

DEVELOPMENT OF ANALYTIC RUBRICS FOR COMPETENCY ASSESSMENT

Nikita Dawe, Gayle Lesmond, Susan McCahan, and Lisa Romkey

University of Toronto
susan.mccahan@utoronto.ca

Abstract –This project aims to create and validate generic rubrics that can be used to authentically assess learning outcomes in core competency areas. As these rubrics are intended for ongoing use by students and educators who have had no involvement in their development, ensuring consistent interpretation and application is a challenge. This paper describes the rubric development methodology and progress to-date on Teamwork, Communication, and Design rubrics including the refinement of outcomes, indicators, and descriptors in response to expert feedback. We also discuss challenges that have delayed testing and deployment as well as future steps.

Keywords: Rubrics, Assessment, Competency, Learning Outcomes, Indicator

1. INTRODUCTION

The Higher Education Quality Council of Ontario (HEQCO) is supporting multiple projects aimed at the development of general, non-discipline-specific learning outcomes assessment tools through their Learning Outcomes Assessment Consortium. At the University of Toronto, the Faculty of Applied Science and Engineering initiated the Development of Analytic Rubrics for Competency Assessment (DARCA) project to create valid rubrics that can be used to monitor student achievement of learning outcomes in key competencies and to compare students in different programs and year levels. The final deliverable will be a bank of validated rubric content from which instructors can select rows that are relevant to their particular assignment.

The purpose of this paper is to share our rubric development methodology and report on current progress. There is little documentation of rubric design for post-secondary STEM disciplines or non-discipline-specific rubrics [6] [9]. The literature that is available rarely addresses rubric quality [10]. Thus, we believe an account of our requirements, methodology, results, and challenges

will be of value to other educators interested in defining learning outcomes and creating valid assessment tools in the form of analytic rubrics.

One of the primary challenges faced in this work is the inconsistent terminology in the literature on academic assessment. This necessitates a brief discussion on the cognitive capacities addressed in this paper. The term “competency” is used for the highest level of categorization. The competencies examined in this project are Design, Communication, Teamwork, Investigation and Problem Solving. The second level, “learning outcomes”, is used to specify “cognitive capacities that comprise the competencies” [2]. For example, “Demonstrate the ability to build a collaborative team environment” might be a learning outcome under the Teamwork competency. By virtue of their general nature, learning outcomes cannot be measured directly. Thus, more specific “indicators” define “particular measurable and quantifiable actions or results” [2].

This paper focuses on draft analytic rubric content that is in development for the Design, Communication, and Teamwork competencies. These competencies align with the following working definitions:

- “Design refers to the process to arrive at [...] a representation, plan, or convention for constructing an object or system. Design activity could lead to a piece of art, the blueprint for a building, a scale model, a book, or a survey instrument.” [3]
- “Communication refers to activities involving the transfer of information from one party to another.” [3]
- Teamwork “refers to activities undertaken or performed by parties consisting of more than one agent, where the party is working, at some level, toward a common purpose.” [3]

Analytic rubrics were chosen for this project because they allow discrete (e.g. low, medium, high) evaluation of the quality of the student’s work, which lends itself to analysis; as opposed to holistic rubrics which are purely qualitative. This scoring strategy requires assessors to provide a Likert scale-like evaluation on each criterion or indicator [7]. The descriptors for performance levels

“should define a continuum of quality throughout each [indicator]” and “be detailed enough to limit subjectivity yet concise enough to avoid confusion or ambiguity” [11]. For this project, it is also important to ensure that the descriptors use general language rather than jargon specific to any one discipline.

The rubrics must reliably and validly assess student performance so that their implementation can provide meaningful data and feedback. Inter-rater reliability data should be collected and analyzed. If unsatisfactory, rubric content and/or training resources must be revised to ensure consistency in grading. Evidence of face and content validity is also important to ensure authentic assessment. Face validity means that to its users, a rubric appears to measure the competency for which it is intended. Content validity refers to how well the rubric indicators and descriptors capture the skills and behaviours of a learning outcome, and how well the learning outcomes represent a competency. Our activities so far have focused on research and expert consultation to iteratively define outcomes and indicators relevant to each competency.

