Higher Education Quality Council of Ontario

An agency of the Government of Ontario

Development of Analytic Rubrics for Competency Assessment

Gayle Lesmond, Susan McCahan
and David Beach, University of Toronto

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:

Lesmond, G., McCahan, S., Beach, D. (2017) *Development of Analytic Rubrics for Competency Assessment* Toronto: Higher Education Quality Council of Ontario.

Executive Summary

This document describes the development of analytic rubrics for competency assessment project. The purpose of this report is to describe the process of developing a set of general analytic rubrics to assess competencies in design, communication and teamwork, and a set of outcomes and indicators to assess problem analysis and investigation.

The work to develop the rubrics was structured into three main phases. In the first or planning phase, a review of the literature was carried out to create a comprehensive list of learning outcomes in the five competency areas under investigation. A list of more specific, measureable learning outcomes, called indicators, was also compiled. The resulting comprehensive list of learning outcomes and indicators was distilled by removing redundancy between the systems, filling content gaps, and grouping indicators into common learning outcome categories.

In phase two, rubric descriptors for design, communication and teamwork were drafted and modified through consultation with instructors and departmental administrators. The outcomes and indicators for problem analysis and investigation were validated through a systematic Delphi technique and are presented in this report. Work on a set of descriptors for the problem analysis and investigation indicators is ongoing.

The third and final phase involved testing of the design, communication and teamwork rubrics. In the case of design and communication, shadow testing sessions were conducted with graduate students with grading experience (assessors). In particular, assessors were asked to evaluate samples of student work using the rubrics and provide feedback through focus groups. Testing of the teamwork rubric consisted of quasi-implementation with teaching assistants (TAs) and think-aloud sessions with instructors (experts). The objective of testing was to further validate the outcomes, indicators and rubric descriptors and to obtain feedback on how they could be improved.

Detailed data analyses were conducted after all testing was completed. Findings from the analysis of focus group discussions with the assessors revealed the following:

a) Rubric content lacked clarity.
b) Rubric levels lacked sufficient distinction.
c) Rubrics omitted important criteria.

Following analysis of focus group data, the rubrics underwent extensive revision to address the feedback provided by the research participants. The result is a set of revised rubrics that can be applied to a wide range of courses and assessment types.

The work of developing the rubrics revealed three key lessons:

a) The Delphi method is an efficient alternative to one-on-one consultation for this type of work.
b) Assessor training is critical.
c) Teamwork skills are particularly difficult to assess, in part because they require observation and interpretation of behaviour.

Final versions of the communication, design and teamwork rubrics are presented in Appendix A. The indicator sets for problem analysis and investigation are included in Appendix B.

Table of Contents

List of Figures

List of Tables

1. Introduction

In 2012, the revised *Quality Assurance Framework* document, produced by the Ontario Universities Council on Quality Assurance (OUCQA), provided the basis for quality assurance of university academic programs to support both national and international academic graduate mobility (Ontario Universities Council on Quality Assurance, 2012). Related efforts in the same year addressed the need for international requirements for academic qualification methods and standards (Canadian Information Centre for International Credentials, 2012). In the wake of the UNESCO and Lisbon Conventions (Division of Higher Education, 2005; Council of Europe, 1999), the need for quality assurance aimed at university graduates was understood to be of paramount importance to ensure accountability, mobility and recognition of Canadian and international university degree-level programs. Similar questions and concerns also convinced OECD Education Ministers in 2008 to consider the assessment and evaluation of the quality of higher education. These discussions have resulted in an increased interest in the development and use of universal, non-discipline-specific learning outcomes assessment tools at the postsecondary level.

Following on from these initiatives, the Higher Education Quality Council of Ontario (HEQCO) initiated the Learning Outcomes Assessment Consortium (LOAC) project with the goal of developing tools for assessing learning outcomes for cognitive skills considered relevant to graduates of higher learning institutions. As part of this effort, the University of Toronto Faculty of Applied Science and Engineering (FASE) planned to develop valid rubrics leveraging work on learning outcomes assessment accomplished through implementation of the Canadian Engineering Accreditation Board (CEAB) Graduate Attributes system. The primary research question that motivated this project was: Can we create valid and reliable analytic rubric items that provide information on learning outcomes in the selected areas? The goal was to thus develop a set of universal analytic rubrics that would assess learning outcomes in design, communication, teamwork, problem analysis (used interchangeably with problem solving) and investigation. The rubrics were intended to be universal so that they could be used across various contexts (courses, programs, disciplines etc.) and would ideally provide information on learning to all relevant stakeholders, including students, instructors and university administrators.

The work to develop the rubrics was broadly structured into three phases. The first phase (planning**)** involved the following:

- Development of a common framework for defining terminology
- The selection and definition of key areas for assessment
- A review of existing work to create a comprehensive list of specific, measurable skills for each area under investigation

The second phase (rubric construction) consisted of the development of draft rubrics, and expert consultation to revise rubric components. The third and last phase (testing) was conducted with groups of assessors to verify that their interpretation of the rubrics matched the intended meaning and to obtain additional feedback for rubric modification. Accordingly, the report described here is organized around the three main phases of the project. The final sections summarize the lessons learned and provide recommendations for effective rubric deployment.

