UPDATE ON THE DEVELOPMENT OF ANALYTIC RUBRICS FOR COMPETENCY ASSESSMENT (DARCA)

Gayle Lesmond, Nikita Dawe, Susan McCahan, and Lisa Romkey
University of Toronto
susan.mccahan@utoronto.ca

Abstract *–The shift towards outcomes-based assessment in higher education has necessitated the exploration and development of valid measurement tools. Given this trend, the current project seeks to develop a set of generic analytic rubrics for the purpose of assessing learning outcomes in the core competency areas of design, communication, teamwork, problem analysis and investigation. This paper will provide an update on the original paper presented at CEEA 2015, in which the approach to rubric development for communication, design and teamwork was discussed. The current paper will detail the process of testing the communication, design and teamwork rubrics. In particular, it will report on the progress achieved in shadow testing, where teaching assistants and/or course instructors with grading experience ("assessors") are asked to evaluate samples of student work using selected rows from the rubrics. The results of shadow testing will be presented.*

Keywords: Rubric, Assessment, Learning Outcome, Focus Group.

1. INTRODUCTION

The assessment of learning outcomes has become an increasingly fundamental component of higher education. To this end, the Higher Education Quality Council of Ontario (HEQCO) has commissioned a group of post-secondary institutions to develop and pilot valid tools for assessing learning outcomes. As a member of HEQCO's Learning Outcomes Assessment Consortium (LOAC), the Faculty of Applied Science and Engineering at the University of Toronto is developing a non-discipline specific analytic rubric bank to assess learning outcomes in design, communication, teamwork, problem analysis and investigation.

At the Canadian Engineering Education Association's (CEEA) annual conference in 2015, the Development of Analytic Rubrics for Competency Assessment (DARCA) project was introduced [2]. The purpose of that paper was to describe the process of designing, modifying and validating the rubrics for design, communication and teamwork including expert consultation, performance level definition and descriptor development. Specific attention was also paid to the challenges of rubric development, namely, faculty involvement and the assessment of process-oriented activities such as teamwork.

In the past year, the project has entered its final phase and focused on two primary areas: 1) expert consultation for the development of rubrics in investigation and problem analysis and 2) testing of rubrics in design, communication and teamwork. For the former, a two-round Delphi survey was conducted to identify the specific skills, behaviours or attitudes ("indicators" on the rubric) that were agreed upon by experts as important for assessing investigation and problem analysis. The Delphi method is a systematic survey technique used to generate and synthesize expert opinion. It is typified by four key characteristics; an expert panel, anonymity, controlled feedback, and multiple survey rounds [5] [6] [3]. The results of the Delphi will be used to finalize the rubric criteria for investigation and problem analysis, and support the development of a set of descriptors at different levels of competency. For a detailed description of the Delphi process employed, see [4].

Subsequently, focus group sessions were conducted with graduate students and sessional instructors to determine the utility, clarity, and reliability of the rubrics. The purpose of this paper is to describe this process and to present the preliminary results of this activity (for a detailed analysis of the feedback received on the design rubric, see [1]).

2. TESTING

2.1. Shadow Testing

Shadow testing involves the assessment of student work already submitted and graded for the purpose of feedback generation outside the context of an existing course. It is distinguished from full deployment in which rater assessment determines the students' grades. Focus groups were chosen as a means of generating multiple insightful perspectives through participants' experiences and interaction within the group. The objective of this methodology was to 1) determine if the group most likely to be the primary users of the rubrics, the assessors, understand it in the way that it is intended, and 2) get a detailed view of how specific components of the rubrics could be improved. Focus groups typically comprised two to five assessors.

Shadow testing began with the research team identifying a list of courses in which the target competencies were either taught or assessed. The instructors were then contacted and asked to provide five to 10 samples of student work (ungraded versions), representing a variety of skill levels. Once a sufficient number of student samples was collected, the process of participant recruitment began. Instructors were asked to either forward a recruitment message to their Teaching Assistants (TAs) or provide a list of TAs whom we could contact ourselves. Participation was initially limited to graduate students who held current TA positions. However, as recruitment proved challenging, the inclusion criteria was revised to include any graduate student or sessional instructor with formal grading experience.