In the following sections, we describe in detail the design process we are following to iteratively create, test, refine, and validate rubrics for competency assessment. We also discuss some of the changes we have made to our initial draft rubrics in response to expert feedback as well as the challenges involved in developing generic rubrics that are statistically valid and accepted by faculty.

2. METHODOLOGY

This project aims to develop valid tools to assess learning outcomes. The methodology involves developing a set of general, non-discipline specific rubrics. To validate these rubrics, the research team is consulting academic professionals and competency/assessment experts, and will be hosting benchmarking sessions with Teaching Assistants and focus groups with students.

2.1 Literature Review and Definition of Learning Outcomes

The first phase of this project involved the selection and definition of outcomes and indicators based on a review of the literature on learning outcomes assessment. In particular, a comprehensive list of outcomes and indicators from various initiatives and institutions (such as the Association of American Colleges and Universities (AAC&U) [1] and the Graduate Attributes Committee at the University of Toronto) was generated. This list was distilled to eliminate overlap and duplication, with additions made to address gaps (for example, in the area of reflection). This initial list of outcomes and indicators

provided the starting point from which rubrics were formulated.

In some instances, such as for the AAC&U VALUE rubrics, descriptors were already available from the literature. However, in many instances, the lists of indicators are given without descriptors that express the quality of the student’s work at each level. For each indicator on the compiled list that did not already have descriptors from the literature, a draft set of descriptors was created.

2.2 Expert Review

Central to rubric development process in this project is faculty and expert consultation. For the Communication rubric, for example, the research team has consulted faculty in the Engineering Communication Program (ECP) at the University of Toronto. Having employed rubrics in many of their courses, ECP instructors have tremendous expertise in the development of assessment tools and are thus a valuable resource. Discussions with experts typically center on clarifying the wording used to define outcomes, indicators and performance descriptors. In particular, experts are asked the following questions:

- Are the outcomes and indicators representative of what you know about the competency? Is there anything missing?
- Are the outcomes and indicators adequately organized? In other words, is it accurate to include these indicators under this outcome?
- Is the language used to identify the outcomes and indicators clearly expressed? How can the language be improved?
- Is the language used to describe each performance level clearly expressed? How can the language be improved?

3. RESULTS AND DISCUSSION

To date, rubrics for Design, Communication and Teamwork have been drafted, and are currently under review by faculty and other experts. Following expert review, rubrics will undergo a process of testing to validate the descriptors, check for consistency among graders, and determine whether students understand them in the way that they are intended.

3.1 Rubric Development

In developing rubrics, many factors must be considered, the first of which concerns the number of descriptive levels. The literature on academic assessment provides no consensus on the ideal number and kind of

performance categories to use in a rubric. For example, Popham [7] suggests that a rubric include three to five levels, Callison [4] recommends a limit of four, and Dunbar, Brooks, and Kubicka-Miller [5] found that using three levels resulted in reduced inference and, therefore, a high level of agreement among raters. Using the rubric models published by the Canadian Engineering Accreditation Board (CEAB) as guides, we decided that a four-level rubric (Fails, Below Expectations, Meets Expectations, Exceeds Expectations) was appropriate for this project.