2. Phase One: Planning

2.1 Terminology

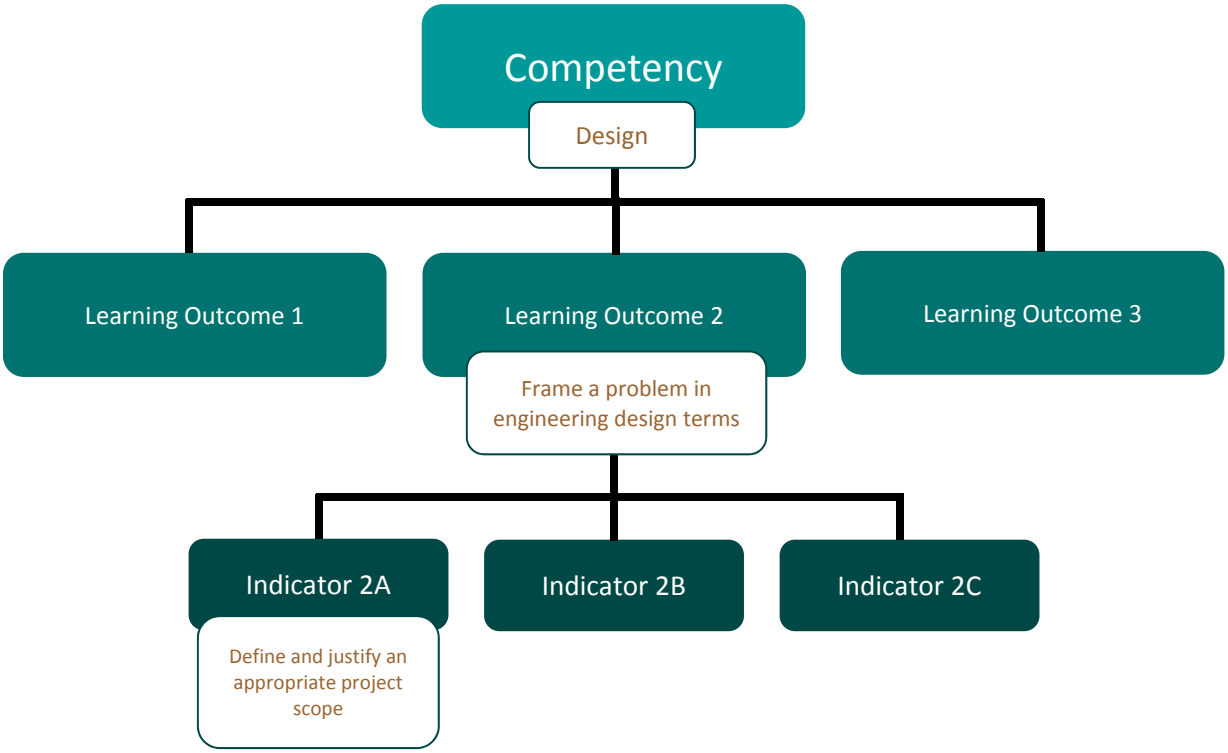
One of the first obstacles that we faced in reviewing the literature stemmed from the differences in terminology associated with learning outcomes assessment. Thus, the first step in rubric development was to create a common framework of reference to describe the goals of the educational process. For the purposes of clarity, we distinguished between three levels of demonstrated capacities. Figure 1 provides a graphical representation of this framework of reference.

- Competencies will be the term used to refer to the most global level of designation or categorization and are typically assessed at the program level. Five competency areas were explored in this research: design, communication, teamwork, problem analysis and investigation.

- Learning outcomes will be the term used to refer to the mid-level of categorization, and specify cognitive capacities that comprise the competencies. They typically describe the learning taking place at the level of a course. For example, “Frame a problem in engineering design terms” might be a learning outcome under the design competency. Learning outcomes are expressed both actively and precisely — that is, they are often expressed with the stem, “The students will demonstrate the ability to…” Although this requirement was not standard in most of the reviewed literature, it followed the largely current consensus, ensuring that this work could be compared and connected to most extant or future learning outcomes projects.

- Indicators will be the term used to refer to the most granular level of learning, specifying particular measurable and quantifiable actions or results that demonstrate specific learning outcomes typical on an assignment or learning module within a course — for example, “Define and justify an appropriate project scope” or “Define appropriate, measurable requirements for evaluating potential engineering design solutions” might be two indicators under the learning outcome described above. An indicator should be specific enough that it can be assessed through a single question or one aspect of an assignment. Specification of indicators follows the same phrasing requirements as specification of learning outcomes. This ensures that indicators can be assessed for reliability, and that measurement precision of indicators can be determined.

Figure 1: Graphical representation of the terminology used in this study



Typically, most authors refer to competencies as "learning outcomes." Throughout this document, we will maintain a consistent use of the terminology described here. Where necessary when referencing the terminology used by others, we will explicitly state our equivalent terminology and use it in place of the original authors' terminology as appropriate.

