Higher Education Quality Council of Ontario

An agency of the Government of Ontario

Scaling the Development and Measurement of Transferable Skills: Assessing the Potential of Rubric Scoring in the Context of Peer Assessment

Steve Joordens, Dwayne Paré, Robin Walker
University of Toronto Scarborough;
Jim Hewitt and Clare Brett
Ontario Institute for Studies in Education

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:

Joordens, S., Paré, D., Walker, R., Hewitt, J. & Brett, C. (2019). *Scaling the Development and Measurement of Transferable Skills: Assessing the Potential of Rubric Scoring in the Context of Peer Assessment.* Toronto: Higher Education Quality Council of Ontario.

Table of Contents

List of Tables

List of Figures

Acknowledgements

The authors wish to thank the Higher Education Quality Council of Ontario (HEQCO) for its support of this research.

Executive Summary

Postsecondary institutions are responsible not only for teaching their students the subject-specific knowledge related to their field of study, but also providing them with transferable skills. These skills can transfer well from learning to career, and from job to job; skills like critical or creative thinking, and clear effective communication. While the subject-specific knowledge is taught formally, the teaching of skills often is not.

In an attempt to provide solutions that would allow educators to both develop and measure core transferable skills, Joordens (2018) argued for a peer-assessment approach; specifically one that involved the use of rubrics to evaluate work amongst peers. The proposed approach rests on the assumption that if students apply a skills-based rubric to the work of their peers, and each student has their work assessed with the rubric by multiple peers, the average of a number of peer ratings would provide a good measure of the skill the rubric was created to assess.

Joordens's proposed approach can answer two questions related to developing core transferable skills: How do we develop these skills? And how do we know these skills are being developed? The act of assessing the presence of a transferable skill such as critical thinking in the work of multiple peers in itself can be part of improving that skill in students. Once a student uses a rubric to assess a peer's work, they can carry that knowledge into their own work, a concept that can be applied across a wide range of learning contexts.

Phase 1

In the first phase of this study we test that assumption directly with the key element being the incorporation of a validated rubric into the assessment phase, in this case the VALUE critical thinking rubric developed by the Association of American Colleges and Universities (AAC&U). The results show that while measures based on peer ratings do capture the rank ordering of quality, they also have a tendency to overestimate the actual amount of the skill present at least when the students are first-year students. This was determined in comparison to the average ratings given by three groups: TAs, expert assessors and the professor of the course.

These findings do not represent a major problem for the proposed approach, but they do lead to suggestions for student training that could help make the approach more optimal. The first phase also provides qualitative data showing that students understand and appreciate the value of developing and measuring these skills, and that they would willingly do the additional work involved in peer assessment if this were possible.

Four different sets of participants contributed data:

- 694 students (recruited from 1,800 students enrolled in an Introduction to Psychology class at UTSC)
- Two TAs (two selected from a pool of current and previous TAs for the Intro to Psych class)
- Two expert assessors (recommended by AAC&U, trained in the use of VALUE rubrics)
- A course professor (no previous experience or training using VALUE rubric)

peerScholar

The procedure used in this study was managed by a cloud-based program called peerScholar, which was developed at the University of Toronto Scarborough (UTSC). It combines the process of peer assessment with self-assessment and feedback. The peerScholar program includes three phases: Create, Assess and Reflect

Create Phase: In this phase, students submit an argument piece that's been researched and is supported by evidence.

Assess Phase: Each student's work is assessed by six of their peers, and each student is required to assess the work of six peers and provide feedback (in this study, after watching an instructional video on how to apply the rubric)

Reflect Phase: Students see feedback on their own work and are asked to rate the quality of the feedback they received. Then students have the option to revise their own work and write a short reflection piece justifying their revisions.

Phase 2

Phase 2 of the project built on the use of peer assessment in Phase 1 by looking at the scalability of the method and asking if having students apply rubrics as they conduct peer assessment could work across a wide range of educational contexts, levels, institutions and even countries.

The research team reached out to professors through a number of different methods and initially were able to attract 53 professors. However, as the project progressed, it became challenging to move these commitments forward past the initial expression of interest and the recruitment process had to adapt in order for the project to proceed. Eventually, a group of 14 professors agreed to participate in Phase 2. The group comprised professors from U of T (including the Ontario Institute for Studies in Education), Lakehead University, Humber College, Assiniboine Community College, Urbana University (US) and Utrecht University (the Netherlands). The 14 professors recruited used the method in 20 different courses and surveyed students' opinions and reactions.

Qualitative data from students replicated and extended the results from Phase 1, showing that students like and value the approach. This data is further complemented by faculty data attesting to the usability of the approach as being non-disruptive and easy to implement, and their perceptions that the approach had a positive impact on learning and should be used more widely. Taken together, the findings bolster the potential of this approach as a way to both develop and measure virtually any transferable skill in virtually any educational context.

Discussion and Findings

The validity of peer-based rubric scores remains something of a question. Despite receiving brief instruction on how to apply the rubric in the assessment process, students still marked higher than the TAs, professor and expert assessors in this study. One potential solution to this problem would be to provide more training using rubrics to students. Another possible solution would be to measure the extent to which a bias is present and then statistically control for it.

Other challenges that may need to be overcome in order for this approach to be more widely adopted include enticing faculty to try a method that may be unfamiliar to them; making the process easier to use for both faculty and students; and convincing students that the perceived "extra work" involved in this process is worth it.

Conclusions

By and large, the faculty involved in Phase 2 of this project were very positive in their assessment of the approach's potential. Perhaps most relevant, virtually all agreed they would continue to use the approach and that it should be used more widely. It is clear then that the proposed approach absolutely has high potential despite the tweaking needed to ensure an appropriate level of validity.

Glossary of Terms

"the initial 53" – This term refers to a group of 53 faculty members (mostly from Canada with a few from Australia, the UK and Africa) who filled out a Google form indicating some degree of interest in participating in a peer assessment project. As the project progressed, most from this group did not proceed but the data they provided through the Google form did provide some insight into the use of peer assessment.

"the participating 14" – This term refers to the group who participated in the study to assess scalability of the peer assessment approach. The group was mostly made up of other University of Toronto professors in the hopes of maintaining a high level of involvement in the project.

Assess Phase – The Assess Phase involved students logging into the peerScholar platform to complete a round of peer and self-assessment. Specifically, they were given access to compositions submitted by six of their classmates and they were expected to assess and provide constructive feedback to each composition. The compositions were selected randomly from the pool of submissions and were presented anonymously.

Create Phase – The Create Phase involved students submitting an argument piece related to a query. Specifically, they were instructed to find and read at least three reliable sources related to the topic and then compose a persuasive argument related to a sub-issue of the larger questions.

creative thinking (creative thought) – Creative thinking is both the capacity to combine or synthesize existing ideas, images or expertise in original ways, and the experience of thinking, reacting and working in an imaginative way characterized by a high degree of innovation, divergent thinking and risk taking.

critical thinking – Critical thinking (or critical thought) is a habit of mind characterized by the comprehensive exploration of issues, ideas, artifacts and events before accepting or formulating an opinion or conclusion.

expert assessors – In this paper, the term expert assessors refers to assessors recommended by and facilitated through the AAC&U. They have undergone rigorous training on the application of the Critical Thinking VALUE rubric to student work.

peerScholar – peerScholar is a learning management platform that combines peer assessment with two other evidence-based educational practices: self-assessments and the formative use of feedback. peerScholar is a cloud-based platform and its process of peer assessment is made up of three phases: the Create Phase, the Assess Phase and the Reflect Phase.

Phase 1 – Phase 1 primarily focused on whether averages derived from peer-provided rubric scores could be used to predict expert ratings (i.e., those provided by TAs and professors).

Phase 2 – Phase 2 represents an attempt to directly assess previously made claims of scalability, primarily by asking faculty to try peer assessment in their class, whatever their class may be and at whatever level it is taught. Phase 2 also includes student and faculty questionnaires that were completed after the peer assessment exercise. Submissions were submitted through the peerScholar platform.

Reflect Phase – The Reflect Phase involved students being given access to the feedback they had received from six of the peers on their own composition. Then they were asked to rate the feedback on the basis of its usefulness, and on its tone of presentation. After rating the feedback, they were given the opportunity to revise their composition based on the feedback they received. Once revisions had been made, students wrote a short reflection piece wherein they justified the revisions they made as well as the suggested revisions they chose not to implement.

VALUE rubrics –The Association of American Colleges and Universities (AAC&U) developed 16 rubrics related to different transferable skills. They have all have been shown to have acceptable levels of reliability along with clear face validity and are publicly available.

Introduction

Educators and educational institutions have two distinct responsibilities to their students. First, they must provide the information that is critical to a student's field of study. In Psychology, for example, we are expected to teach the primary theories and experiments of the field, the important figures associated with them and the current state of "knowledge." Second, we must give our students repeated practice using core transferable skills, skills like critical thinking, creative thinking, effective communication, and the metacognitive awareness they need to understand their strengths and weaknesses at a given time.

As university educators we do a very good job teaching information, partly because we know how, partly because we can measure success objectively, but mostly because it's relatively easy to do within the constraints of our institutions. When it comes to teaching skills in a formal way, things become more difficult. We have no way of measuring skill development, and we cannot currently see how to do either well given the time and resource constraints of our institutions. As a result, we formally teach and measure information, and take a much less formal approach to the development of skills. That is, we rely on great lectures or readings to stimulate critical or creative thought, and we give students assignments that we hope will help them develop communication skills. We assess performance on those assignments, but we don't formally recognize those grades as reflecting skill development. Rather, they are typically added into a course grade that is primarily influenced by a student's ability to retain information.

Recently Joordens (2018) suggested a process that could provide a solution to both the "how?" and the "how do we know?" issues, and can do so without requiring much alteration of an instructor's typical approach to teaching their class. By "how?" we are referring to the question of how best to develop core transferable skills like critical thinking. By "how do we know?" we are referring to the need to evaluate these skills so we know they are being developed, and so we can further hone the development process in an evidence-based manner. The approach suggested by Joordens (2018) combines technology-enabled peer assessment with the use of rubrics specifically created to assess core transferable skills.

The next section describes the proposed process in some detail, highlighting how it provides a powerful learning experience, while also allowing skills to be quantified directly in the context of in-class activities. We then move to the primary focus of the research reported here which is the following: The proposed approach assumes that peer assessment provides a valid measure of transferable skills. More specifically, an average of peer assessment scores derived from having students apply rubrics related to transferable skills to the work of their peers accurately measures the skills in question; is this in fact the case? A number of studies have demonstrated that averaging peer grades yields a statistic that approximates grades given by profs (Cho, Schunn & Wilson, 2006; Paré & Joordens, 2008), but does this also hold for rubric-based scores?

Combining Peer Assessment and Validated Rubrics

There is an ever-growing body of literature showing that asking students to assess the work of their peers (peer assessment) is an extremely powerful pedagogical approach in many ways (Dochy, Segers & Sluijsmans, D., 1999; Liu & Carless, 2006; Topping, 1998). Perhaps most relevant to this work, it is one of the few educational processes that has an experimental evidence base supporting its ability to develop critical thinking and enhance metacognitive awareness (Joordens, Paré & Collimore, 2014). Specifically, Joordens et al. (2014) provide data showing that with each experience reviewing the work of six peers, student ratings of quality come to more closely match those provided by experts and those based on peer averages. Joordens et al. attribute this to enhancements in "quality-based discrimination," which they claim is at the heart of critical thinking. In addition, they demonstrated that students' assessment of the quality of their own work became significantly more accurate following just one analysis of the work of six of their peers. Given that peer assessment requires students to expend significant time and effort reviewing peer work, it is also relevant that it provides an opportunity for the deep learning that students value and enjoy (Collimore, Paré & Joordens, 2015; Hanrahan & Isaacs, 2001).

There is an ever-growing list of peer-assessment technologies that can allow instructors to add peer assessment steps to any assignment given to students. We have developed one such system within our lab, a system called peerScholar (see vision.peerScholar.com for detailed information) that combines peer assessment with two other evidence-based educational practices; self-assessments (Boud, 2013) and the formative use of feedback (Bennett, 2011; Nicol & Macfarlane-Dick, 2006). Given we know it best, and that it will be the system that will be used in the current research, our presentation here will relate to the process it was designed to support. That said, the following core idea could likely be implemented in many technology-enabled peer-assessment solutions.