It is important to note that the rubrics were used in much the same way that they were intended for use once implemented. In particular, unique rubrics were created for each focus group with relevant rows selected from various competencies based on the assignment instructions. For example, in an assignment that asked students to develop a contract with their clients outlining the activities to be completed in a design project, rows from the design as well as communication rubrics were selected.

In a typical focus group session, participants were first introduced to the project and the purpose of the focus group in particular. They were then asked to complete a short demographic questionnaire where they indicated their primary language, grading experience and background in rubrics. They were then given approximately one hour to assess two to four samples of student work. The objective of this activity was to allow assessors to familiarize themselves with the rubric content before the final activity, where they were asked to complete an exit survey and take part in a group discussion. The following prompts provided a framework for the focus group discussion:

1. Identify indicators that you think should have been assessed for this assignment and were missing from the rubric.
2. Identify indicators that were included in this rubric but were not relevant to this assignment.
    a. Please explain why you think they are not relevant to the assignment
3. Identify all confusing indicators
    a. What was confusing about them? Were there words that were not clear to you?
4. Identify all the descriptors that you found confusing.
    a. What was confusing about them? Were there words that were not clear to you?
    b. Did the descriptor not seem to relate to the indicator?
    c. Did it not work for the assignment?
5. What (if any) rubric training have you had in the past?
    a. What resources were provided?
    b. What was useful? What was not useful? OR What do TAs need from training materials (e.g. sample work that demonstrates performance expectations, definitions of terminology, general tips for using rubrics in assessment)?
6. If you were to use this rubric again, tell us one change that you would make (to the rubric or to the process of using the rubric).

2.2. Teamwork

In our previous paper presented at CEEA's 2015 conference, we detailed the challenges of testing teamwork. In particular, we differentiated between competencies like communication and design, which can be easily assessed using student artefacts, and process-based activities like teamwork, that can only be measured through observation of the entire *process* and not merely the *product*. It was thus impossible for teamwork to be authentically reproduced for the purposes of shadow testing.

To address this issue, we asked TAs responsible for courses in which teamwork was a central component, and where they have regular contact with the teams through weekly meetings and/or tutorials, to use the rubric. This process began with TAs reviewing the entire teamwork rubric and selecting relevant criteria based on the course content. These rows were reviewed throughout the semester before a final assessment was made. TAs then selected three teams that they regularly supervised, and observed them throughout the duration of the semester. After filling out the rubric for each of the teams that they selected, they completed an exit survey (similar to the one used in shadow testing) where they were expected to provide detailed feedback on their experience using the rubric.

Informal discussions were also held with University of Toronto staff who work with teams in curricula and co-curricular environments. In these sessions, the rubric was reviewed with participants identifying the rows most relevant to their course or co-curricular activity. Participants also identified a team with whom they were familiar and used selected rows from the rubric to assess that team. They were asked to explain their thoughts as they rated each of the selected rows on the rubric. These discussions were typically guided by the following questions:

1. Why is the indicator relevant?
2. Why is the indicator not relevant?
3. What do you think the indicator means?

4. RESULTS AND DISCUSSION

To prepare focus group data for analysis, data sheets were produced for each indicator highlighting both the frequency of selections for each assignment, and the qualitative feedback generated from group discussions and exit surveys. The former provided some preliminary quantitative data on the degree of agreement among assessors. Below, we provide a sample data sheet for the communication indicator, "Integrate visuals into text or talk so that they support and clarify key points", used in the assessment of a Project Requirements and Project Management Plan (PRPMP) assignment. The PRPMP is a design assignment in which students are asked to develop a detailed design process plan explaining their proposed activities.