2.2 Competency Set

A set of competencies was loosely defined before a review of the literature was performed. Initial investigation was guided by the five competencies identified in the research proposal:

- Design
- Communication
- Teamwork
- Problem analysis (or problem solving)
- Investigation

A full definition of these competencies followed the literature review, which explored definitions offered by other researchers in order to arrive at consensus. To guide the literature review, the following rough definitions were used:

- Design refers to the process of arriving at a "... specification of an object, manifested by an agent, intended to accomplish goals, in a particular environment, using a set of primitive components, satisfying a set of requirements, subject to constraints" (Ralph & Wand, 2009, p. 108). The result of design work is a representation, plan, or convention for constructing an object or a system. Design activity can lead, for example, to a piece of art, the blueprint for a building, a scale model, a book (more precisely, a story) or a survey instrument.
- Communication refers to activities involving the transfer of information from one party to another.
- 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.
- Problem analysis (used interchangeably with problem solving) refers to a process that involves defining and executing a solution pathway to the attainment of a goal, usually under constraints. Frequently, the constraints are such that the number of possible solutions is small. The result of problem analysis is a thought product (e.g., a number, an algorithm, a definitive course of action or judgement, a process, a structure or a relationship), that is a synthesis of existing knowledge. Problem analysis activity might give rise to a data-sorting algorithm, a well-defined work-around for an experimental issue or an equation (mathematical or otherwise).
- Investigation refers to the process roughly understood as "research" in the science or engineering sense — that is, activities aimed at increasing knowledge stock, involving hypothesis, experimentation and conclusion.

The five competencies were modeled after CEAB Graduate Attributes (Canadian Engineering Accreditation Board, 2008), and can be mapped to the Ontario Council of Academic Vice-Presidents (OCAV) Undergraduate Degree Level Expectations (UDLEs) (Ontario Universities Council on Quality Assurance, 2012; Council of Ontario Universities, 2011).

2.3 Literature Review

The definition of outcomes and indicators for particular competencies required not only investigation into past learning outcomes assessment efforts, but also investigation into the general definition of competencies. To narrow the scope of the required literature review, the following questions were posed around the topics of learning outcomes, assessment and rubrics:

- What other institutions have undertaken assessment of competencies or learning outcomes?
    - How do other institutions define them? What is considered a "spanning set"?
    - How do the proposed competencies or outcomes compare to the set of competencies or outcomes compiled by the team?
    - Are any case studies available?

- What academic literature exists on the definition or assessment of competencies similar to the tentative competencies roughly adapted from CEAB and OCAV UDLEs?

    - What is the purpose of the work?
    - What are the concepts used (e.g., models, theories, definitions, etc.)?
    - What particular assumptions are made (e.g., age, educational level, discipline)?
    - What data or experiments are used to support the work?
    - What are the conclusions of the work?
    - What are the implications of the work?
    - What general characteristics apply regardless of discipline?
    - How does the work compare to the framework of the Graduate Attributes Committee (GAC) at the University of Toronto?
    - Are any rubrics available? If so, have they undergone validity testing?

The goal of the literature review was to create a list of learning outcomes and indicators in the five competency areas under investigation. The list of learning outcomes and indicators was not intended to provide a complete list that must be used exclusively for every assessment or course, but rather to act as a set of indicators from which those relevant to the course material, learning outcomes and instructor's teaching strategy could be selected and compiled to create a customized rubric for a given assignment.

Below are the primary sources used to develop this initial list of outcomes and indicators. These sources represent commonly used systems of learning outcomes in the literature, such as the VALUE rubrics, NILOA, and the Tuning project, as well as "homegrown" learning outcomes systems. Some context for each is provided here. These systems do not have strengths or weaknesses per se, but merely represent different approaches to the challenge of describing learning goals.

- The Valid Assessment of Learning in Undergraduate Education (VALUE) rubrics were developed by the Association of American Colleges and Universities (AACU), as part of the Liberal Education and America's Promise (LEAP) initiative. The VALUE rubrics are not discipline specific and assess student learning at four distinct levels; the first (benchmark) measures student performance at entrance into university and the final level (capstone) measures performance upon completion of an undergraduate degree. There are 16 VALUE rubrics grouped under three main categories: intellectual and practical skills, personal and social responsibility, and integrative and applied learning. The most relevant VALUE rubrics for our purposes were written communication, oral communication, teamwork, problem solving, and inquiry and analysis (Association of American Colleges & Universities, 2013).

- The Tuning project was developed by the European Union. The name "Tuning" was given to the faculty-driven process to ensure transparency and quality in higher education by stipulating learning outcomes (competencies) for a chosen discipline at a particular degree level. The Tuning project recognizes the variation in disciplines and attempts to allow flexibility under a common framework

and process. The initiative distinguishes between three categories of generic competencies; instrumental, interpersonal and systemic (Bulgarelli, Lettmayr and Menéndez-Valdés, 2009).

- The National Institute for Learning Outcomes Assessment (NILOA) worked with the framework set out in the Lumina Foundation's Degree Qualifications Profile (DQP) to provide institutions with a means of specifying what students should be expected to know and be able to do regardless of discipline (Adelman, Ewell, Gaston, & Schneider, 2011). The DQP specifies five major axes of learning processes: applied learning, intellectual skills, specialized knowledge, broad knowledge and civic learning. It provides a general framework for accommodating many different types of institutions by allowing for relative emphasis on one axis with respect to other axes, for a particular degree. This allows different institutions to specify different bachelor-level requirements, while maintaining an overall quality level.

