertical consistency between samples. The exact agreement between the two raters ranged from 48% to 78%. None of the paired ratings for any of the samples differed by more than one level on the rubric. For each of the course artefacts, the raters scores were averaged to form the final rating per sample, dimension. The comparative means for the first and fourth year course-based artefacts, for each of the rubric dimensions, are presented in Figure 7, with standard deviations displayed.

Principal axis factoring was used to confirm a single factor solution for each of the constructs of problem solving, critical thinking, and written communication. Based on each of the dimensions within the rubric, their single factors accounted for 70%, 76%, and 78% of the variance respectively. Each of the factors was found to be highly interrelated, with the same coefficient for Pearson’s correlations between problem solving, critical thinking, and written communication of $r(315) = .91$, $p < .01$. Based on these correlations, the scores for problem solving, critical thinking and written communication were aggregated for comparison between assessment tools.
Figure 7: VALUE rubric mean scores for 1st year (n=150) and 4th year (n=23) cross sectional samples.
TLO Results

The results here represent a beta run of a new instrument. Continuing research has since refined the tool for the purpose of facilitating awareness of learning orientations and self-regulation, and trigger meta-cognitive processes in our students.

The scale reliability for the beta version were calculated using Cronbach's alpha, which was Motivation $\alpha = .455$, Learning Belief $\alpha = .414$, Self efficacy $\alpha = .717$, Transfer $\alpha = .446$, Self-regulation $\alpha = .501$, and Organization $\alpha = .530$. The scale scores were converted to a four-point scale and averaged with the rubric rating to form a dimension score. The self-rated scores represent a continuum from one to four. A score of one describes a surface level engagement, fixed mindset, limited confidence, minimal self-monitoring and lack of organization, whereas four describes a deep learner, flexible mindset, confident, self-control and high level of organization. For the first year Engineering sample, the dimension means ranged between 2.4 (SD .57) for organization and 2.9 (SD .51) for learning belief.

Pearson’s $r$ correlations were calculated using the CLA+ total score, the aggregate VALUE rubric scores for the first year course assignments, and dimensions of the TLO (see Figure 8). The relationship between CLA+ and VALUE rubric scores, and self-reported learning orientations provide an indication of the effect of attitudes and behaviours on critical thinking, problem solving, and communication. Results show significant moderate correlations between learning belief, and self-efficacy, and CLA+ scores, suggesting that students with a more flexible mindset and greater confidence in their ability in Engineering perform better in critical thinking, problem solving and written communication. In addition to this, correlations between the CLA+ and VALUE rubric assessment of Assignments 1 and 2 supports content validity of the instruments.

Cost Benefit Analysis

The costs for implementation of, and perceived benefits for the CLA+ and VALUE rubric data were analyzed. As we have yet to fully finish the development of the TLO, we have yet to compile a definitive cost of all development and implementation. The costs were calculated using 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+ $n= 598$, for VALUE rubric marking, $n= 621$. The CLA+ costs included the test provider fee, and proctoring; the VALUE rubric marking costs included marker training (salaries), and salaries for markers (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
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 4th year samples were marked by graduate students. The average cost per valid n for the VALUE rubric marking was $14.82. Figure 9 shows a comparison of cost per valid n for the CLA+ and VALUE rubrics.

Figure 8: Correlations between TLO, CLA+ and VALUE Rubric ratings. Coefficients calculated at the 95% confidence level using Pearson’s r.
Instructor feedback

Instructor debriefing sessions were audio recorded, transcribed and analyzed to provide insight into the utility of the CLA+ and VALUE tools. The goal was to determine the perceived benefit of these tools from the viewpoint of the instructors, to improve student learning. As part of the first phase in the study, researchers employed inductive content analysis to group the comments into common themes. The themes that emerged from the sessions were: *Ease of logistics; Confidence in the reliability and validity; Alignment to the course; and Applicability to affect course improvement*. Figure 10 below is a summary of feedback from course instructors, which is explained in more detail in the subsections below.