The next step in designing the rubrics involved creating the rubric descriptors. Because the rubrics are intended for use in a variety of courses and programs, the language used to describe performance levels was very broad. Rubric descriptors were initially drafted using the following as a guide:

- **Fails:** Indicator is not demonstrated i.e. work is “not context relevant”, not identified”, “not presented” etc.
- **Below Expectations:** Indicator is “somewhat”/ “partially” demonstrated or is demonstrated “sometimes”
- **Meets Expectations:** Indicator is “generally”/ “mostly” demonstrated
- **Exceeds Expectations:** Indicator is “always”/ “completely”/ “very well” demonstrated

After drafting the rubric descriptors, the research team met with faculty and other experts to obtain feedback on the language used in the rubrics. The key issues raised concerned the absence of a qualitative difference between performance levels. For example, an earlier version of the Communication rubric used the following descriptors for the indicator, “Describe the purpose of communication”

- **Fails:** Purpose is entirely improperly described
- **Below Expectations:** Purpose is somewhat correctly described
- **Meets Expectations:** Purpose is mostly correctly described
- **Exceeds Expectations:** Purpose is completely correctly described

The descriptions are parallel at each level, the only difference being a change in quantity denoted by overly vague terms like “somewhat”, “mostly” and “completely”.

As discussed in our most recent report to HEQCO (dated March 2015), this approach is problematic because:

“...it undermines the clarity and precision of rubric descriptors, making it difficult for stakeholders (instructors, TAs, students etc.) to clearly identify the differences among score levels. A rating of “meets expectations”, for example, is not better from that of “fails” or “below expectations” simply

because a piece of work exhibits more (or less) of some criterion (for example, number of grammatical and spelling errors) but rather because the former is fundamentally and qualitatively different from the latter. Another problem with using comparative language is that it depends heavily on the inference of the grader to determine what behaviour constitutes a particular level. Relying on subjectivity and personal bias compromises inter-rater reliability and ultimately the effectiveness of the rubric. Put differently, the more qualitatively distinct the evaluation levels, the greater the likelihood of consistency in grading.”[3]

More recent iterations of the rubrics have sought to provide enough detail to establish usability while also remaining general enough to ensure their applicability to a range of contexts. The following guide was used to draft the revised rubrics:

- **Fails:** Complete lack of quality and demonstration of opposite quality
- **Below Expectations:** Lacks quality; work must be revised significantly for it to be acceptable
- **Meets Expectations:** Definition of quality. Often, the descriptor at this level mirrors the relevant indicator.
- **Exceeds Expectations:** “Student goes over and above the standard expectations to produce superior work, for example, by influencing others.” [3].

Based on this analysis and feedback from the experts, the descriptors were revised (presented below). In this new version the descriptor levels are far more clear, and rely less on subjective assessment of word meanings for terms such as “somewhat”.

- **Fails:** Purpose is unclear or unrelated to the work
- **Below Expectations:** Purpose is vague, imprecise or not clearly related to the work
- **Meets Expectations:** Purpose is clearly described or stated
- **Exceeds Expectations:** Purpose is explicitly identified in a clear statement that captures the full goal of the document

Expert consultation has also resulted in changes to the outcomes and indicators, in particular, the inclusion of new criteria, the removal of duplicates and modification of wording. For example, the outcome “Deliver an effective visual presentation” (along with its indicators), was incorporated into the most recent version of the Communication rubric after concerns were raised that the original indicator “Incorporate various media effectively” did not sufficiently capture all aspects of visual communication. The outcome “Use advanced (higher-

level) skills” was also problematic. Many experts suggested that this was unnecessary, as it was expressed under the “Exceeds Expectations” level of each indicator. The outcome was thus removed.

Other issues were unique to the nature of the competency being assessed. One such competency was Teamwork. The problem with team-based work, as relayed by the experts, is that it is not easily observable by people external to the group like an instructor or teaching assistant. Unlike other competencies like Design and Problem Solving, Teamwork involves a behavioral process that requires interpersonal interaction and is, therefore, difficult to assess through traditional means. The Association of American Colleges and Universities (AAC&U) offers the following suggestion for assessing teamwork:

“It is recommended that work samples or collections of work for this outcome come from one (or more) of the following three sources: (1) students' own reflections about their contribution to a team's functioning; (2) evaluation or feedback from fellow team members about students' contribution to the team's functioning; or (3) the evaluation of an outside observer regarding students' contributions to a team's functioning. These three sources differ considerably in the resource demands they place on an institution. It is recommended that institutions using this rubric consider carefully the resources they are able to allocate to the assessment of teamwork and choose a means of compiling work samples or collections of work that best suits their priorities, needs, and abilities.” [1]