- The Engineering Graduate Attributes framework was developed at the University of Toronto in response to the outcomes-based accreditation system instituted by the CEAB. The following describes the relevant CEAB Graduate Attributes (i.e., competencies):

    - Design: An ability to design solutions for complex, open-ended engineering problems; and to design systems, components or processes that meet specified needs with appropriate attention to health and safety risks, applicable standards, and economic, environmental, cultural and societal considerations.
    - Communication skills: An ability to communicate complex engineering concepts within the profession and with society at large. Such abilities include reading, writing, speaking and listening, the ability to comprehend and write effective reports and design documentation, and to give and effectively respond to clear instructions.
    - Individual and teamwork: An ability to work effectively as a member and leader in teams, preferably in a multi-disciplinary setting.
    - Problem analysis: An ability to use appropriate knowledge and skills to identify, formulate, analyze and solve complex engineering problems in order to reach substantiated conclusions.
    - Investigation: An ability to conduct investigations of complex problems by methods that include appropriate experiments, analysis and interpretation of data, and synthesis of information in order to reach valid conclusions.

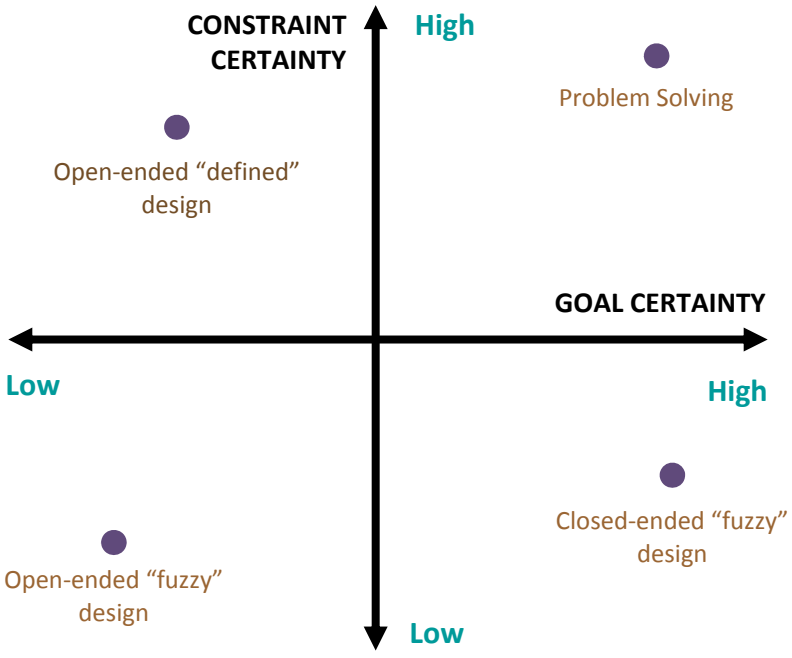
    (Canadian Engineering Accreditation Board, 2008)

Our review of the literature identified that the tentative definitions of the five competencies were sufficient to begin development and validation of assessment instruments. The use of the five competencies also allowed leveraging past experience with the CEAB Graduate Attributes process, in particular, the use of developed indicators to guide outcome and indicator selection for future work.

Noticeably, the selection of a complete set of competencies was heavily influenced by the vision of the particular institution. Institutions that emphasized general liberal education understandably perceive competencies differently than institutions with a more targeted view of education (e.g., professional schools). Thus, we could expect that definitions of competencies by a particular institution or department would be unique to accommodate expected institutional differences in educational emphases.

Analysis of the learning outcome systems revealed confusion over the boundary between problem analysis and design. A number of frameworks for problem analysis were explored including Woods's (Woods et al., 2001) model and Deek's (Deek et al., 1999) problem analysis assessment instrument. An artificial boundary between these two competencies was drawn to keep them as separate entities. The inclusion of both problem analysis and design in the competency set was expected to create some difficulty, particularly in distinguishing between the two, but also in adjusting for differences across programs such as engineering, physical sciences, arts, social sciences or humanities. However, by characterizing design and problem analysis activities along orthogonal axes based on the certainty of goals and constraints, or by use of a similar system to distinguish the two, the breadth of activities covered by the two competencies was much broader and more general (see **Error! Reference source not found.**). By including both problem analysis and design in the compilation of indicators and allowing overlap, the list covered the spectrum rather than trying to find a clear boundary between these two competencies.

Figure 2: Characterization of Design and Problem Analysis/Problem Solving by Constraint and Goal Certainty



Once the literature review was completed, exhaustive lists of indicators were assembled. Through discussion within the research team, the resulting comprehensive list was distilled by removing redundancy between the systems and grouping indicators into common learning-outcome categories. This long list of outcomes and indicators was reviewed and a number of gaps were identified by the research team. In particular, we identified a lack of indicators in the area of competency awareness. Woods and colleagues (2001; 2002) emphasize the importance of awareness of the problem analysis process in students, in particular. This self-awareness generally describes the amount of knowledge that the student has not only about terminology, but also the foundational components of the particular competencies. With an increased awareness, students are better prepared for moments of insight as well as fruitful reflection and critical evaluation of their abilities. In addition, it is expected that heightened awareness leads to students benefiting more from feedback and evaluation of their performance, as they can now relate feedback more strongly to particular areas of improvement. It was thus determined that competency awareness should be assessed as a dimension of each competency.