It should be noted that *fails* is distinguished into two categories. The first, *not demonstrated*, represents a failure by omission. For the communication indicator, "Integrate visuals into text or talk so that they support and clarify key points", for example, such a rating would be received if *no* visuals were included in the work. The second, *fails by misconception*, refers to a fundamental misunderstanding of the concept or assignment requirements. In the example of the above indicator, this option would be selected if visuals were disconnected from the key points and/or were poorly located or formatted.

Table 1: Integrate visuals into text or talk so that they support and clarify key points

	F-ND	F-M	B	M	E	No selection
A	6	1	1			1
B	6	1				
C	4	2		2		1
D	6	1				
E	5		3	3		

F-ND: Fails (not demonstrated)
F-M: Fails (misconception)
B: Below expectations
M: Meets expectations
E: Exceeds expectations

Comments on indicator
- Indicator irrelevant for assignment
- The PRPMP could be well-written without any visuals
- Unsure what "visuals" refer to

Comments on descriptors
- *Meets expectations* similar to *meets expectations* for another indicator
- Does not specify what *not demonstrated* means

By way of explanation, table 1 indicates that for the aforementioned indicator in assignment A, *fails-not demonstrated* was selected 6 times, *fails-misconception* and *below expectations* were selected once. No selection was made once.

Once data sheets were completed, a comprehensive analysis was conducted in which the entire research team reviewed all assessor feedback, making decisions on which suggestions to adopt and which to set aside. Typically, analysis would begin with a review of the quantitative data, which provided a limited view on inter-rater reliability. A selection of scores that spanned several performance levels and/or in which no selection was made, indicated greater inconsistency in rating. In the example below, the levels selected for one PRPMP assignment are presented for the design indicator, "Document appropriate engineering design requirements using a suitable model (e.g. goals-functions-constraints or objectives-metrics-criteria-constraints)". The results of this indicator are widely spread with assessors choosing every level of the rubric.

Table 2: Document appropriate engineering design requirements using a suitable model (e.g. goals-functions-constraints or objectives-metrics-criteria-constraints)

F-ND	F-M	B	M	E	No Selection
	2	3	2	4	

Review of qualitative feedback often provided potential reasons for inconsistent grading. Decisions on incorporating feedback were generally based on whether it was apparent that assessors had misinterpreted the rubric or the assignment instructions. In cases where assessors had clearly misinterpreted the rubric or the instructions, a note was made to further define the term or phrase in the training materials to be later developed and used in conjunction with the rubric.

Our review of the focus group data revealed three major themes: 1) rubric terminology 2) differences between levels of proficiency and 3) missing criteria.

4.1 Unclear rubric terminology

Most group discussions focused on the language used to define the indicators and rubric descriptors. Assessors provided several examples of rubric terms that needed further clarification. A recurring example was the communication indicator, "Incorporate evidence from diverse sources". Diversity was originally intended to indicate the range in the type of sources used. A research paper in which only popular websites were employed, for instance, could receive a rating of *fails*. While most assessors commented that "diversity" of sources was not as important as quality or credibility, their discussions revealed tremendous uncertainty around its meaning. Some participants, for example, interpreted diversity as the number of sources used, while others interpreted it as the acknowledgement and use of diverse perspectives. Others simply stated that they were unsure of what it entailed. For example:

> "I mean there was one assignment that stood out to me as having diverse sources, like they had done a YouTube search…other than that, I didn't really see what was meant by diverse."

> "I don't think that diverse is nec[essary]—if by diverse sources it means some websites, some interviews, some books, if that's what that means, which I wasn't even clear on what that mean[t], then that's okay. Although, I still don't think, I mean in today's day and age, let's be honest, they don't use books at all; they're all going to be websites. And so, I mean, I don't know, it seemed out of place."

Not surprisingly, a review of the data sheets revealed significant disagreement among assessors. Below we provide an example data sheet for the PRPMP (a half score was assigned when two levels were selected).