As an initial assessment of the rubrics, we had planned to use “shadow” testing. In shadow testing, a subset of student work that is submitted and graded for a course is later re-graded using the rubrics to compare the rubric feedback and assessment with the scoring done by the instructor initially. Because of its problematic nature, the Teamwork competency cannot be assessed through this type of testing. As a result, we will be exploring opportunities to integrate the rubric into an existing course.

One of the major challenges that we encountered in developing the rubrics was faculty skepticism. In particular, many instructors were unconvinced that a universal rubric would adequately capture the needs of their assignments. To address these concerns, the research team reminded instructors that the rubric was intended to provide a starting point (not necessarily a ready-to-use rubric) from which they would select relevant rows and make necessary changes. It would then be adjusted to fit the particular assignment. Although this addresses the faculty concern, extensive modification of the rubrics

would undermine the possibility of using the resulting data to compare outcomes across course boundaries. So we also appealed to the instructors for understanding of the larger goal: program assessment.

Another concern centered on the fact that many instructors have already developed their own rubrics that perfectly fit their assignments. In addressing this concern, the research team presented the rubrics as an opportunity for consistent assessment so that student performance can be compared across a range of contexts. Despite their initial reluctance, faculty acknowledged the importance of creating a valid measure of student academic development that can be used to compare learning longitudinally.

3.2 Next Steps: Testing

Following expert review, the rubrics will be piloted in various programs in the Faculty of Applied Science and Engineering. Rather than integrating the rubrics into existing courses, the research team has proposed a strategy of shadow testing. This approach will involve “background assessment” of student work i.e. testing will be conducted after student work has been graded and submitted. As an advantage, this approach allows us to compare the quality of the feedback and the level of assessment given on the rubric to the feedback and assessment provided by the instructor using their current marking practices. Rubric testing is expected to occur in two phases:

Benchmarking sessions with Teaching Assistants (TA's): Groups of teaching assistants will be asked to evaluate a set of samples of student work using the rubrics that we have developed. In cases where no TA's are available (we see this being a potential problem in courses that have already been completed), other TA's with sufficient knowledge in the course material will be asked to participate. TA's will not assess work in its entirety, instead, they will provide feedback on sections of the students' work. For example, they may be asked to assess the introduction of a paper using select rows of the Communication rubric. In addition to assessing samples of student work, the TA's will be asked about their interpretation of the rubrics to determine whether their understanding is consistent with the intended meaning. This process will be repeated using additional sets of projects, assignments and tests to validate the transferability of the rubrics. This will also provide valuable data on the inter-rater reliability of the rubric descriptors, and will be used to further modify the rubrics.

Focus groups with students: Focus group sessions will be conducted with volunteer undergraduate students. Students will be given the rubrics and asked a range of questions to determine whether they understand the rubric

language in the way that it is intended. Discussion questions will include the following:

- What do you understand the descriptor, “Writing is not appropriate for the audience” to mean?
- What are some synonyms for “flow” in the statement, “Create ‘flow’ in communication through organization”?
- What is you think the term “critical thinking” means in the descriptor, “Clear evidence of critical thinking”?

Results from our discussions with students will be used to refine the rubric.

Although testing will begin in the Faculty of Applied Science and Engineering, the research team will explore opportunities to test the rubric in other faculties and institutions.

4. CONCLUSION

Draft rubrics have been created for three (Design, Communication, Teamwork) of the five competencies that will be addressed in this project. The research team has also identified potential courses for rubric deployment. Testing of these rubrics and work on the remaining two competencies (Investigation and Problem Solving) will begin once the process of expert consultation has reached saturation. With the help of instructors, the research team has begun to gather a range of assignment instructions and student work from past semesters to use during the shadow-testing phase.