3. Phase Two: Construction

3.1 Develop Rubric Descriptors

Following compilation of the outcomes and indicators, the development of the specific draft assessment instruments (rubrics) was undertaken for the design, communication and teamwork competencies. Although the literature provided no clear consensus on the ideal number of levels to include in a rubric, the CEAB suggested four-level rubrics in the examples it published for institutions. For this reason, rubric descriptors (see Table 1 for the basic components of an analytic rubric) were developed using a four-level scale (fails, below expectations, meets expectations and exceeds expectations).

Table 1: Basic structure of an analytic rubric

Indicators/Rubric criteria	Scale (Level of Mastery)			
	Fails	**Below Expectations**	**Meets Expectations**	**Exceeds Expectations**
Indicator 1	Descriptor 1a	Descriptor 1b	Descriptor 1c	Descriptor 1d
Indicator 2	Descriptor 2a	Descriptor 2b	Descriptor 2c	Descriptor 2d

The fails level was ultimately subdivided into two separate but related categories as defined in Table 2. This was done to illustrate that if a student fails to complete the necessary work, or if there is insufficient work to assess, the indicator is not demonstrated. However, the student may have completed the work, but in such a way that it lacks sufficient quality or demonstrates a significant misconception that places it in the fails category for the indicator. These two different types of failure to demonstrate have been parsed out in the

rubrics to allow the grader to give clearer feedback to the student as to the nature of the failure. These rough definitions were used to guide rubric development.

Table 2: Performance levels

Fails		Below	Meets	Exceeds
Not Demonstrated	***Misconception***			
Indicator is not demonstrated because of insufficient work to assess	There is a complete lack of quality and/or demonstration of a fundamental misunderstanding of the concept	Lacks quality; work must be revised significantly for it to be acceptable	Definition of quality. Work is acceptable and demonstrates some degree of mastery	Student goes over and above the standard expectations to produce superior work

In developing rubric descriptors, the research team paid specific attention to two key principles. The first concerned the differences in performance levels, in particular, highlighting qualitative rather than quantitative differences. Sometimes, differences in rubric levels primarily signal a difference in quantity; for example, the student is rated as exceeding expectations because he or she had done more. In an early version of our communication rubric for instance, reaching the highest level of performance on the indicator "create 'flow' in communication (oral or written) through organization," meant that all ideas (as opposed to most ideas) were presented in a clear, logical progression. The problem with relying on comparative or quantitative terms such as "none," "most" or "all," is that they undermine the clarity and precision of descriptors, making it difficult for assessors to clearly identify the boundaries between rubric levels. A rating of exceeds expectations, for example, is higher than that of meets or below not only because an assignment exhibits more (or less) of some criterion (for example, number of grammatical and spelling errors) but also because the former is fundamentally and qualitatively different from the latter. To give a more specific example, the way in which the grammatical errors interfere with the understanding of the document goes to a deeper evaluation of the quality of the writing, not just an absolute number of errors. In Table 3 we provide an example row from an earlier draft of the communication rubric in which quantitative terms were used exclusively to differentiate between performance levels. Table 4 outlines suggested changes to clarify the differences between levels.

Table 3: Excerpt from first draft of Communication rubric

Indicator	Fails	Below	Meets	Exceeds
…formulate, in written, visual and/or spoken form, credible and persuasive support for a claim	No claims are formulated or most claims are not credible in themselves or are not supported logically with evidence from reliable sources	Some claims and conclusions are credible in themselves or are supported logically with evidence from reliable sources	Most claims and conclusions are credible in themselves or are supported logically with evidence from fairly reliable sources.	All claims and conclusions are credible in themselves or are supported logically with evidence from highly reliable sources.

Table 4: Suggested changes to excerpt from Communication rubric

Indicator	Fails		Below	Meets	Exceeds
	ND	***Misconception***			
…formulate, in written, visual and/or spoken form, credible and persuasive support for a claim	☐Claims are unsupported	☐ Claims are supported by irrelevant evidence and reasoning ☐ No linkage between evidence and claims	☐ Claims are supported by weak evidence and reasoning ☐ Claims are inconsistently supported ☐ Weak linkage between evidence and claims	☐ Claims are supported by credible evidence, (or evidence used credibly) and sound reasoning ☐ Claims are consistently well supported ☐ Sufficient linkage between evidence and claims	☐ Meets + ☐ Makes claims based on sophisticated handling of evidence and reasoning ☐ Acknowledges opposing evidence and claims and presents a reasonable counterpoint

The second principle of developing rubric descriptions relates to rubric dimensionality. For a rubric to be valid, each row must be unidimensional; this is sometimes referred to as parallelism. Unidimensionality refers to the existence of a single underlying trait across each level along the continuum (i.e., from fails to exceeds expectations). A unidimensional rubric row does not introduce novel constructs or constructs not included in the indicators or rubric criteria. It also does not introduce new constructs as the level changes. Table 5 provides an example of a multidimensional rubric row in which several issues were introduced, including being critical in cell one and demonstrating indifference in cell two, both of which are not parallel to the idea of exhibiting a "positive attitude" identified in the indicator.