Within peerScholar, students complete work across three phases. The first phase, "Create," is one in which students simply submit some digital composition in response to a prompt and materials provided by their instructor. This part might align with an assignment the instructor already uses in their course. The peer assessment occurs in the second phase, "Assess." Within this phase students see a set of anonymously presented, randomly selected compositions submitted by their peers. In our usage, they typically see six peer compositions, and they are asked to complete an assessment exercise in accordance with questions or rubrics that the instructor has defined within the tool. This phase will be discussed later in more detail; for now, as a given student is assessing their peers, the same number of peers will assess their work, and in the third phase, "Reflect," they see the feedback applied to their work, rate that feedback and are then allowed to revise their work for final submission by incorporating the feedback they found useful.

A detailed pedagogical discussion of this process and how it supports the development of core transferable skills is provided by Joordens, Paré and Collimore (2014). Briefly, the process requires students to repeatedly exercise critical thinking, creative thinking, expressive communication and receptive communication, and does so in a feedback-rich context that enhances metacognitive awareness. In addition to a pedagogical breakdown, the report also details experiments that demonstrated that the described process has significant positive effects on measures argued to reflect critical thinking and metacognitive awareness as described earlier.

For current purposes the focus will be on a potential augmentation of the process or, more specifically, a principled inclusion of validated rubrics within the Assess phase. Thus, the central idea is that as students are assessing the work of their peers, one of the things they do is score the work using a validated rubric associated with the transferable skill that is intended to be measured. Given that each composition is assessed by six peers, the average of these scores — hereafter described as the "peer average" will provide a measure of that skill. With respect to the use of six peers specifically, previous studies have shown that peer averages of overall quality ratings are valid and reliable when based on as few as four peers in one report (Cho et. al, 2006), and as few as five peers as reported in another (Paré & Joordens, 2008). Thus six is a preferred number when relying on peer averages as a slight over-delivery of what may be necessary.

When we refer to validated rubrics, one clear example is the VALUE rubrics that have been developed by the Association of American Colleges and Universities (Rhodes, 2016). It has developed rubrics related to 16 different transferable skills, and all of these rubrics have been shown to have acceptable levels of reliability along with clear face validity (Finley, 2011). If a given instructor was interested in developing, say, creative thinking in her students then she would make sure students understood that demonstrating creative thought in their compositions is the core goal, and when they assessed one another's work they could use the Creative Thinking VALUE Rubric to quantify how much creative thought is evident in a given composition.

This approach is much more than a means to an end. In addition to providing measures of transferable skills, it may also provide an enhanced way of teaching them. In the realm of music, students quickly become aware that a critical step in being a better musical performer is to "develop one's ear" (Woody, 2012). That is, they must learn how to more accurately listen to, and appreciate, the subtle but important things other musicians do when they play. They must move beyond "that sounds nice" or "that makes me feel happy" in order to get to "the use of silence is interesting" or "I'm intrigued by how they create phrases and/or how they use notes outside of those expected by the scale." Once they hear this in the performance of others, and feel the effect those techniques have, they have a clearer sense of how they could add those elements to their own performance.

Applying rubrics associated with transferable skills is, in a sense, a method for "developing one's ear" for that skill. Rather than trying to produce critical thinking, for example, students are being asked to detect it in the work of others, to find cases where a student is doing something well, to recognize it and to value the effect it causes. Then, just like the musician, they are in a better position to use those techniques in their own work.

Another way of thinking of this is opening the students up to the ability to switch perspectives between performer and audience. Every good performer is able to leave their body, and consider their performance from the perspective of the audience. Typically in universities we primarily ask students to perform, or at least to compose work. We seldom ask them to assess work, let alone do so in the sort of structured manner that applying rubrics provides. Once the student is able to see compositions from an assessment perspective, they can carry that with them when they compose, and that should make them better composers.

Critical to scaling, this process of applying rubrics is what is often described as content agnostic. Any attempt to formalize and measure skill learning in a way that will allow it to compete with information learning must describe a process that could be broadly applied across a wide range of learning contexts. Given that the instructor determines what sort of compositions students initially write, those compositions can be in any area of study and students can be at any level of study. For example, at the University of Toronto technology-enabled peer assessment is now used in over 50 courses per term that range from our first-year Introduction to Psychology course to much smaller graduate-level courses in medicine. Further, given that a rubric can be tailored to any transferable skill — for example the AAC&U rubrics cover 16 possible transferable skills — the focus of the assignment is also very free to vary.

On the face of it then, the proposed process appears to have several benefits. It provides students with structured practice assessing transferable skills, which may be especially helpful given the addition of applying rubrics to peer compositions. It does so in a way that encourages deep processing and thus powerful learning. It also provides a student-by-student measure of the strength of a given skill as apparent in that student's work. Finally, it can be applied in virtually any class context irrespective of level, subject area or method of course delivery.

While the potential of combining technology-enabled peer assessment with the principled use of validated rubrics sounds impressive, clearly it could benefit from a direct assessment of that potential. The study is designed to provide such an assessment and, across the two phases, it does so by assessing two assumptions upon which this potential rests; the first is the assumption that averages based on peer-

provided rubric ratings are valid, the second is that the process is flexible enough to support wide-ranging application across the higher education ecosystem.

Phase 1

The proposed process assumes that the measure provided by averaging peer ratings provides a valid measure of the skill in question. Previous research has shown that when the ratings of at least four or five peers are combined, they provide a good predictor of the ratings that would be provided by teaching assistants, ratings that are comparable to those provided by a different teaching assistant (Cho, Schunn & Wilson, 2006; Paré & Joordens, 2008). Another study of high school students showed that peer ratings also predicted expert scores at a high level (Tseng & Tsai, 2007). However, those previous studies only asked students to provide overall ratings of quality; they did not specifically assess the validity of rubric-based ratings, let alone rubrics focusing on the somewhat complex transferable skills we are primarily interested in here.

Thus the primary focus of Phase 1 is to directly assess the validity of transferable skill estimates based on peer averages as students apply the Critical Thinking VALUE rubric to the work of their peers. These estimates based on peer averages will be compared to estimates provided by teaching assistants and by experts trained by AAC&U to apply the VALUE rubrics. When assessing validity we examine both agreement in terms of the rank ordering of quality (i.e., Pearson correlations) and in terms of actual agreement of scores (i.e., kappa measures), and we compare the agreement of peer averages with teaching assistants and experts with agreement levels within the two expert groups. Our findings suggest that while peer averages provide good predictors of the grades given by both teaching assistants and trained experts, they do not do as well in terms of actual agreement, primarily because peer averages tend to be somewhat inflated relative to the ratings provided by the other two groups.

In addition, we also collected some qualitative data related to peer assessment via questionnaire data. Specifically, immediately after completing the activity, students were asked a number of questions related to the proposed imbalance between information training and skills training as well as questions related to the process they just participated in as a method for addressing this imbalance (see Appendix B). The need to provide some context made these questions somewhat more complex than typical questionnaire questions. For example, one reads: “In most course you are taught a lot of information about some subject area, but it’s also important that we develop your so-called transferable skills, skills like critical thinking, creative thinking, or clear effective communication. To what extent would you agree with this statement: ‘I feel that typical university courses challenge me to think critically and creatively, and they give me experience communicating my ideas effectively.’” To our knowledge no one has previously tried to assess

the student perspective on these issues and, given that we were throwing our students into the thick of new approaches to resolve them, we thought their perspective would be interesting and potentially useful.

Method

Participants

Four distinct sets of participants contributed data to this study.

Students – Student participants for this study were recruited from the 1,800 students enrolled in the Introduction to Psychology class offered at the University of Toronto Scarborough. All students completed the technology-enabled peer assessment activity as part of their course. Immediately after completing the peer-assessment phase of the online activity, an on-screen query appeared requesting consent to use their data for research. The query emphasized that their participation was voluntary and that not only was their participation in the study independent of the course, but also that the course instructor would not be informed or aware of whether or not any given student consented to participate. They chose one of two options via a pulldown menu: either "I consent to allow my data to be used for research purposes" or "I would prefer that my data NOT be used for research purposes." Of the 1,683 students who completed the activity, 694 participants consented to allow us to use their data, an acceptance rate of 41%.

Teaching Assistants (TAs) – Two experienced teaching assistants were selected from the pool of current and previous assistants for the introductory psychology course and asked to rate student work using the Critical Thinking VALUE rubric. These TAs were selected for their noted diligence and high-quality work in previous assessments of student performance. They received hourly compensation for their participation in this study.

Expert Assessors – Two expert assessors as recommended by and facilitated through the AAC&U were hired to participate in this study. These experts have undergone rigorous training on the application of the Critical Thinking VALUE rubric to student work.

Course Professor – In addition to the groups or pairs of assessors above, it was decided that it would be interesting to have the professor of the course also apply the Critical Thinking VALUE rubric in this study. Despite being a researcher active in the study of critical thinking, the professor had no previous experience or training applying the VALUE rubrics.

Procedure

Student data was collected in the context of an online peer-assessment activity supported by peerScholar, a cloud-based educational technology. This activity was worth 12% of the course grade, and the composition step required students to take a position on a relevant debate and argue their position well, showing as much evidence of critical thinking as possible as they did so.

Each peerScholar activity consists of three phases: a Create Phase, an Assess Phase and a Reflect Phase. Each phase unfolded as follows.

The Create Phase – Each student was asked to submit a one to two page (double-spaced) argument piece related to the query "Is the dual standard across animal versus human research ethics justifiable?" Specifically, they were instructed to find and read at least three reliable sources related to this topic, and then to eventually compose a persuasive argument related to a sub-issue of this larger question. For example, a student might ultimately argue that the dual standard is justifiable because animals have no self-awareness. In so doing they would be expected to back up their argument with explicit research and convincing arguments. Once constructed, they would log into peerScholar and submit their composition prior to the deadline for this phase.

The Assess Phase – After all students had submitted their compositions, they logged in again to complete a round of peer and self-assessment. Specifically, they were given access to compositions submitted by six of their classmates and they were expected to assess and provide constructive feedback to each composition. The compositions were selected randomly from the pool of submissions and were presented anonymously. The specific assessments they were asked to perform were determined by the instructor. In this case they consisted of the following: (a) students could attach in-line comments directly to the peer work to indicate lower-level writing issues like grammar corrections, citation corrections, etc., (b) students were asked to fill out a matrix that mapped onto the Critical Thinking VALUE rubric (see Appendix A), and (c) students were asked to give each peer some constructive advice by highlighting the one thing that — if improved — would most improve the overall work, and they were to give the peer some sense of how they might improve it. For current purposes, the VALUE matrix ratings are the most critical.

Prior to performing their peer assessments using the rubric, students were required to watch a 24-minute video (http://heqco.s3-website.ca-central-1.amazonaws.com) that instructed them on how best to apply the rubric. Following the video they were asked five questions related to the content, and were not allowed to proceed to the Assess Phase until they had answered all five questions correctly. This step represented an attempt to at least somewhat calibrate our students in terms of how they applied the rubric.

The Reflect Phase – As each student was providing feedback to six peers during the Assess Phase, six peers were also providing feedback to each given student's original composition. In the Reflect Phase a number of things happen. First, students are given access to the full set of feedback provided by their peers. Then, one by one, they are asked to rate the feedback on the basis of its usefulness and on its tone of presentation. Once they have rated each piece of feedback, they are then given the opportunity to revise their composition for final submission based on the useful feedback they received. Once they have revised their submission, they wrote a short reflection piece wherein they justified the revisions they made as well as the revisions suggested by peers that they chose not to implement.

Students were ultimately graded based on four components. Their draft composition was worth 2%, their final composition was worth 4%, the quality of their feedback to peers was worth 3%, and their ability to justify the revisions they did or did not make in light of the peer feedback they received was worth 3%. All components were graded by TAs upon completion of the activity.

As a final step at the end of the Reflect Phase, students were given a set of questions (see Appendix B) within the same assignment context. The questions they were asked were meant to capture their opinions around (a) the importance of exercising and measuring transferable skills, and (b) the extent to which they endorsed the specific approach being investigated in this work.