The rubrics will first be tested in the Faculty of Applied Science and Engineering, after which they will possibly be deployed in other faculties at the University of Toronto and partner institutions. Figure 1 outlines the sequence of project deliverables. It is expected that the full project will be completed by March 2017.

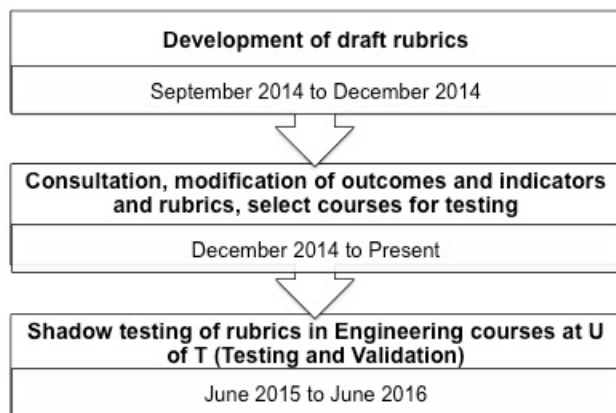


Figure 1. Expected project timeline.

Acknowledgements

This work is supported by the Higher Education Quality Council of Ontario (HEQCO) Contract #12/13-RFP-006-06.

References

- [1] Association of American Colleges and Universities, *VALUE: Valid Assessment of Learning in Undergraduate Education*. <http://aacu.org/value/rubrics/index.cfm>, 2013. [Online; accessed 15Aug13]
- [2] David Beach, *Development of Analytic Rubrics for Competency Assessment: Phase 0 (1PR)*. Higher Education Quality Council of Ontario, 2013, 44 pp. (For a copy of the report, contact Susan McCahan)
- [3] David Beach and Gayle Lesmond, *Development of Analytic Rubrics for Competency Assessment: Phase 1 (3 IPR)*. Higher Education Quality Council of Ontario, 2015, 104 pp. (For a copy of the report, contact Susan McCahan)
- [4] Daniel Callison, “Rubrics,” *School Library Media Activities Monthly*, vol. 17, no. 2, pp. 34-36, 2000.
- [5] Norah E. Dunbar, Catherine F. Brooks, and Tara Kubicka-Miller, “Oral communication skills in higher education: Using a performance-based evaluation rubric to assess communication skills,” *Innovative Higher Education*, vol. 31, no. 2, pp. 115-128, 2006.
- [6] Tara L.S. Kishbaugh, Stephen Cessna, S. Jeanne Horst, Lori Leaman, Toni Flanagan, Doug Graber Neufeld, and Matthew Siderhurst. “Measuring beyond content: A rubric bank for assessing skills in authentic research assignments in the Sciences,” *Chemistry Education Research and Practice*, vol. 13, no. 3, pp. 268-276, 2012.
- [7] W. James Popham, “What’s wrong and what’s right with rubrics,” *Educational Leadership*, vol. 55, no. 2, pp. 72-75, 1997.
- [8] Paul Ralph and Yair Wand, “A proposal for a formal definition of the design concept,” In K. Lyytinen, P. Loucopoulos, J. Mylopoulos, and W. Robinson, editors, *Design Requirements Workshop (LNBIP 14)*. Springer-Verlag, 2009.
- [9] Malini Y. Reddy, “Design and development of rubrics to improve assessment outcomes: A pilot study in a Master's level business program in India,” *Quality Assurance in Education*, vol. 19, no. 1, pp. 84-104, 2011.

- [10] Malini Y. Reddy and Heidi Andrade. “A review of rubric use in higher education,” *Assessment and Evaluation in Higher Education*, vol. 35, no. 4, pp. 435-448, 2010.
- [11] Brian C. Wesolowski, “Understanding and developing rubrics for music performance assessment,” *Music Educators Journal*, vol. 98, no. 3, pp. 36-42, 2012.