Table 5: Example of a multidimensional rubric row

Indicator	Fails	Below	Meets	Exceeds
Convey a positive attitude about the team and its work (e.g., through tone and expression)	☐ Conveys a negative attitude about the team and its work through negative vocal or written tone, facial expressions and/or body language ☐ Often publically critical of the team and its work	☐ Conveys an indifferent attitude to the team and its work	☐ Uses positive vocal or written tone, facial expressions and/or body language to convey a positive attitude about the team and its work	☐ Meets + ☐ Encourages other team members to have a positive attitude about the team and its work

In the revised version of the rubric (see Table 6), a more parallel construction was used; the indicator was modified to reflect the idea of being "constructive." This concept is then more or less repeated at each level of the rubric.

Table 6: Suggested changes to a multidimensional rubric row

Indicator	Fails	Below	Meets	Exceeds
ND*	***Misconception***			
Convey a constructive attitude about the team and its work	☐ Conveys a negative attitude about the team and its work in a way that hinders team cohesiveness and effectiveness	☐ Attempts to convey a constructive attitude about the team and its work but is inconsistent	☐ Consistently conveys a constructive attitude; demonstrates an understanding of when to be positive and appropriately critical about the team and its work	☐ Meets + ☐ Encourages other team members to adopt a constructive attitude about the team and its work

**ND: Not demonstrated*

3.2 Consultations with Content Experts

An important part of rubric validation was expert consultation. If most experts[1] agreed that the outcomes, indicators and rubric descriptors were worded clearly and accurately, and were representative of the relevant competency, one could argue that the rubrics were valid at the face level (i.e., they subjectively appeared to measure what they purported to measure) and at the content level (i.e., they reflected the skills and behaviour that they intended to measure).

Once the rubric descriptors were drafted, the research team conducted one-on-one discussions with content experts in design, communication and teamwork from the University of Toronto's Faculty of Applied Science and Engineering (FASE). Design experts consisted of teaching staff in design-based courses. Communication experts primarily included faculty in the Engineering Communication Program (ECP) and teamwork experts, though very few, included academic professionals with experience in teamwork research. The main objective of these sessions was to obtain feedback on the draft rubrics and to make revisions where necessary. Expert consultations were guided by the following questions:

- Do the outcomes reflect the major aspects of the design/communication/teamwork process? If not, what outcomes/indicators are missing?
- Have the outcomes and indicators been categorized appropriately?
- Are all the indicators necessary for each outcome? Are there any redundancies?
- How can the language used to describe the outcomes and indicators be improved?
- How can the language used to describe the performance levels be improved?

For problem analysis and investigation (for which descriptors are still being developed), expert consultation focused on the outcomes and indicators and was conducted using a different approach: the Delphi method. The Delphi method is a systematic, interactive survey tool used to elicit input from experts and stakeholders through a series of questionnaire rounds (Linstone and Turoff, 1975). It is based on the idea that the collective knowledge of a panel of experts is more valid than that of an individual. The objective of the Delphi study was to identify the skills, knowledge and behaviours that were agreed upon by experts to be important for assessing problem analysis and investigation skills. See Table 7 and Table 8 for the number of Delphi experts by program. The Delphi study was conducted in two rounds. The first survey asked experts to provide a complete list of indicators to assess problem analysis or investigation. In the second round, the indicators compiled from round one were presented to experts. They were asked to rate how likely they were to use each indicator in their teaching (1: very unlikely, 2: unlikely, 3: likely, 4: very likely) and the importance of each indicator to the curriculum as a whole (1: not at all important, 2: somewhat important, 3: important, 4: very important). The list of indicators generated from the Delphi method was then combined with the original list derived from the literature review conducted in the first phase of the project.

1 The experts consulted in this study are faculty who teach courses related to the competencies in question. Both for the Delphi method work reported here, and the less formal consultations, the experts are people who have multiple years of experience creating assessments (e.g., tests, assignments), and providing feedback and marks to students at the undergraduate level at the University of Toronto.

Table 7: Investigation experts by program

Program	Number of Experts
Biomaterials and Biomedical Engineering	4
Chemical Engineering	1
Electrical and Computer Engineering	2
Materials Science and Engineering	1
Mechanical and Industrial Engineering	2
Physics	1

Table 8: Problem analysis experts by program

Program	Number of Experts
Chemical Engineering	5
Civil Engineering	3
Electrical and Computer Engineering	2
Mechanical and Industrial Engineering	5

Feedback received through expert consultation led to significant changes in the phrasing and organization of the outcomes, indicators and descriptors. For the final list of outcomes and indicators for problem analysis and investigation, see Appendix B. Once the rubrics for design, communication and teamwork were modified based on expert feedback, they were considered ready for testing.

4. Phase Three: Testing and Analysis

4.1 Shadow Testing

The rubrics were evaluated through shadow testing, an approach in which student work that had already been submitted and graded was reassessed by raters (or assessors) outside of the context of an existing course. The assessor group (focus group) provided feedback on the rubrics. The assessors used in this process were graduate students who all had previous experience as teaching assistants in courses that relate

to the competencies being assessed. The objective of shadow testing was to assess the utility, clarity and inter-rater consistency of the rubrics.