Apparatus, Materials and Context

The activity central to this study was one of several online activities that are a normal part of the Introduction to Psychology class. All are described in the course syllabus, with details regarding due dates and system use also highlighted therein. The application is very intuitive, and no student indicated any issue navigating it as intended.

The collection of student data (i.e., the procedure described above) occurred online and was managed by the peerScholar system. Each phase had a specific opening and closing date. Students were free to complete each phase of the activity across multiple sittings using whichever device and browser they preferred. As the activity was online, they could complete it from home, school or wherever they chose.

One of the critical materials used in this study is a rubric designed to assess critical thinking skills: the Critical Thinking VALUE rubric (see Appendix A). It is the result of significant research by the Association of American Colleges and Universities (AAC&U) to maximize reliability, validity and applicability to general contexts (Finley, 2011). The AAC&U also offers training programs that faculty members can undergo to increase their expertise and consistency in terms of applying the rubrics. Our students had no such training, but as noted

above we did produce an online module to roughly calibrate them, and that module can be accessed online (http://heqco.s3-website.ca-central-1.amazonaws.com).

We also created a questionnaire specifically for this study. Discussions focused on identifying core transferable skills and methods for developing them rarely include a strong student voice. Given the truly unique nature of the process we are evaluating here, it seemed wise to ask students for their opinions both on the specific process, and on the larger goals of developing transferable skills. No such questionnaire existed so we created one ourselves; it is presented in Appendix B. At this point, the goal of this questionnaire is simply to gain an initial glimpse into the student perspective.

TA and expert data was collected using an online tool we built, called Rubricaun, to optimize the process of applying the rubric to student work. This tool presents users with the work to be assessed and an interactive form of the rubric (see Appendix C). This process greatly reduced the amount of time and effort required to assess large collections of documents (compared to assessing documents with paper and/or spreadsheets). Additionally, TAs watched the same training video that students watched prior to completing the assessments.

TAs and experts assessed the same randomly selected set of 200 compositions drawn from the pool of 665 compositions: the 694 students who consented to participate excluding any students who did not perform six peer assessments or who were not assessed by six peers. To be clear, if a student fails to complete the Assess Phase then up to six students end up receiving one less review than would otherwise have been the case. This happens rarely, especially because students receive a zero on the activity when this happens without cause. In the current activity, 29 of the 694 students who consented (4.2%) did not have a complete set of reviews.

Results

We analyzed the results of this work in two sections. The first focuses on assessing the reliability of student-provided rubric ratings, and it includes data only from the 200 randomly selected compositions described in the previous section. The second focuses on the questionnaire findings and includes data from the entire fall 2017 offering of UTSC's Introduction to Psychology course.

Reliability Data

Ultimately, our central goal is to assess the reliability of rubric scores acquired by averaging scores provided by untrained undergraduates as they assess their peers. Recall that 200 compositions were randomly

selected from the available compositions as described previously and each selected composition was assessed by six peers, a pair of graduate level TAs and a pair of faculty-level experts who underwent training provided by the AAC&U. The professor of the course also scored all compositions simply to provide a sense of where a faculty member's ratings may fall, and how they may compare to others on the expertise continuum. By examining agreement levels between specific sets of assessors we can gain a clear sense of peer-based rubric assessment reliability.

As a first step, we examined the descriptive statistics related to the ratings associated with each group as presented in Table 1. The averages in Table 1 are based on the total rubric score associated with each composition. Each rubric score actually reflected five distinct factors of critical thinking, specifically explanation of issues, evidence, influence of context and assumptions, student's position, and conclusions and related outcomes. Each of these could earn a maximum score of 4. Thus the means reflect these five factor scores summed, then divided by 5 to produce an overall average score.

It should be noted that the AAC&U training allows for ratings of 0, whereas the raters in this study were required to provide a minimum rating of 1 (i.e., the lowest category on each rubric component). This was simply a result of confusion on our part as the rubrics themselves (see Appendix A) do not show a 0 rating and thus we assumed that the lowest ratings were intended to be 1. By the time we reached out to the experts and learned of our mistaken assumption, our students had already been through the process without the zero option. Thus we required all other scorers to proceed in the same manner as our students in order to keep things consistent.

As suggested by the averages associated with the four groups (i.e., the right four columns of the table), lower levels of expertise are associated with higher average ratings. A one-way between-factors ANOVA confirmed that these three means are reliably different ($F(3,796) = 199.18$, $p < .001$). Followup independent-sample t-tests further revealed that the average rating provided by experts was significantly lower than that provided by teaching assistants ($t(398) = 13.14$, $p < .001$), which was itself significantly lower than the average rating provided by peers ($t(398) = 16.18$, $p < .001$). As implied, the average expert rating was also significantly lower than the peer average rating ($t(398) = 31.35$, $p < .001$). The professor ratings were no different from the average TA ratings ($t(398) < 1$), but were significantly higher than the average expert rating ($t(398) = -8.80$, $p < .001$), and significantly lower than the average peer rating ($t(398) = 12.17$, $p < .001$).

Table 1: Means and Standard Deviations of Ratings Provided by Scorers

	Expert 1	Expert 2	TA 1	TA 2	Expert Average	Professor	TA Average	Peer Average
Mean	1.68	1.59	2.35	1.82	1.64	2.08	2.09	2.73
Standard Deviation	0.35	0.30	0.44	0.48	0.24	0.29	0.39	0.40
Note: Ratings composing these averages can range from 1 to 4								

The suggestion of these findings is that as one gains more experience applying the rubric, the scores one provides decrease. Recall that the experts underwent an entire course sponsored by AAC&U that was designed to teach experts how to apply the rubrics to work in a consistent and valid manner. The TAs and students watched a short video that attempted to enhance consistency and understanding with respect to rubric application. The course professor produced the short video. Thus all lie on a continuum of expertise that ranges from student and TA, though to professor, through to expert. The pattern of results seen here is likely due to enhanced discrimination abilities that come with expertise, allowing an expert to see differences in quality that the novice eye is not yet able to distinguish. As one becomes more sensitive to these differences one is less likely to give a composition that is just OK a higher rating (e.g., MacLean et al., 2010).

With these differences in overall group ratings noted, we now turn to an examination of the reliability of rubric scores that are based on averaging peer ratings. As we do so we will compute two statistics for each analysis. The first is a simple Pearson correlation that primarily assesses the ability of one rating group (e.g., students, TAs, etc.) to predict the ranking a composition will receive from another rating group. Pearson correlations capture associations in the rank ordering of compositions and are not affected by the specific values of the rankings themselves. In addition, we also compute kappa scores which assess true agreement between raters and thus capture the extent to which specific scores provided by rater groups match.

To better understand the manner in which these stats capture "prediction" differently, a quick re-examination of Table 1 is helpful. Note that while the experts gave lower scores than the TAs on average, within the experts (especially) and with the TAs (to a lesser extent) the ratings were similar. Thus when we compute kappas we should expect higher intergroup scores than cross-group scores, because the different level of the ratings overall will decrease the kappa score. However, as long as a composition that scored relatively highly within the context of a specific rater also scored highly in the context of a different rater,

the correlation could be high. The correlation is not sensitive to the overall rating level, but rather focuses on the rank ordering of the rating within the context of a specific rater.

Although our focus will ultimately be on the quality of the peer averages as measures of critical thinking, the within rater measures associated with the expert rater group and the TA groups provide relevant benchmarks for comparison. The expert raters give us a sense of what is likely the highest possible standard, and the TAs provide a benchmark reflecting the current standard practice in higher education, that of typically relying on teaching assistance to perform assessments. The correlations and kappa values associated with each of these groups is presented in Table 2.

Table 2: Correlations and Kappas within Expert and TA Pairs

	Experts	Teaching Assistants
Pearson Correlation	0.57**	0.45**
Cohen's kappa	0.40**	0.15**

Note: The kappas Reported are weighted Cohen's kappa values.
* indicates a result significant at $p < .05$ ** indicates a result significant at $p < .01$

The AAC&U trained experts did show the highest level of agreement in their assessments of work, both in terms of relative (i.e., correlation) and absolute (i.e., kappa) agreement levels. The TAs showed a high relative agreement, but a much lower kappa which reflects the fact that one TA gave much higher ratings than the other. Both correlations are in line with or, in the case of the experts, somewhat higher than correlations that have been previously reported for TAs grading subjective work (Cho, Schunn & Wilson, 2006; Paré & Joordens, 2008). In terms of absolute agreement, the Cohen's kappa value for experts is in line with, if not better than, that previously observed for AAC&U trained experts. For example, previous research on the reliability of the Critical Thinking VALUE rubric reported a Cohen's kappa of .29 between pairs of expert assessors (Finley, 2011).

With these values as benchmarks we now consider the peer averages. There is an additional statistical complexity that warrants discussion at this point. When TAs and experts rated the 200 compositions, the same two TAs and the same two experts rated all 200 compositions. However, when the peers assessed these compositions, a different set of six peers contributed to each peer average. This is statistically sub-optimal as is sometimes the case when one is researching real work occurring in a real class. However, it is important to emphasize the effect this could have on the results. The analyses we use assume a consistent set of raters, as is the case for our TA and expert data. When the raters change, as they do for our peer data,

that introduces an additional source of variance that, if anything, will reduce the degree to which the peer averages will predict — either relatively (i.e., correlation) or exactly (i.e., kappa) other datasets. Thus, to the extent this is affecting our data it will result in underestimates of the actual relationships making any evidence of relations we do see more trustworthy.

Before considering the correlations and kappa values reflecting the relative and absolute agreement levels of these peer averages with TA and expert ratings, we first quantified the agreement within the individual peer ratings that went into each average. This value is good to document as future studies may be interested in say, tracking how the agreement of values that make up an average change as expertise is acquired. To compute this value we computed weighted Cohen's kappas for all 15 distinct pairs of raters, then averaged those values. The resulting measure of within student agreement is 0.095. It is no surprise that of the three groups we examined, the first-year undergraduates are showing the lowest level of absolute agreement.

Of course the proposed process is one that is based on the average peer ratings, not on individual peer ratings. The correlations and kappa values relating average peer ratings to the ratings provided by our other groups (and averages thereof) are presented in Table 3. Once again, as evidenced in Table 1, the peers were much more generous in the ratings they provided and that alone will have a major negative impact on absolute agreement as captured by the Cohen's kappa values. Thus the near zero values we observe here are not surprising, though they clearly do reflect a bias in ratings that optimally would not occur.

Table 3: Correlation and Kappas between the Peer Average and the Experts and TAs

	Expert 1	Expert 2	TA 1	TA 2	Expert Average	Professor	TA Average
Correlation	0.22**	0.42**	0.35**	0.49**	0.36**	0.35**	0.49**
Kappa	-0.05**	-0.06**	-0.05**	-0.04**	-0.03*	-0.02	-0.03*

Note: The specific kappa computations were weighted Cohen's kappa and Fleiss' kappa for pairs of raters and groups of raters, respectively.
* indicates a result significant at $p < .05$ ** indicates a result significant at $p < .01$

When it comes to the relative ratings (i.e., the correlations), the peer averages are doing quite well. Focusing on the correlations with the expert averages, the TA averages and the score provided by professor, the observed values are not in the same range as those observed within the expert and TA groups. In fact, the peer average correlates a little better with the TA average than the two sets of TA ratings correlate with one another.

Recall that the analyses above were all conducted on the total rubric score applied to each composition. For completeness, we also analyzed the results broken down by the five distinct factors of critical thinking; explanation of issues, evidence, influence of context and assumptions, student's position, and conclusions and related outcomes. Rather than describe all these findings in the text we have summarized them in Tables 4 and 5, which correspond to Tables 2 and 3 but break the data down by factors. Overall, the general trends apparent in the overall data are represented and reliable at the factor level as well. No factor stands out as being differentially reactive in any manner that warrants deeper consideration.