Table 3: Incorporate evidence from diverse sources

	F-ND	F-M	B	M	E	No selection
A		1	3	4		1
B			3	1	2	1
C		2	4		2	1
D			2.5	3.5	1	
E		1		5	5	

Another example that seemed to cause much confusion for assessors was the design indicator, "Create prototypes, models, or simulations to meaningfully explore and analyse design components". Table 4 is an excerpt from the data sheets for the Preliminary Consultant's Report (PCR), an assignment where students are asked to present the results of a streamlined life cycle assessment (SLCA) and an analysis of their functions, objectives and constraints (FOC). Group discussions revealed that participants had misunderstood the terms prototypes, models and simulations. Although it was intended that prototypes also include screenshots of computer-based models that students had developed (and included in many of the samples provided), assessors believed that the indicator was not relevant for the assignment ("I felt the assignment was asking more for descriptions vs. creation of prototypes/models"). This misinterpretation is clearly evident for assignment B where 2/3 assessors did not select a performance level.

Table 4: Create prototypes, models, or simulations to meaningfully explore and analyse design components

	F-ND	F-M	B	M	E	No selection
A			1	0.5	0.5	1
B				1		2

At times, participants were unfamiliar with terms that the research team had assumed were widely across the Faculty. One such example was "engineering design priorities" and "design for X" in the design indicator, "Identify and describe engineering design priorities (i.e. Design for X) and/or social and professional concerns relevant to the problem" (D2B in the quote below). Our discussions revealed that, although common in courses like Praxis in the Division of Engineering Science, these phrases were not common knowledge for students from other programs:

> "I found one that I think was confusing was D2B…I didn't quite get what I was supposed to do there."

4.2 Distinguishing between levels of proficiency

Assessors often found it difficult to differentiate between performance levels, particularly those of immediate proximity, for example, *fails* and *below*, and *meets* and *exceeds*. In particular, they commented that, in the case of *meets* and *exceeds*, differences were signified through the use of overly subjective and "generic" terms:

> "Now I could not decide between *meets* and *exceeds* because there wasn't much context to decide…it was like one or two sentences and I could not decide on the basis of that."

> "This was a subjective assessment more than an objective one. Something was okay and something was greater than okay, and I don't know what is that. So I feel one change that could be made to this is instead of a subjective one, maybe for each column we can include a numeric value from one to 10 so you can have more leeway to mark the assignment, instead of deciding *below* and *meets*."

Participants adopted various strategies for distinguishing between levels, such as comparative assessment ("I will say okay this one's better than that one so this is *exceeds*"), or relied on their their own understanding of what it meant to achieve a certain level of performance rather than using the descriptors provided.

Also problematic for assessors was the progression from one level to the next. They provided examples where there seemed to be a major leap between descriptors. One example identified by several participants was the communication indicator, "Support a claim with use of evidence and reasoning" where the descriptors "jump" from "claims are supported by evidence that lacks credibility" in the *below* level to "claims are well supported by credible evidence" in *meets*.

> "…there were really big gaps between two different sections and they didn't provide a gradual scale…I think that's such a big jump from they're "well supported by credible evidence", so there's no in between…"

They also expressed frustration with the rubric when student work seemed to fall under multiple performance levels. They attributed this to very rigid descriptions provided in the descriptors. They usually resolved this by merely placing checkmarks in all appropriate boxes or by drawing arrows that spanned two proximate levels on the rubric.

4.3 Missing criteria

When asked to provide examples of indicators that should have been assessed for the assignment, but were missing from the rubric, assessors provided a range of responses some of which were believed to be already included in the rubric. For the communication rubric, for example, they noted the absence of criteria that assessed the mechanics of writing.

> "I just noticed there was nothing to do with grammar and I think writing style is a big thing in terms of technical writing."

> "Grammar was also a major issue, not just using the correct vocabulary."

> "I also thought there was a vocabulary box but I didn't see any sentence structure."