Shadow testing was used, rather than implementation in an actual course, for two reasons. First, it was useful to have several assessors evaluating the same piece of work so inter-rater reliability could be examined. In a normal grading situation, each piece of student work is evaluated only by one rater, and this would not allow for inter-rater testing. Second, by using student work that had already been graded we were able to select examples at the high, middle, and low end of the quality scale. This was useful for examining the effectiveness of the range of descriptors across all levels of the rubrics. The goal, ultimately, is to have more polished rubrics to deploy in an actual course situation.

The first step in shadow testing involved identifying courses in which the competencies were taught and assessed. Student samples representing a range of proficiencies were then collected from these courses. Once relevant courses were identified and sample assignments compiled, graduate students with grading experience were recruited to serve as assessors.

Shadow testing was conducted through a series of focus groups. Focus groups provided an opportunity for the researchers to obtain in-depth information through concentrated discussion and group interaction. In the focus groups, participants completed an assessment activity where they assessed similar samples of student work using a rubric developed from rows of the communication and/or design rubrics. Participants also completed a survey that asked their feedback on the indicators, descriptors and overall rubric. The final activity was a discussion guided by the following questions/prompts:

- Identify all the indicators that you think should have been assessed for this assignment and were missing.
- Identify all the indicators that were included in this rubric but were not relevant to this assignment.
    - Please explain why you think they are not relevant to the assignment.
- Identify all confusing indicators.
    - What was confusing about them? Were there words that were not clear to you?
- Identify all the descriptors that you found confusing.
    - What was confusing about them? Were there words that were not clear to you?
    - Did the descriptor not seem to relate to the indicator?
    - Did it not work for the assignment?
- What (if any) rubric training have you had in the past?
    - What resources were provided?
    - What was useful? What was not useful?
- 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)?
- If you were to use this rubric again, tell us one change that you would make.

A total of 17 focus group sessions were conducted with 46 assessors to test the design and communication rubrics.

For teamwork, however, the shadow testing approach could not be employed. Unlike communication and design, where products of student work such as essays and design reports could be easily gathered, teamwork is essentially performance based and could only be assessed by those able to observe the process. Thus, in lieu of shadow testing, the research team adopted two methods to assess teamwork. The first, quasi-implementation consisted of graduate students 1) observing three teams that they regularly supervised (acting as a TA) throughout the semester, 2) using the teamwork rubric to assess those teams, and 3) providing detailed feedback based on their experience using the rubric. A total of six graduate students participated in this process. The second method, think-aloud sessions, involved discussions with instructors responsible for courses and co-curricular activities in which teamwork was a central component. In these sessions, instructors were asked to provide detailed walk-throughs of typical assessment activities using the teamwork rubric. In particular, they were asked to identify all relevant indicators for their assignments, explain why they were relevant (or why they were not relevant) and to describe how they interpreted particular items on the rubric. Three instructors participated in the think-aloud sessions.

The following table outlines the total number of times that the rubrics were tested for each program or department in the Faculty of Applied Science and Engineering. As indicated, rubric testing was conducted in about half of the Engineering programs offered at the University of Toronto.

Table 9: Rubric testing by program

Program/Department	Total number of times tested
Aerospace Studies	N/A
Biomaterials and Biomedical Engineering	N/A
Chemical Engineering and Applied Chemistry	N/A
Civil Engineering	N/A
Electrical and Computer Engineering	1
Engineering Science	6
First Year Program	7
Institute for Leadership in Undergraduate Education	2
Materials Science	N/A
Mechanical and Industrial Engineering	4

4.2 Data Analysis

Two fundamental questions drove the analysis process: 1) How can assessor feedback be used to modify the rubrics? 2) What can assessor ratings tell us about the inter-rater reliability of the rubrics? Accordingly, data sheets were created for each indicator and assignment. These highlighted assessors' quantitative ratings and qualitative feedback on the indicator, descriptors and overall rubric as obtained through focus group discussions and surveys. These data sheets were then analyzed in detail by the research team.

Analysis of testing data began with a review of assessor ratings for each indicator. If assessors selected multiple performance levels (i.e., their selections collectively spanned three to four levels of the rubric), this suggested an inconsistent rating and possible poor inter-rater agreement. A review of the qualitative feedback was then conducted. The qualitative data often explained why ratings for a particular indicator appeared inconsistent. Figure 3 and Figure 4 are example data sheets for select indicators from the communication and design rubrics respectively.

The numbers in the table represent the number of assessors who rated the piece of work at the given level. For example, in Figure 3, two assessors found that the work meets expectations on the Ethics 3 indicator. A half score was used when the assessor indicated that the assignment fell between two adjacent levels. For example, in Figure 3, one assessor found that the work falls between below expectations and meets expectations on the Ethics 1 indicator.

In Figure 3, assessor ratings for the communication indicator, "Provide a clear introduction that states the topic and previews the material" are presented for "Ethics and Technology," an assignment in which students were asked to select and analyze a piece of technology. The data sheet indicated that, with the exception of assignment five, scores were fairly consistent.