Table 4: Correlations and Kappas within Expert and TA Pairs across the Five Rubric Factors

	Experts	**Teaching Assistants**
Explanation of Issues	R: 0.60** K: 0.59**	R: 0.26** K: 0.13**
Evidence	R: 0.54** K: 0.60**	R: 0.37** K: 0.11**
Context & Assumptions	R: 0.52** K: 0.54**	R: 0.28** K: 0.21**
Student's Position	R: 0.50** K: 0.58**	R: 0.35** K: 0.16**
Conclusions and Related Outcomes	R: 0.59** K: 0.59**	R: 0.27** K: 0.08**
Note: The kappas reported are weighted Cohen's kappa values. * indicates a result significant at $p < .05$ ** indicates a result significant at $p < .01$		

Table 5: Correlation and Kappas between the Peer Average and the Experts and TAs

	Expert 1	Expert 2	TA 1	TA 2	Professor	Expert Average	TA Average
Explanation of Issues	R: 0.16* K: -0.10	R: 0.27** K: -0.10	R: 0.29** K: -0.11	R: 0.38** K: -0.16	R: 0.28** K: -0.12	R: 0.23** K: -0.07	R: 0.42** K: -0.11
Evidence	R: 0.09 K: -0.14	R: 0.20** K: -0.15	R: 0.24** K: -0.17	R: 0.39** K: -0.15	R: 0.35** K: -0.04	R: 0.15* K: -0.11	R: 0.39** K: -0.08
Context & Assumptions	R: 0.05 K: -0.16	R: 0.12 K: -0.22	R: 0.27** K: -0.18	R: 0.36** K: -0.14	R: 0.27** K: -0.1	R: 0.09 K: -0.16	R: 0.40** K: -0.10
Student's Position	R: 0.02 K: -0.12	R: 0.13 K: -0.16	R: 0.18* K: -0.08	R: 0.25** K: -0.12	R: 0.27** K: -0.13	R: 0.07 K: -0.11	R: 0.26** K: -0.06
Conclusion and Outcomes	R: 0.09 K: -0.14	R: 0.28** K: -0.15	R: 0.28** K: -0.13	R: 0.36** K: -0.12	R: 0.32** K: -0.11	R: 0.19** K: -0.11	R: 0.41** K: -0.09

Note: The specific kappa computations were weighted Cohen's kappa and Fleiss' kappa for pairs of raters and groups of raters, respectively.
* indicates a result significant at $p < .05$ ** indicates a result significant at $p < .01$

Questionnaire Data

Just as they completed the activity, and still within the peerScholar platform, students were presented with the six statements listed in Appendix B and asked to rate their agreement with each on a 7-point Likert scale. Given that the issues we are asking them about are somewhat novel to them, the statements often provide some context prior to presenting the critical statement students are asked to consider. As a result, the questions are relatively long and not easily represented within a cell of a table. As a result we present the question numbers here and the full questions in Appendix B.

The resulting data is represented in Figures 1A through 1D. Figures 1A and 1C include the frequency of responses within each Likert value. Figures 1B and 1D show the distribution of responses that could be classified as "agree" or "disagree." The Disagree percentage reflects the sum of the 1, 2 and 3 responses divided by the total responses for that question. The Agree percentage reflects the sum of the 5, 6, and 7 responses divided by the total number of responses for that question. Figures 1A and 1B present the data for all students who completed the activity, who consented to allow their data to be presented, and who answered the question. Students were allowed to skip questions if they desired, which is why the total

number of points is in the 587 to 588 range for the most part (a question skip rate of about 12%). The consistency of this number suggests it's the same students who didn't answer any questions rather than a bias to avoid any specific question.

Figures 1C and 1D present only the responses that came from the students who produced the 200 compositions that were the focus of our previous analyses. The two sets of data are highly consistent and, as such, our commentary will not discriminate between the two.

Figure 1A: Frequency of Responses from all Participants within Each Value of the 7-point Likert Scale (n=587)

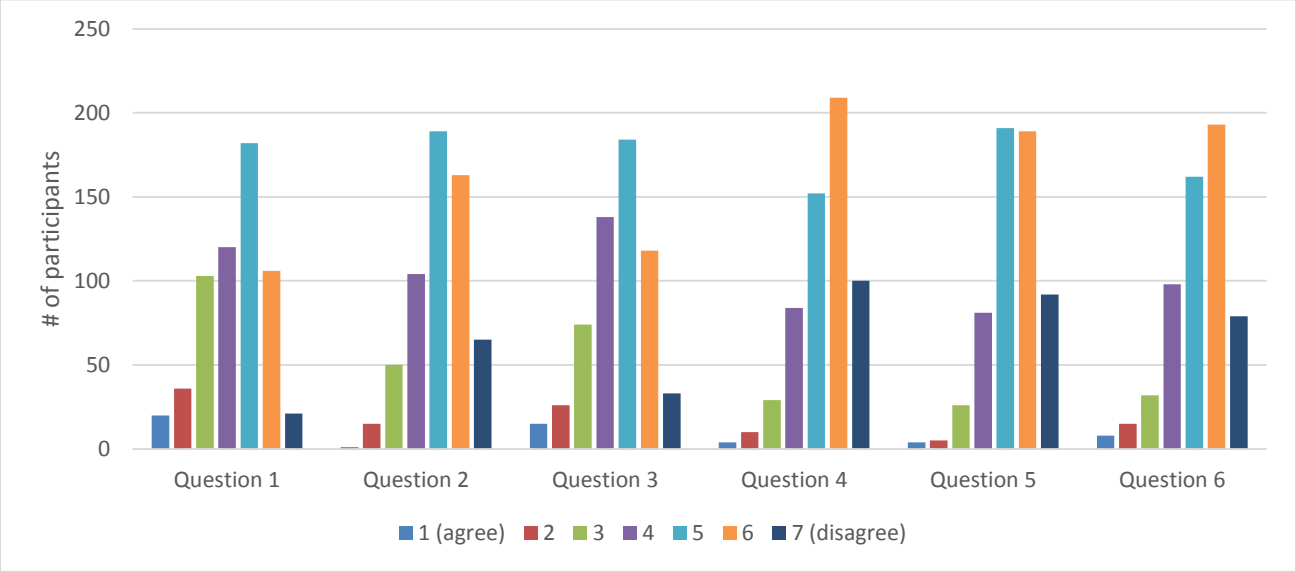


Figure 1B: Agreement and Disagreement Values for Each Question (n=587)

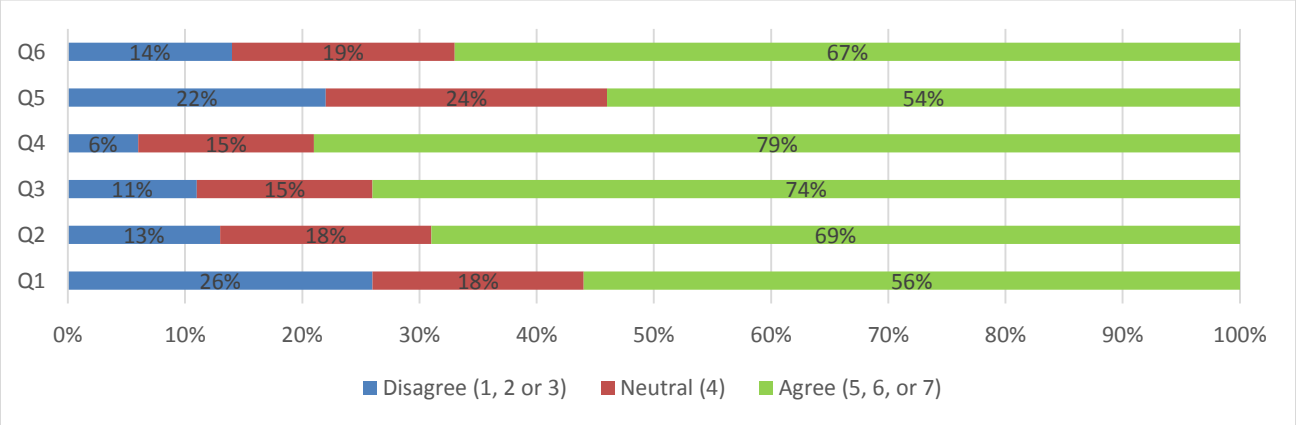


Figure 1C: Frequency of Responses from all Participants within Each Value of the 7-point Likert Scale (n=200)

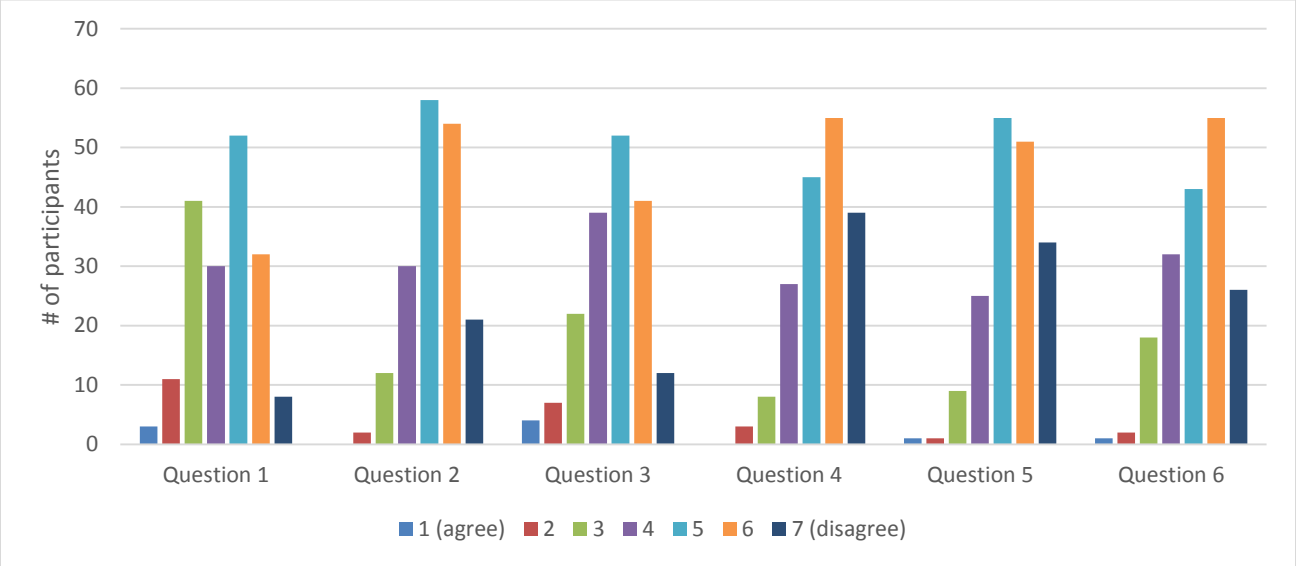
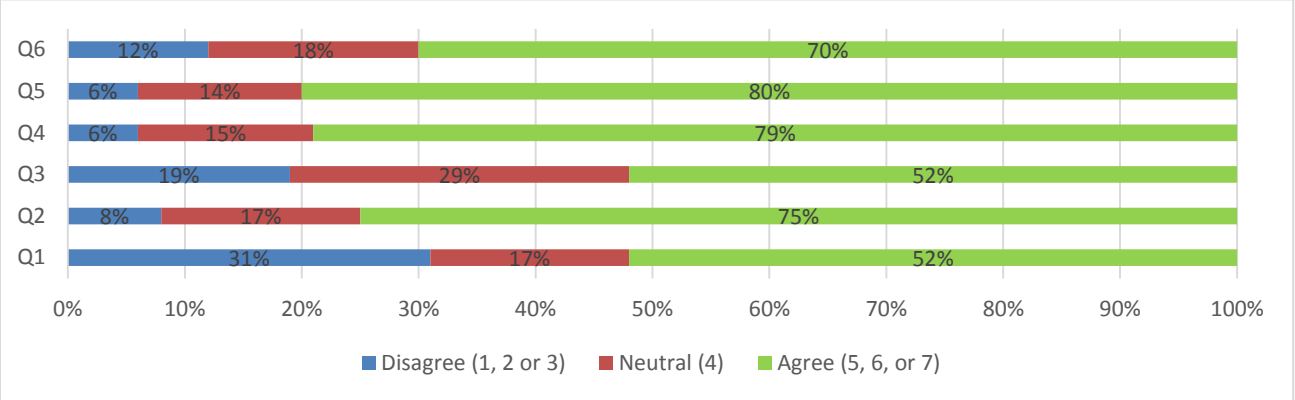


Figure 1D: Agreement and Disagreement Values for Each Question (n=200)



A brief summary of these findings: First, students generally agreed (52%) that the grades they receive in courses do reflect their learning, but this agreement level is among the lowest of the questions asked. They agreed more (71%) with the claim that typical university courses, as currently delivered, provoke them to think critically and creatively, and to communicate with others about their ideas. Taking these two questions together, then, we are not seeing deep discontentment with the current way of doing things in university settings. The next two questions focus on the need for structured practice to develop skills like critical thinking. Based on question 3, students agree somewhat (57%) that they are currently getting practice with

these skills, and yet the result of question 4 suggests that they strongly agree (78%) that they are willing to do additional work, of the sort supported by peerScholar, to develop these skills in a more formal and structured manner. The last two questions focus more on the need to measure these skills. They strongly agreed (80%) that the goal of measuring and documenting skill levels of transferable skills is necessary. With respect to doing so via the process that is the focus of this work (peer assessment), most agreed (71%) that this was a good approach.