It was originally intended that the indicator, "Make clear and appropriate vocabulary choices" would encompass spelling and grammar. However, many participants thought this to be insufficient. The focus group discussion thus highlighted the importance of having an indicator that specifically addressed writing mechanics.

Other suggestions included professionalism, feasibility of the research proposal, display of academic misconduct (plagiarism), understanding of designers and their values, formatting, scoping in design, solution independence etc. In many cases, missing criteria would have been covered by rows from other competencies (such as investigation and teamwork) not included in the rubric used for testing.

5. CONCLUSIONS

The focus groups produced a wealth of information that has enhanced our understanding of the utility and clarity of the rubrics. The next steps of rubric review and modification will address the major themes detailed in the previous discussion, namely, unclear rubric terminology, the ambiguous relationship between performance levels and missing criteria.

First, to employ a more universal rubric vocabulary, less detail will be added to the rubric criteria and descriptors. A more generic rubric will allow instructors greater flexibility to incorporate course-specific terms (such as "design for x") and provide explanations where necessary. Training materials will also be developed, including a glossary that will provide descriptions and examples of key terms (such as "diverse"). The importance of in-depth assessor training will also be emphasized.

To provide a more gradual progression from one performance level to the next, the review process will incorporate a detailed analysis of all descriptors to ensure that the "distance" between each level is consistent. More concrete modifiers will also be included to help users better distinguish between performance levels.

Lastly, instructors will be encouraged to customize the rubrics to suit their needs. This could include, for example, adding missing criteria, modifying existing rubric criteria, using more specific course-based definitions of rubric terms, applying weights, or changing the order in which the levels of performance are arranged (for example, by starting with the highest level of performance).

The immediate next steps of the project will include continued testing of the communication, design and teamwork rubrics. One larger testing session (with 15-20 assessors) will be held so that a more formal and definite assessment of inter-rater reliability can be conducted. Rubrics will be modified to address the issues raised by assessors.

Opportunities for testing outside of the Faculty of Applied Science and Engineering at the University of Toronto and at partner institutions will also be explored.

Acknowledgements

This work is supported by the Higher Education Quality Council of Ontario (HEQCO) Contract #12/13-RFP-006-06. For more information on HEQCO's Learning Outcomes Assessment Consortium, please visit http://www.heqco.ca/en-ca/OurPriorities/LearningOutcomes/Pages/Assessment-Consortium.aspx.

References

[1] Nikita Dawe, Lisa Romkey, Susan McCahan, and Gayle Lesmond, "User testing with assessors to develop modular rubrics for assessing engineering design," in *Proc. ASEE American Society for Engineering Education Conf., ASEE16*, (New Orleans, LA; 26 – 29 June 2016), 12 pp., 2016.

[2] Nikita Dawe, Gayle Lesmond, Susan McCahan, and Lisa Romkey, "Development of analytic rubrics for competency assessment," in *Proc. CEEA Canadian Engineering Education Association Conf., CEEA15*, (Hamilton, ON; 31 May- 3 June 2015), 6 pp., 2015.

[3] Chia-Chien Hsu and Brian A. Sandford, "The Delphi technique: Making sense of consensus," *Practical Assessment, Research & Evaluation*, vol. 12, no. 10, pp. 1-8, 2007.

[4] Gayle Lesmond, Susan McCahan, Nikita Dawe, and Lisa Romkey, "Using a delphi study to develop criteria for an analytic, competency-based rubric," in *Proc. ASEE American Society for Engineering Education Conf., ASEE16*, (New Orleans, LA; 26 – 29 June 2016), 12 pp., 2016.

[5] Harold A. Linstone and Murray Turoff, *The Delphi method: Techniques and applications*. Reading, MA: Addison-Wesley, 1975 (Vol. 29), 616 pp. {ISBN 0-201-04294-0}

[6] Ian Masser and Paul Foley, "Delphi revisited: Expert opinion in urban analysis," *Urban Studies*, vol. 24, no. 3, pp 217-225, 1987.