*Ease of Logistics*

The CLA+ presents logistical issues due to the minimum 90 minute time requirement, the need for a computer and wireless connection. In addition to this, the test takers need to turn up at a particular time, in a particular place to do a test that isn’t directly connected to coursework. This results in low response rates unless course instructors encourage participation. In contrast, the VALUE rubrics are applied to a course artefact, so present no additional work for either the student or instructor.
### Figure 10: Summarized feedback from instructor focus groups about the utility of the CLA+ and VALUE tools

<table>
<thead>
<tr>
<th></th>
<th>CLA+</th>
<th>VALUE rubrics</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Variable test conditions</td>
<td>Interesting</td>
<td>Need for training</td>
</tr>
<tr>
<td>Divorced activity</td>
<td>Face validity</td>
<td>Face validity</td>
</tr>
<tr>
<td>Red flag</td>
<td></td>
<td>High inter-rater reliability</td>
</tr>
<tr>
<td>Lack of transparency</td>
<td></td>
<td>Useful for aligning tasks to assessment criteria</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Helped clarify thinking</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Useful information</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Really effective for course improvement</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Assessment for learning</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Useful for instructional design</td>
</tr>
</tbody>
</table>

**Reliability and Validity**

Instructors were wary of standardized testing, which they felt involved paying a testing service to tell you how good your students are at doing something. There was concern about external validity, and the need for an institution to prove itself in wider circles. Concern about the construct and criteria, i.e. what the test or rubric measure, was a lesser concern for instructors, but the uncertainty in terms of content was apparent.

Instructors were provided with empirical evidence to support the reliability of the CLA+. They were impressed with the inter-rater reliability of the VALUE rubric marking completed as part of the research, but expressed concern about the possibility of doing the ratings themselves.

**Alignment to the course**

The CLA+ was seen to provide interesting overarching feedback about student ability, but no particular connection to their course. In contrast, the VALUE rubrics functioned as an effective tool for instructors to aid in designing assessments. Through feedback, course instructors became
aware of the problem of the course artefact not giving students the opportunity to demonstrate particular criteria on the VALUE rubric.

**Applicability to affect course improvement**

Instructors found results from the CLA interesting, but did not find very much information from the CLA+ assessment they considered effective for course improvement, whereas they suggested that the VALUE rubric results were diagnostic for the particular task evaluated.

**Conclusions**

The results of the first year of the study are showing some interesting results, including correlations between measures, discrimination of instruments, and differences in response rates and resources to implement. Some of the preliminary conclusions are:

- There are statistically significant correlations between critical thinking and communication dimensions measured on the CLA+ and the VALUE rubrics, and correlations with measures of learning belief and self-efficacy.
- CLA+ and VALUE rubrics scores show significant increases in critical thinking, problem solving, and communication between the first year and fourth year samples.
- Participation rates between among fourth year students on the CLA+ are much lower than first year rates, primarily because the first year testing was conducted in scheduled course, which was not possible for the fourth year students. This problem has been identified in prior CLA studies, and reflects the significant disadvantage of standardized tools compared to using VALUE rubrics and in-class artefacts.
- The cost of implementing the VALUE rubric testing was approximately \( \frac{1}{3} \) that of the CLA+ testing.
- The CLA+ results provide an ability to compare institutional performance with other schools. This is not possible with VALUE rubric scoring without common training and calibration procedures.

It is valuable to know how well, and where, students are developing transferable intellectual skills including critical thinking, problem solving and written communication. The first phase has identified that both tools are measuring significant gains, and significant correlations between instruments. Instructors feel that the alignment of teaching, learning and assessment is necessary to enable the measurement of cognitive skills within these courses. The most successful strategy has been to facilitate student’s engagement with tasks that specifically encourage critical thinking and problem solving. Focus groups with instructors showed that they felt that the VALUE rubrics could be more effective in the short term, but there may be benefits in the CLA+ for longer-term evaluation and comparison with other institutions.
Additionally, this study has been a valuable catalyst for change, and helps to build capacity across a department by engaging interested instructors in all four years of a program.

References

2. Rhodes T, Finley A. Using the VALUE Rubrics for Improvement of Learning and Authentic Assessment. AAC&U; 2013.