These questions should not be seen as providing a definitive sense of students' perspective on these issues. The questionnaire has not been validated for reliability, some of the questions could be worded in more neutral ways, and it is generally a better idea to include some "reflective" questions. However, these responses do give us a sense of the student perspective, and it is gratifying to know that, when it is explained to them and when they experience it first-hand, they see the value both in measuring transferable skills in general and in the specific approach assessed here.

Discussion

Joordens (2018) argued that the current education system in unbalanced, emphasizing information acquisition over core skill learning. He further suggested a specific approach that could be used to right this imbalance, an approach that could be used in any course context and at any level. The described approach involved asking students to apply validated rubrics related to core transferable skills to the compositions submitted by their peers. This approach, he argued, could be delivered via any technology that supports peer assessments.

In Phase 1 we primarily focus on a central assumption of the proposed approach: that rubric scores derived by averaging a set of peer responses would provide acceptable measures of the skill in question. To assess this assumption we compared the ratings based on peer averages to ratings (and benchmarks) provided by two expert groups; teaching assistants and trained raters. To the extent that the ratings based on peer averages can predict expert ratings, in either relative or absolute terms, then further support for the proposed approach would be provided.

Our results suggest that, at least when the raters are first-year Introduction-to-Psychology students, the ratings based on peer averages provide good relative ratings, but not-so-good absolute ratings. That is, as suggested by the correlations, the compositions that the undergraduates rate highly are also those that the other groups rate highly, so there certainly is a sense that the students are sensitive to the differences in compositions and that their ratings reflect them. However, the low levels of absolute agreement as shown by our kappa analyses highlight that the undergraduates are simply less tough. They give significantly higher

ratings, the highest of all the groups. The conclusion, then, is that the ratings do reflect the amount of critical thinking that is present, but they do so in a way that is biased to overestimate.

Our assumption based on these findings is that as one gains more experience or training applying a rubric, one is in a better position to see things that could be improved. As a result, ratings of overall quality would reduce with expertise. The fact that the experts in our study showed both lower ratings, but high within-group kappa values suggests that the ratings they are giving are likely much more accurate. That is, the enhanced agreement suggests that experts are not just getting randomly tougher in their ratings but, rather, that there is a systematicity to why they rate compositions lower, which again suggests an enhanced sensitivity to the skill being assessed.

In previous work, Paré and Joordens (2008) similarly found that undergraduate students tended to give slightly higher grades to peer work than TAs did, at least in Phase 1. However, by simply describing to students the desired distribution of grades and the rationale to support it, undergraduate marks fell in line with those provided by TAs (i.e., Experiment 2). Thus, it's possible that a similar approach of specifying a desired distribution similar to that demonstrated by our experts may similarly reduce the scores that undergraduates provide and thereby increase their levels of absolute agreement with experts.

That said, expertise itself is likely an important issue related to this pattern of results. Once again, as one gains experience within a discrimination context, one's discriminations become more fine-grained and nuanced (MacLean et. al, 2010). They see problems that they did not see when they were less experienced, and this is likely why our experts give lower scores on the same work relative to TAs and undergraduates. Thus, it may be the case that undergraduates could only reach high levels of absolute agreement with experts by gaining more experience applying rubrics.

If the application of rubrics in the context of self-assessments became common, then as students progressed through their courses their expertise would naturally increase with experience. Thus, we would expect higher kappa levels if we compared second years to experts, even higher for third years and higher still for fourth years. However there is a worry with this sort of passive approach to waiting for the development of expertise. That is, if experience results in decreased scores, then even if quality of, say, critical thinking is improving from first- to fourth-year students, if these students are always assessed by their peers, and if these peer assessments are becoming lower with additional experience, the gains in critical thinking might be masked by reduction in scores related to experience. These two factors could perhaps be teased apart by a principled consideration of kappa values, but this example does highlight the complexities one must keep in mind when measuring transferable skills using student ratings.

A different approach would be to more actively train students early on. We did attempt to provide some training in the current experiment but perhaps a much more formal training experience is needed, one that includes pre-scored exemplars and more formally requires students to calibrate their scoring with these examples before assessing peer work. In a sense, this is an argument for a sort of training that is similar to the kind of training that our AAC&U expert raters experienced. As argued earlier, this sort of training and the practice applying rubrics likely represents a powerful pedagogical approach for teaching these skills, so the need to train may actually be a good thing to add to the proposed approach.

One last important note about this bias of undergraduates to provide relatively high rubric scores is that, with forethought, that tendency itself could be measured and even corrected for. Specifically, as students are assessing peer compositions it could be possible to include some compositions that have been pre-scored by experts. These differences in the ratings that a student provides to these compositions relative to the experts could be used to actually correct the student's ratings. That is, the extent of bias could be quantified and corrected for, and this approach would allow one to explicitly see the bias hopefully reducing as a student's experience increases.

Do average ratings, based on peer-provided rubric scores provide a valid and reliable measure of the transferable skills related to the rubric in question? The current work suggests that they certainly have the potential to do so, although in our data there is a bias of undergraduates to provide slightly elevated ratings, even while predicting expert ratings well. Additional work that either finds ways to remove this bias, or that finds ways to estimate and correct for it, would be extremely useful to maximize the validity of peer averages.

In addition to assessing the reliability of the peer averages as measures of transferable skills, we also described some questionnaire data designed to elucidate student opinions about the need to develop and measure transferable skills more aggressively than we do now. At a very general level the resultant data suggested that while students are not completely dissatisfied with their current educational experiences, they do see the value in both developing and measuring transferable skills; they endorse the approach they experienced, and they are willing to do more work of this sort if it means their skills would be better developed.

Phase 2

Phase 1 primarily focused on whether averages derived from peer-provided rubric scores could be used to predict expert ratings. The results were generally heartening as the ability of a peer average to predict expert scores was in the same range as the ability of TAs to do so, and to predict one another's scores. Yes,

there are still some issues regarding a bias to score high that still need to be worked out, but generally the approach seems to have great potential.

The second phase we report here shifts to a different but also relevant question: Can this approach of having students apply rubrics as they conduct peer assessments work across a wide range of educational contexts? That is, is this approach truly scalable to virtually any educational context in the manner needed for widespread use in our current educational systems? Such scalability is necessary if this approach is to be viewed as a means of rebalancing complete educational systems in the way advocated by Joordens (2018) and thus it is a question worthy of assessment.

What will follow is perhaps best considered a "proof of concept" assessment. The approach is straightforward. The authors of this report, primarily Joordens, reached out to colleagues entreating them to please try the described process within their class, whatever that class may be. The goal was to get as many colleagues as possible, across institutions, and in fact across countries, to try it and to have them and their students report back about their experiences with it. The students were asked the same questions asked of the students who participated in Phase 1. The faculty were asked a separate set of questions (see Appendix D).

One of the primary issues we hoped to address was scalability in terms of adoption and continued use of the approach. That is, for a new approach to be truly scalable within an education system, certain truths have to hold. First, from the faculty perspective, the new approach must be easy to implement. All the hype about disruptive technologies aside, faculty are already time-challenged and for them to seriously consider a change to their teaching that change must be easy to implement. In addition, those same faculty have to believe the time invested was worth it, that students learned something valuable or that some other relevant goal (e.g., the measurement of a skill) was achieved. These were the issues we hoped to measure in our faculty questionnaire.

Also relevant to scalability though is the student reaction to the process. Educators love to highlight the merits of active learning (Bean, 2011), so why is every classroom not relying heavily on active learning already? Why are so many still characterized by a textbook, lectures, multiple-choice exams and nothing else? Part of the reason is that, from a student perspective, active learning means time and effort, often in a relatively novel context. It's less familiar, it requires time management, and it requires them to work through something in a way studying for a final exam may not. If students don't like it then faculty evaluations may suffer, making faculty less likely to adopt these approaches.

All of this to say that student reaction to the learning experience matters. If they enjoy or see the value of some form of learning, they may become one of the forces that makes that form of learning more common. Likewise, if they find a form of learning tedious, difficult, uncomfortable or for any reason undesirable, they may become one of the forces that make that form of learning less common. Thus learning more about students' reaction to the process of asking them to apply rubrics related to transferable skills to the work of their peers is relevant to the question of how scalable that approach may be.

In Phase 1 we sought some initial data from the student perspective, and it was largely positive, suggesting that students enjoy the process and see its value to their education. While these findings are heartening it is perhaps important to note our involvement with these particular students and the potential that they were providing answers that they assumed their professor would like. Acquiring similar data from students less directly connected to an advocate of the approach could tell us the extent to which students truly value the process.

With all this as context, the remainder of this paper will recount a progressive approach to eventually realizing the goals just described. This was the first time that this team attempted to perform research with cross-intuitional faculty, the vast majority of which they did not know personally. Doing so can be very challenging, and part of this report will highlight some of these challenges as they were part of a journey that provided different sets of data relevant to the process under investigation.

Overview of the Recruitment Process

Perhaps we all cannot help but believe that others playing a role similar to us will feel as passionately about certain things as we do. With this sort of mindset in play, we began reaching out to other faculty members via posts to LinkedIn groups, advertisements in Academica (an e-newsletter that reaches Canadian academics daily), listserv groups such as the one the serves the STLHE (Society for Teaching and Learning in Higher Education) community, social media outreach, word of mouth, and literally any other means possible beginning in the summer of 2018. Interested participants were asked to please fill out a Google form to indicate their interest and to answer a number of questions.

Interest appeared high with 53 different faculty members who took the time to fill out the initial form which, among other things, asked for their email addresses. All who filled out the form were sent an experiment protocol that explained the steps necessary to participate, and each was advised to reach out to facilitate provision of the necessary software, and to get any necessary assistance with ethics forms or any other formality their institution may require.

Well into the fall 2018 term things remained quiet, too quiet as they say. As we began reaching out to them one by one, most simply did not respond at all. Of those who did, some said they didn't fully understand the requirements or, for various other reasons, were no longer willing or able to participate. The few who were still interested seemed unsure of what to do despite all the information that had been provided to them. At this point it was obvious that most of these participants would not actually implement the process or even try, and a new approach was necessary. That said, the data initially gathered from these 53 was gathered and will be described as "the initial 53."

Although disheartened, we revised our approach and once again cast out recruiting nets, this time being much clearer up-front about what was expected, and investing more time early in the process forming relationships with the potential participants and being clear about expectations of participation throughout. We also reached out more heavily to colleagues within our own institution in hopes they would be more likely to carry through once they committed to doing so. Those participants, 14 professors in total, make up the primary group used to access scalability of the approach. They also filled out a Google form that provided some information but did so at the end of their participation. Their data will be referred to as that from "the participating 14" given they actually participated in the study as intended.

Methods: The Initial 53

Participants

Of "the initial 53," 45 were from Canada, three from Australia, two from the UK, two from Africa, and one from India. They taught courses that ranged across all university levels (including graduate) and spanned nearly all disciplines. In addition to providing these basic demographics they answered two questions the answers to which were deemed sufficiently interesting to include in this report.

Materials

All respondents provided their data via a Google form. It was a simple form primarily designed to assist the investigators in terms of connecting with the participants. However, it included a consent form to use the data, and beyond the basic demographics it included these two questions:

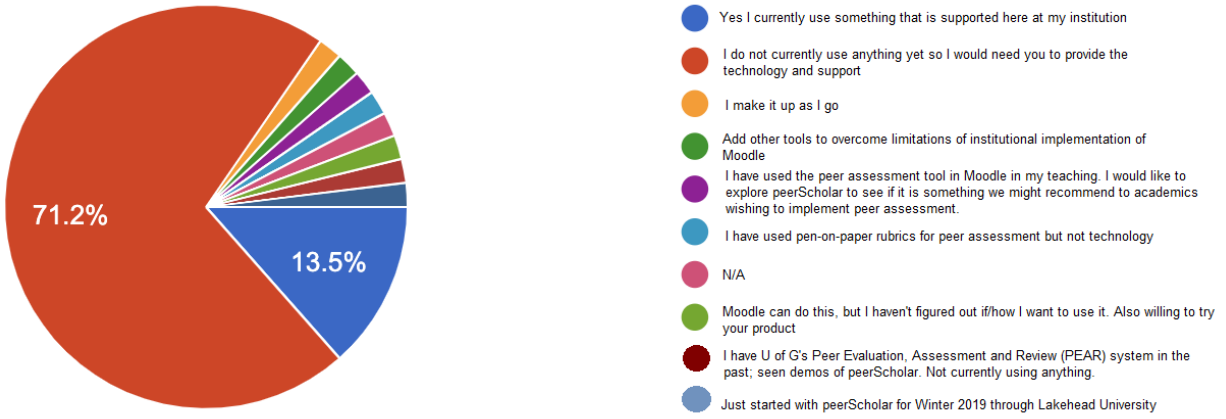
1. Do you currently use some form of technology-enabled peer assessment or would you like us to provide the technology and support?
2. Which of the following transferable skills are especially important to you?
    - Critical thought

- Creative thought
- Metacognitive Awareness
- Written communication
- Receptive communication
- Oral communication
- Numeric literacy
- Scientific literacy
- Other

Results and Discussion

To the best of our knowledge no one has published data related to the two non-demographic questions that were posed and thus we decided to include the findings in this report. With respect to the first question, the results are depicted in Figure 2. About 13.5% of the faculty who responded to our invitation to participate currently use some form of technology-enabled peer assessment at their institution. This figure is relevant to the issue of the scalability of the process under investigation as some form of technology-enabled peer assessment system is needed to manage the process.

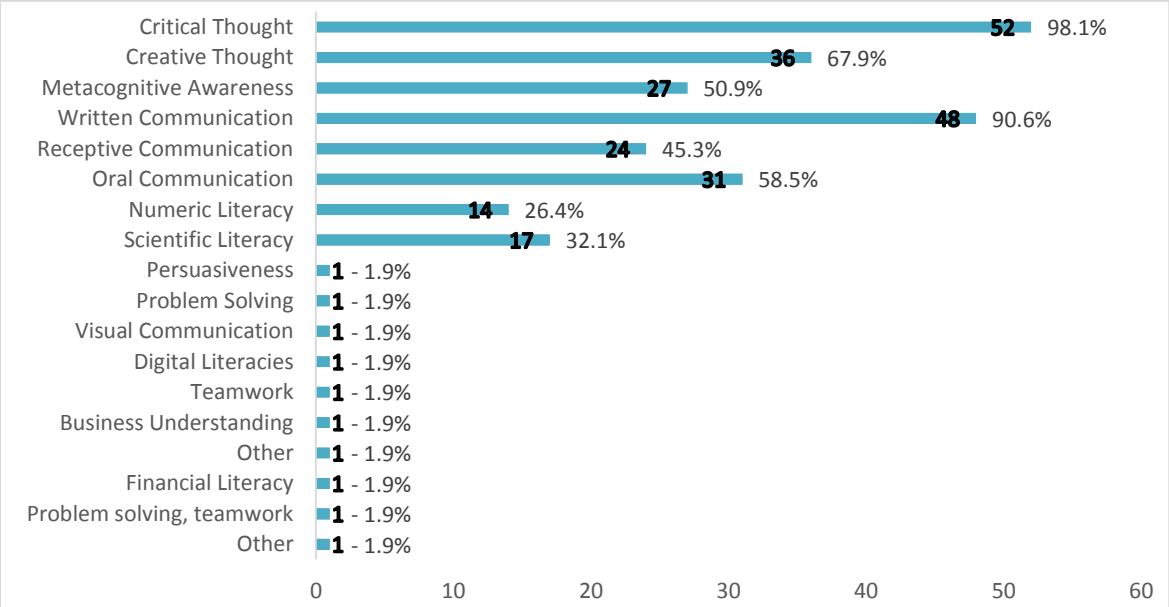
Figure 2: Responses to the question: Do you currently use some form of technology-enabled peer assessment or would you like us to provide the technology and support? (n=53)



The second question asked which transferable skills are most relevant to the given faculty member. We allowed faculty members to check as many as they wished from a list of eight provided skills, and we also allowed them to write in their own if it was not on the list. The results are depicted in Figure 3. Clearly, critical thinking and written communication lead the list with over 90% choosing each as relevant. These were followed by creative thinking, oral communication, meta-cognitive awareness and receptive communication. Of the list, the remaining two, scientific literacy and numeric literacy, had the lowest "relevance" rates, but this is likely a function of the fact that only courses in specific areas — the sciences especially — focus on these skills. All of the items with a frequency of 1 were written in by respondents.

Once again, the intention was that all of these participants would try the proposed process of having their students apply a relevant transferable-skill rubric to the work of their peers in the context of peer assessment but, for the most part, that step was not taken. In fact, some of these faculty are part of the next 14 to be discussed, but most are not. Nonetheless, their responses to these two questions do highlight two points. If this process is to be used widely, one critical step will be getting the technologies that support this sort of peer assessment used more widely than they are now. In addition, were one to focus implementation of this approach on just one or two transferable skills initially, critical thinking and written communication are the two that seem most relevant to those potentially interested in the approach.

Figure 3: Responses to the question: Which of the following transferable skills are especially important to you? (n=53)



Methods: The Participating 14

Participants

The 14 professors who implemented the described process in their class all did so in the winter 2019 term. They comprised eight professors from the University of Toronto (four of which were from the Ontario Institute for Studies in Education), two from Lakehead University, one from Humber, one from Assiniboine, one from Urbana University (US), and one from Utrecht University (The Netherlands).

Of these 14 faculty, nine used it in one class, four used it in two classes, and one used it in three classes so, in total it was used in 20 classes. These classes included classes from first year to master's level in subject areas including Psychology, Neuroscience, Auditing, Advanced Auditing, Human Resource Management, Law, Criminalistics, Oral French, Education, Research and Architectural Design.

Materials

A research protocol was created to guide participating faculty through the setup of their activity. This protocol outlined the need to incorporate three things into any activity (a) an aligned "validated rubric," (b) the consent "opt in/opt out" prompt and (c) the six questions for students (Appendix A). The consent prompt and the student questions were provided in the protocol.

Professors were obviously free to define the student composition in any way they wished but they were strongly encouraged to choose the rubric they would use first, and then ensure the assignment they gave students aligned with the rubric. For example, if they intended to use the AAC&U Critical Thinking rubric, then they were asked to make sure the task they gave students asked them to show evidence of critical thinking. The consent prompt and questionnaire will be discussed in more detail in the "Procedure" section that follows.

Finally, the faculty were provided with a zip file that contained all 16 of the validated AAC&U rubrics. While we encouraged the use of one of these rubrics, a few faculty members pushed back. For example, one faculty member was asking students to produce a reflection of learning, and she had a validated reflection rubric she wished to use. Ultimately we felt it was the process, and not the specific rubric, that was the focus of the study so we allowed those who desired to use any validated rubric they preferred.

Procedure

These 14 participants were chosen by Steve Joordens, in part, because of the fact that they would respond to and interact with him. Each was sent a research protocol that outlined the steps needed to make their peerScholar activity conform to the research needs. Assistance was provided, where necessary, to assist with filling out any necessary forms (e.g., ethics) and to provide access to the necessary software. We also offered to assist with the setup of the activity but all preferred to do that on their own.

The technology used to administer their activities, peerScholar, has quite a few parameters that allow an instructor to set up their activity as they wish. We generally allowed instructors to do as they wished with the caveat of the following items that we insisted they do:

- In the Create Phase, the relevant rubric had to be attached with the instructions such that it was available to students at all times

- In the Assess Phase, a consent prompt was added. Specifically, after performing their peer assessments but before being allowed to submit, a pop-up box emerged that read, “This activity is part of a multi-institutional research project, and we’d like to include your data in our research. We assure you that your data will be kept confidential, specifically it will be anonymized and secured. In order to use your data we require your consent to do so … your professor will not be told whether you consent or not, and your decision will not impact your coursework in any way. With all this in mind …” and below that they had to choose one of two options from a pull down menu, either consenting to their data being used or declining.

- Finally, in the Reflect Phase, after students saw and assessed the feedback they received, but before they could submit, another pop-up emerged that contained the six questions presented in Appendix A. The answers to these questions are one of the primary datasets of interest in this study.

Once again, most instructors did all of the above with little to no help from the research team although, in a few cases, we were asked to take a look and make sure all was set up correctly. At that point all was allowed to proceed until the end of the term. At the end of the term, the participating instructors were contacted again and asked to fill out a Google form that included the Faculty Questionnaire (see Appendix D). The answers to these questions are the second dataset of interest in this study.

Results

Of the 1,907 students who filled out the survey, only 32.6% consented to allow their data to be used for research. This value is lower than expected or hoped for, but perhaps reflects the "distance" between researcher and participant when research is cross-institutional in this manner. The survey results for those students who consented are presented in Figures 4A and 4B.

Figure 4A: Frequency of Responses from all participants within Each Value of the 7-point Likert Scale (n=621)

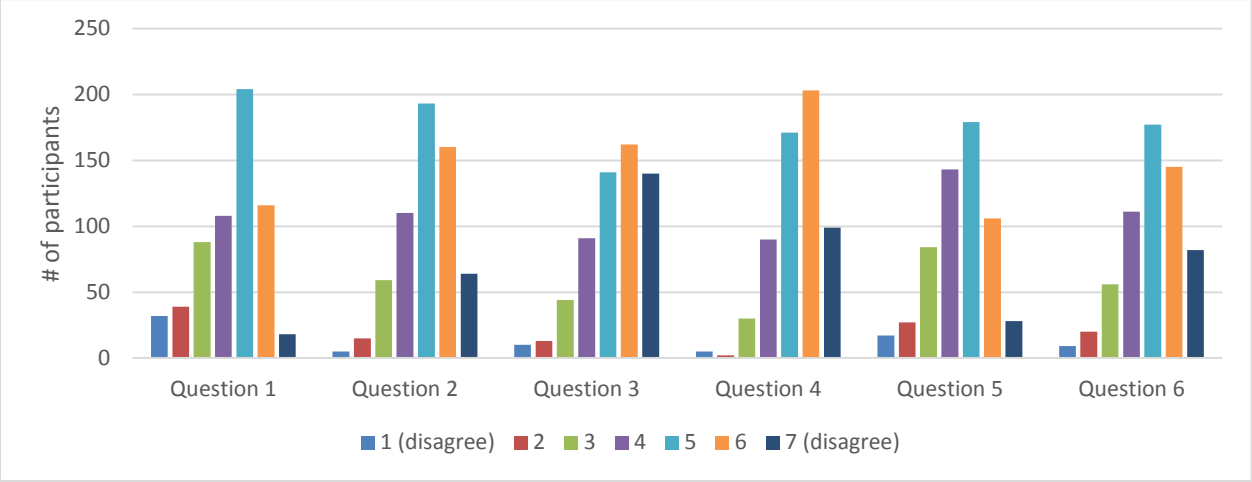
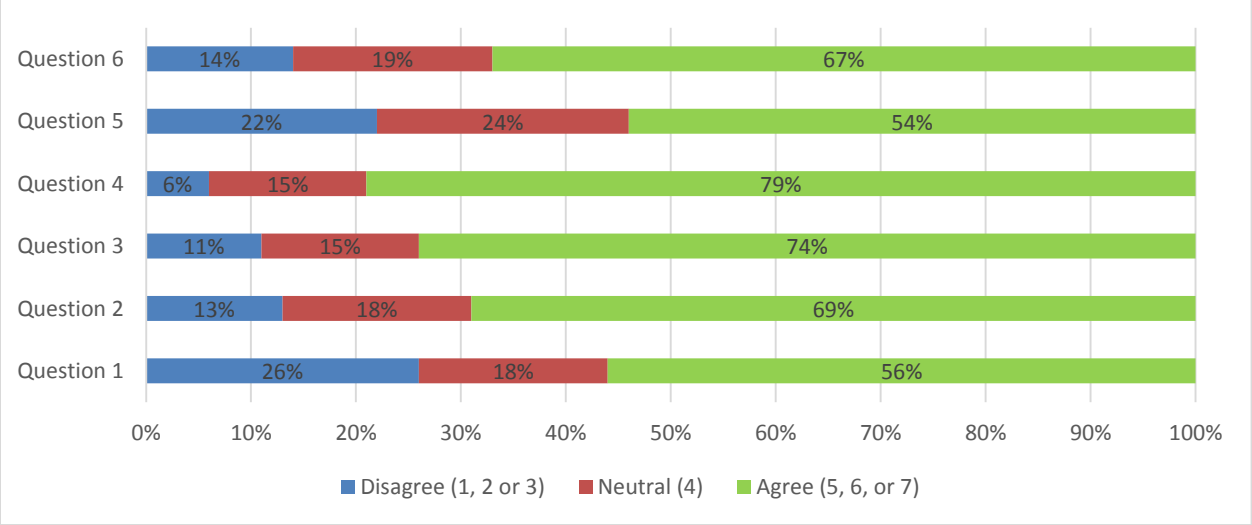


Figure 4B: Agreement and Disagreement Values for Each Question (n=621)



Despite the low consent rate, this data replicates all the patterns seen in Study 1, patterns suggesting that students who experience the process see its value and endorse its wider usage. To the extent that there are differences, they are mostly apparent in slightly lower agreement rates to the last two questions. That is, they were less convinced that measuring skills was important, despite still agreeing overall, and they were less convinced that the process they experienced was a good way of doing so, despite again agreeing with that premise in general. When one steps back a little the critical point is this: There is no revolt and, in fact, students are buying into this approach to skill development and endorsing its wider usage.

Turning now to the faculty survey, all 14 faculty members consented to allow their data to be used and thus the results for the entire sample are summarized in Figures 5A and 5B. For each question we also allowed qualitative comments and given the relevance each question will be discussed in some detail.

Figure 5A: Frequency of Responses from All within Each Value of a 7-point Likert Scale (n=14)

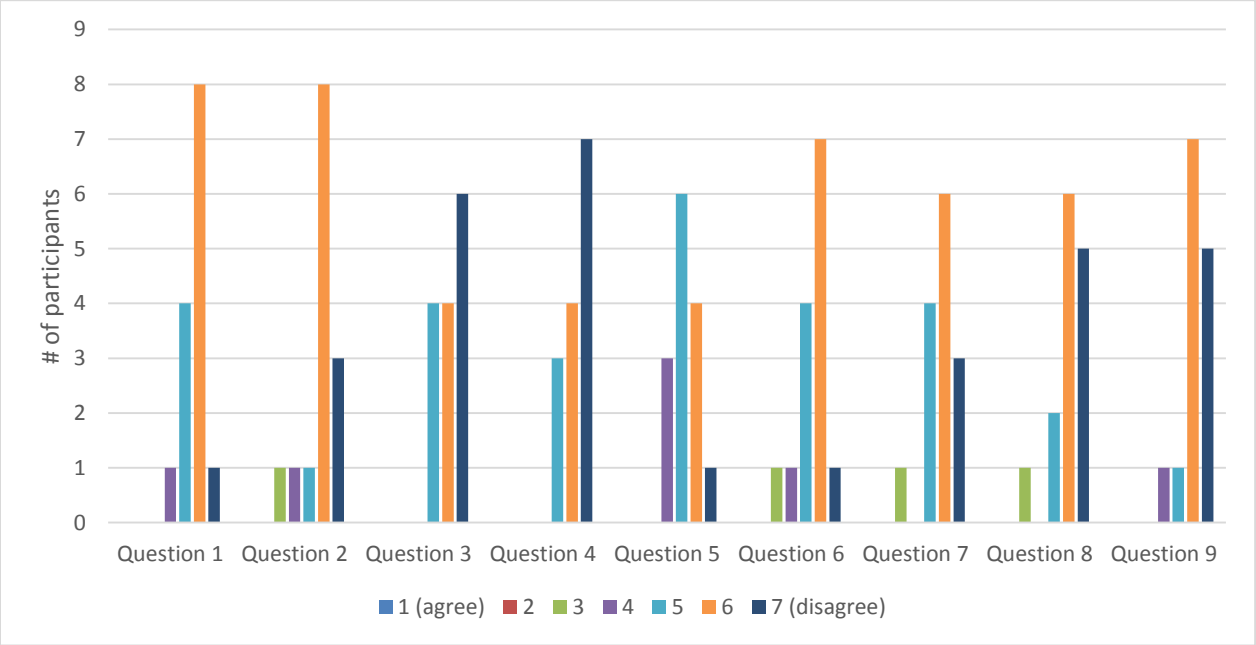
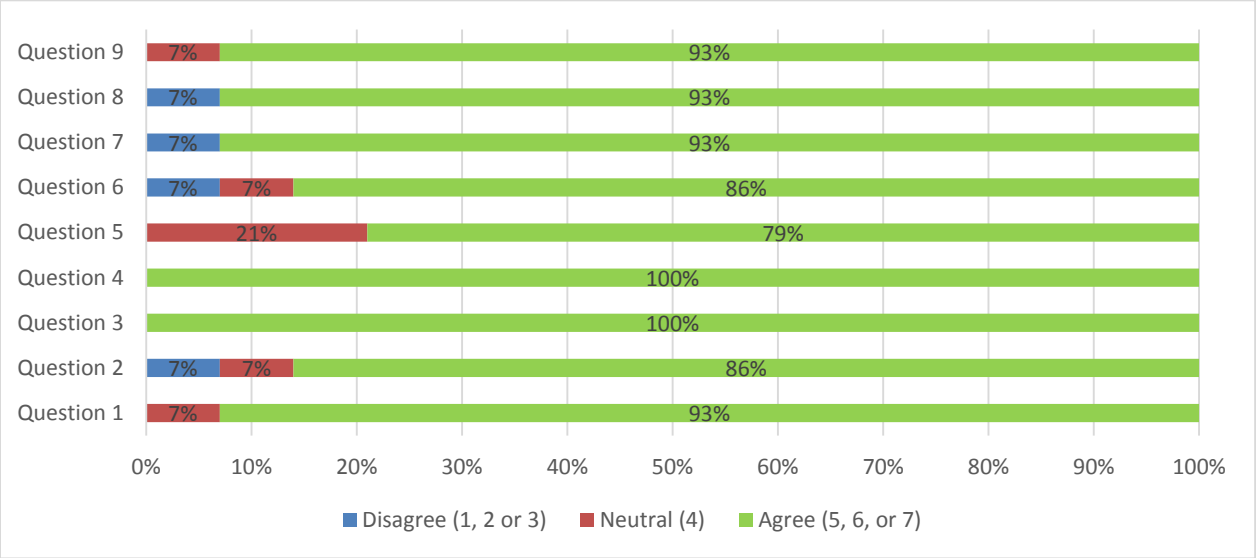


Figure 5B: Agreement and Disagreement Values for Each Question (n=14)



Questions 1 and 2 related primarily to the ease of use of the technological solution. Was the activity easy to learn (Q1) and do they believe the subsequent activities will be fast and easy to set up? A rating of 6 out of 7 was the modal answer in both cases, suggesting that the technology was, in general, well received and not seen as representing a major barrier. The qualitative comments largely corresponded to this reading, though some did point out areas where things could be clearer still.

Question 3 was about perceptions of learning, asking whether the faculty felt students had learned a great deal, more than they would have without peer assessment involved. All faculty agreed with this, with 7 out of 7 as a modal value. Clearly the faculty see the pedagogical value of the process. One comment said, "Some students said it was the most important thing they learnt."

Question 4 focused more directly on the approach being studied, asking faculty if they believed this combination of peer assessment with the application of valid rubrics seemed like a good way to teach transferable skills. Once more, 100% agreed with this and once again the modal reply was 7 out of 7.

Question 5 asked faculty whether they thought students enjoyed the experience. Recall that active learning requires students to spend time on and invest significant effort in their learning. Thus, it would be possible that students would push back and let faculty know they were not happy. Fortunately, we see no evidence of push-back reaching the faculty. In fact, 79% of the faculty believe their students enjoyed the activity, with the remainder feeling neutral about it.

Question 6 combines two central issues: Did the faculty feel that this process provided a powerful learning experience while requiring little time and effort on their part? Twelve out of 14 faculty (86%), agreed with this, with a modal answer of 6 out of 7. Yes, the process certainly requires time and effort on the part of the faculty to set up and manage this process. This high buy-in rate on the part of faculty is encouraging in this respect.

Question 7 is focused on the scalability of this approach to "any educational context and any level of study." Despite the wide-ranging aspect of this question all but one faculty member agreed, with a modal value of six out of seven. The one faculty member who disagreed added a comment suggesting a suspicion that it may not work at the primary level of education.

Question 8 was more focused on continued use; was the faculty member sufficiently impressed that they think they will continue to use this process? This aspect is sometimes referred to as "stickiness" with sticky approaches having more potential to scale and grow. Once again, all but one faculty member agreed that they would continue to use this process, with a modal value of six out of seven on the agreement scale.

Finally, Question 9 asked if they would like to see more courses adopt this approach, thereby giving students additional practice across varying contexts. This would of course be the goal if use of this process would scale within an institution. Once more, all but one faculty member agreed they would like to see more courses using it, again showing their general support of the approach they and their students experienced first-hand.

Taken together, this data from both students and faculty members is extremely positive with respect to the potential of this newly proposed approach for the formal development and measurement of transferable skills. There are no signs of resistance from either faculty or students. Rather, there is a general endorsement of the approach and its value from both, and a desire to see it employed more widely. Critically, this endorsement comes from instructors and students experiencing the process across a wide range of educational contexts.

General Discussion

Joordens (2018) argued that the current approach in higher education of only informally teaching core transferable skills like critical and creative thinking, and clear effective communication needs to be formalized given the importance of these skills in a dynamic work environment. He proposed what may be the first formal approach to both the development and measurement of any skills that can be captured by a rubric. The approach, which combines technology-enabled peer assessment with the principled application

of validated rubrics, is one that theoretically could be applied in any discipline and at any level. The potential of this approach to impact students' future success appears very high but only, of course, if the described process is as scalable as suggested.

The purpose of the research described here was to test several aspects of the approach that are relevant to its scalability. First, the approach assumes that one could rely on averages derived from peer-based rubric applications as a valid basis for measurement of the skill in question. Second, for any educational technology to be scalable, it must be easy to use, its value to learning must be accepted, and it must not provoke any sort of backlash from students or faculty who use it. Better yet if those who enjoy the process also see its value and would like to see it used more widely. Each of these issues will be addressed one at a time in the sections that follow.

Validity of Peer-based Rubric Scores

Note that there are two claims Joordens (2018) makes about the process he highlights. First, he claims it provides a powerful approach to developing transferable skills by adding a form of "ear training" (Woody, 2012) to the students' experience while learning a new skill. For example, the suggestion is that if students are learning about, say, critical thinking then within the described process all of the learning that would occur via the typical peerScholar process (Joordens et al., 2014) would be enhanced further as they look for and score the amount of critical thinking present in the work of their peers. This enhancement in skill development was not directly examined in this paper, but on its face it seems sensical.

Joordens (2018) second claim was that the approach could also provide a measure of any given skill in the context of the assessment. Specifically, the claim was that if a given student's work is assessed for, say, critical thinking by five or more of that student's peers, those five or more scores could be averaged together, and that average may provide a valid measure of the skill just as peer averages have previously been demonstrated to provide valid measures of overall quality of work (Cho et. al, 2006; Paré & Joordens, 2008).

The ability to measure is critical to the adoption and scaling of any formal approach to developing skills within institutions of higher education (Liu, Bridgeman & Adler, 2012). Such institutions use measurements to shape learning, and to credential it which, to some extent, is a major role such institutions play in society. Thus, Joordens' (2018) claim that the proposed approach can provide such measures easily and in the context of regular work is a key feature that gives the approach such high potential impact.

Do rubric scores based on peer averages provide valid measures of skills? This claim was the focus of Phase 1 and it was found to be somewhat true, but not totally true, at least as implemented here. Specifically, the ability of peer-based averages to both relatively and absolutely predict ratings given by teaching assistants and experts was assessed. The peer-based ratings did a good job predicting the relative ratings, meaning that those compositions that peers found relatively strong (or weak) were also the ones that the TAs and experts found relatively strong (or weak). However, when it came to making absolute predictions, that is predicting the actual rating a teaching assistant or expert would give to some composition, the peer-based averages did not do so well. This failure to make accurate absolute ratings was due in large part to the fact that the peers gave reliably higher scores as compared to the more expert scorers.

Before dwelling on the failure to make accurate absolute predictions, it may be wise to first contextualize the data in terms of the merits of the proposed approach. Some approaches to measuring transferable skills, for example the Cornell Critical Thinking Test, require students to log on to a website and spend a significant amount of time answering a wide range of questions (Ennis, 1993). If an institution wishes to measure impacts of learning on critical thinking using this approach, they must set up testing sessions, say at the beginning of Year one and the end of Year four, and entice a sample of students to do the assessment. What's more, this would only measure critical thinking. Tests like this do not exist for every transferable skill, and if they did then the logistical demands of administering this approach would be necessary for every skill the institution valued. A much better approach is to find a way of measuring a skill in the context of natural work the students are doing as part of their education (Simper, Frank, Scott & Kaupp, J., 2018), and that is what the process recommended by Joordens (2018) allows. Moreover, because it can measure whatever skill corresponds to the rubric implemented, it provides a means to measure any skill in any context at any point in the students' university career.

It is also important to note that the ability to make relative predictions accurately is also worthy of celebration. Peers are successfully detecting the presence (or absence) of a skill, and they are reflecting back to students the strength of their skills relative to their cohort. When this is considered in light of the logistical advantages described above, it's clear that progress in the development and measurement of these skills is being made.

Ultimately though, the endeavor would clearly be more impactful if the peer averages provide accurate, absolute measures of the skills in question. From this perspective the current data provides a starting point for more work. We highlight two obvious approaches for enhancing these measurement scores.

The first is to focus on the training of the students prior to them asking them to assess their peers. AAC&U has created a training program and they are currently examining ways of making their training available

online (Terrel Rhodes, personal communication). Perhaps students could be required to attend this training as part of a course. This is not at all absurd as undergoing training that would focus on what, say, critical thinking is, and how it looks when present, would be a fantastic academic experience for students. At the very least we might imagine a shorter course that could be an online module that might scaffold the process of applying rubrics in a manner more intense and perhaps relevant than the short training video we used here. Further research could assess different approaches to training and their effects on validity to determine an optimal way forward.

Even with great training at the outset though, it is still likely the case that experience in applying the rubrics will lead to lower scores as the assessor begins to see finer grade discriminations in the work (MacLean et al., 2010). This may simply be part of the learning process that can never be eliminated. If this is the case, then another good approach might be to measure the extent to which bias is present, and then to statistically control for it. For example, an instructor could have a set of compositions related to some activity scored by experts. They could then include these activities in with the actual student compositions, resulting in students being asked to score those compositions as well. The difference between the students' average ratings and those provided by the experts could be used to estimate the amount of bias present in the student ratings, and this bias could be subtracted from the actual ratings. Although requiring some additional time and effort, this approach not only allows for control of inflated ratings, it also allows one to know the extent to which such a bias is present, and how that bias changes over time.

In conclusion, our assessment of Joordens' (2018) claim that the proposed approach can provide measurement of core transferable skills in the context of natural class work is generally positive. As assessed, peer averages provide strong relative ratings of the skills in question, and with a bit of follow-up work there is every reason to believe that strong absolute ratings can also be obtained, perhaps through a combination of more in-depth training and the use of pre-scored compositions to estimate and control for bias.

Will They Use It?

The success of any educational technology depends of course on its efficacy and relevance to institutional goals, but it also depends on much more human factors. As every Centre of Teaching and Learning knows, when enticing faculty to consider some new technology or approach, a core concern is always ease of use for faculty (Bates, 2000). Faculty simply do not want to completely reinvent what they do, especially on a regular basis. But if a new technology or approach purported to enhance pedagogy can be implemented without major disruption, then its use is much more likely to scale within the institution.

Also relevant is the reaction of students to the new approach. Most new approaches being championed involve some form of active learning and at its heart, active learning requires students to do what many of them consider to be work (Michael, 2006). The simple fact that something is different may already cause some discomfort in students as we all feel most comfortable within familiar contexts (Reis Maniaci, Caprariello, Eastwick & Finkel, 2011). The additional need to do something also brings in time-management issues and requires effort. Thus, it should not be surprising if students sometimes resist new active learning opportunities, even those that may enhance their future success. Of course, student resistance can be reflected in lower course evaluations, and given that course evaluations are a major contributor to decisions related to promotion and salary raises, scalability of a technology is also sensitive to the student reaction (Simpson & Siguaw, 2000).

Given this, both faculty buy-in and student buy-in determine the extent to which new teaching innovations will scale institutionally. When the innovation requires use of a new technology, then ease-of-use of that technology is critical, as are perceptions of learning and perceptions of student acceptance. If an innovation can address all of these concerns, then it has a high chance of gaining widespread use.

Joordens championed a process in his 2018 work, claiming that it can work in any discipline at any level and, depending on the rubric employed, can be used to develop and measure any transferable skill. These are strong claims that, if true, suggest that this is not a niche process but rather one that could have a wide-ranging impact on the assessment of transferable skills across a variety of institutions.

Phase 2 represents an attempt to directly assess these claims of scalability, primarily by asking faculty to try it in their class, whatever their class may be, and at whatever level it is taught. After trying it, both faculty and students filled out questionnaires to assess the issues highlighted above including perceived value of the process, ease of implementation, enjoyment, and agreement that the process was indeed as potentially scalable as suggested.

The approach was tried across a number of varied contexts, from first-year oral French to third-year law, to a master's level class in education. By and large, the faculty were very positive in their assessment of the approach's potential. Perhaps most relevant, virtually all agreed they would continue to use the approach and that it should be used more widely. Crucially, the students also responded positively overall to the experience despite the fact that it required significant time and effort on their part. They also endorsed the claim that it should be used in more of their courses, and as has been found repeatedly, they value the opportunity to learn from and with their peers (e.g., Topping, 1998).

Of course, the challenge with any new approach or technology is to get it used in the first place, but from the current data it appears that once it is tried, the approach described by Joordens (2018) appears "sticky." Knowing that those who have tried it stick with it may itself be an important part of getting others to try, so overall the data reported here is very encouraging with respect to the faculty and student reaction to the proposed approach.

Overall Conclusions

Institutions of higher education have traditionally focused on teaching information formally, and teaching core transferable skills less formally. For example, by asking students to take a position in an essay, or ask questions during a lecture. Movement towards a more formal approach to skill development requires an approach to developing skills that can fit within our constrained university contexts, and one in which skills can be measured accurately, thereby allowing accreditation of the learning. These challenges were referred to earlier as the How? (…do we develop skills effectively) and the How Do We Know? (…when what we are doing is having an impact).

Joordens (2018) proposed a process that claimed to provide a concrete and inherently scalable answer to both questions. In the current report we assessed a number of factors related to this claim. Our conclusion based on our findings is this: The proposed approach absolutely has high potential. On the measurement side of things there is still some work to be done to provide measures that are at the appropriate level of validity (i.e., free from any bias), but the road for doing so is clear and there is every reason to believe it will take us where we need to go. On the development side of things, the core of the proposed process has already been shown to impact core transferable skills (Joordens et al., 2014). The augmentations to this core should enhance learning still, and reactions from faculty and students suggest that the approach can be successfully used across a wide range of learning contexts.

References

American Association of Colleges & Universities (AAC&U) (n.d.). Current VALUE Research. https://www.aacu.org/current-value-research

Bates, A. W. (2000). *Managing Technological Change: Strategies for College and University Leaders.* The Jossey-Bass Higher and Adult Education Series. San Francisco: Jossey-Bass.

Bean, J. C. (2011). *Engaging Ideas: The Professor's Guide to Integrating Writing, Critical Thinking, and Active Learning in the Classroom*. John Wiley & Sons.

Bennett, R. E. (2011). Formative Assessment: A Critical Review. *Assessment in Education: Principles, Policy & Practice, 18*(1), 5–25.

Boud, D. (2013). Enhancing Learning through Self-assessment. Routledge.

Cho, K., Schunn, C. D., & Wilson, R. W. (2006). Validity and Reliability of Scaffolded Peer Assessment of Writing from Instructor and Student Perspectives. *Journal of Educational Psychology, 98*(4), 891–901.

Collimore, L. M., Paré, D. E., & Joordens, S. (2015). SWDYT: So What Do You Think? Canadian Students' Attitudes about peerScholar, an Online Peer-assessment Tool. *Learning Environments Research, 18(*1), 33–45.

Dochy, F. J. R. C., Segers, M., & Sluijsmans, D. (1999). The Use of Self-, Peer and Co-assessment in Higher Education: A review. *Studies in Higher education, 24*(3), 331–350.

Ennis, R. H. (1993). Critical Thinking Assessment. *Theory into practice*, *32*(3), 179–186.

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

Hanrahan, S. J. & Isaacs, G. (2001) Assessing Self- and Peer-Assessment: The Students' Views. *Higher Education Research & Development, (20)*1, 53–70, DOI: 10.1080/07294360123776

Joordens, S. (2018). Developing and Measuring Learning Outcomes at Scale: The Promise of Peer Assessment. In F. Deller, J. Pichette and E. K. Watkins (Eds.), *Driving Academic Quality: Lessons from Ontario's Skills Assessment Projects* (pp. 13–27). Toronto: Higher Education Quality Council of Ontario

Joordens, S., Paré, D., & Collimore, L-M. (2014). *Taking Learning Outcomes to the Gym: An Assignment-Based Approach to Developing and Assessing Learning Outcomes*. Toronto: Higher Education Quality Council of Ontario.

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

Liu, N. F., & Carless, D. (2006). Peer Feedback: the Learning Element of Peer Assessment. *Teaching in Higher Education, 11*(3), 279–290.

MacLean, K. A., Ferrer, E., Aichele, S. R., Bridwell, D. A., Zanesco, A. P., Jacobs, T. L., ... & Wallace, B. A. (2010). Intensive Meditation Training Improves Perceptual Discrimination and Sustained Attention. *Psychological science*, *21*(6), 829–839.

Michael, J. (2006). Where's the Evidence that Active Learning Works? *Advances in Physiology Education*, *30*(4), 159–167.

Nicol, D. J., & Macfarlane-Dick, D. (2006). Formative Assessment and Self-regulated Learning: A Model and Seven Principles of Good Feedback Practice. *Studies in Higher Education, 31*(2), 199–218.

Paré, D. E., & Joordens, S. (2008). Peering into Large Lectures: Examining Peer and Expert Mark Agreement Using peerScholar, an Online Peer Assessment Tool. *Journal of Computer Assisted Learning, 24*(6), 526–540.

Reis, H. T., Maniaci, M. R., Caprariello, P. A., Eastwick, P. W., & Finkel, E. J. (2011). Familiarity does Indeed Promote Attraction in Live Interaction. *Journal of Personality and Social Psychology*, *101*(3), 557.

Rhodes, T. L. (2016). The VALUE of Assessment: Transforming the Culture of Learning. *Change: The Magazine of Higher Learning, 48*(5), 36–43.

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

Simpson, P. M., & Siguaw, J. A. (2000). Student Evaluations of Teaching: An Exploratory Study of the Faculty Response. *Journal of Marketing Education*, *22*(3), 199–213.

Topping, K. (1998). Peer Assessment between Students in Colleges and Universities. *Review of Educational Research, 68*(3), 249–276.

Tseng, S. C., & Tsai, C. C. (2007). On-line Peer Assessment and the Role of the Peer Feedback: A Study of High School Computer Course. *Computers & Education*, *49*(4), 1161–1174.

Woody, R. H. (2012). Playing by Ear: Foundation or Frill? *Music Educators Journal, 99*(2), 82–88.

Higher Education Quality Council of Ontario

An agency of the Government of Ontario