In Figure 4, assessor ratings for the design indicator "Accurately state the engineering design problem and summarize key details (interpret a problem statement if provided)" are presented. The scores were assigned for the "Project Requirements and Project Management Plan" (PRPMP), where students were asked to provide a comprehensive plan of their proposed design activities. In contrast to Figure 3, the scores were widely spread, an indication of inconsistent grading and poor inter-rater agreement. The comments suggested that this inconsistency was likely due, at least in part, to assessors' inability to distinguish between performance levels and the multidimensional nature of the indicator.

For teamwork, consistency in ratings was not assessed as the graduate students selected indicators based on the needs of their specific teams. Thus, analysis of teamwork data focused on the qualitative feedback provided by assessors in the surveys and think-aloud discussions. Figure 5 is an example data sheet for the teamwork indicator, "Accurately report on other team members' contributions to the team activity."

The frequency of selection for each indicator was also reviewed as it provided some insight into the rubric criteria viewed as most important for teamwork assessment. The most selected (i.e., selected 15 or more times) teamwork indicators among TAs were:

- T1A Communicate respectfully with team members, using appropriate tone, body language and facial expressions
- T1G Offer new suggestions that build on the ideas of others
- T1H Articulate the merits of alternative ideas from others
- T2D Attend team meetings regularly and on time
- T2E Complete all assigned tasks by (external and/or internal) deadline
- T2F Produce quality work that advances the team
- T2G Make individual contributions that advance the project, either directly or indirectly

Once all data sheets were reviewed, rubric modifications were made to address assessor feedback.

Figure 3: Sample data sheet for "Provide a clear introduction that states the topic and previews the material"

	Fails		Below	Meets	Exceeds	No selection
	N-D	*M*				
Ethics 1			0.5	4.5		
Ethics 2			1.5	1.5		
Ethics 3			1	2		
Ethics 4		1.5	3.5			
Ethics 5		2	0.5	0.5		
Ethics 6			5			

Comments on the indicator

- Difficult to write introductions and conclusions that meet the requirements outlined in the rubric
- Two factors are confounded – topic & outlines (preview)

Comments on the descriptors

- Inconsistent use of checkboxes
- Inconsistent use of language across rows
- Subjective language in "introduction captures and maintains the reader's interest" in exceeds expectations level

Figure 4: Sample data sheet for "Accurately state the engineering design problem and summarize key details (interpret a problem statement if provided)"

	Fails		Below	Meets	Exceeds	No selection
	N-D	*M*				
PRPMP1			5	3.5	0.5	
PRPMP2			4	1	1	1
PRPMP3	1	1	4	2.5	0.5	
PRPMP4		1.5	3.5	2		
PRPMP5			5	4	2	

Comments on the indicator

- Should be broken down into multiple indicators

Comments on the descriptors

- The descriptor for below expectations applies in many cases to a failing grade
- Difficult to distinguish between meets expectations and exceeds expectations
- Meets expectations is too general

Figure 5: Sample data sheet for "Accurately report on other team members' contributions to the team activity"

Feedback on the indicator

- Confusing; what is meant by "report"?

Feedback on the descriptors

- The meets expectations level is problematic as students often find it difficult to identify specific examples — they usually do not pay attention to this throughout
- Perhaps add an indicator that assesses students' ability to "capture evidence of team practices"

5. Results and Discussion

This project sought to develop valid, analytic rubrics to assess learning outcomes in five key competency areas: design, communication, teamwork, problem analysis and investigation. The results of this work are catalogued in Appendices A and B. Appendix A contains the final version of the rubrics for design, communication and teamwork. Appendix B contains the indicators list for problem analysis and investigation that was compiled and modified using the Delphi process. These materials are also available on a public website: https://sites.google.com/site/uoftlearningoutcomesproject/

The primary goal was to create rubrics that were universal and customizable; universal in the sense that they could be used to assess student learning across various contexts, and customizable so that instructors could adjust them to suit their unique needs and preferences. The team started by developing descriptors for the indicators on design, communication and teamwork. To obtain feedback on how the rubrics could be improved, the researchers conducted a series of structured group discussions with graduate students and instructors (assessors). These group discussions provided tremendous insight into assessors' perceptions about the clarity, relevance and completeness of the rubrics and revealed that significant changes needed to be made before the rubrics could be deployed. The key findings from our analysis of the focus group sessions are summarized below:

- Rubric terminology lacked sufficient clarity: Assessors were asked to identify indicators and/or rubric descriptors that they found difficult to understand. Often, they identified parts of the rubric that were too general such as the communication indicator, "Select appropriate content and approach for audience and purpose," or specific terms that were unclear. An example of the latter was the term "diverse" in the communication indicator, "Incorporate evidence from diverse sources," wherein assessors were either unsure of its meaning ("I wasn't even clear on what that mean[t]") and/or provided multiple, conflicting interpretations. Another example was the design indicator, "Extract and integrate information from stakeholders and other appropriate (reliable, diverse, credible) sources to enhance understanding of the problem," where assessors relayed their confusion about the term "stakeholders" ("Does stakeholders refer to different groups who use it…or the different parties involved in the use of a single application"). To address the incomprehensibility of rubric terminology, the research team identified all unclear indicators and descriptors, and modified them to enhance simplicity and clarity. In the example of the communication indicator mentioned above, for instance, the indicator was ultimately changed to "Incorporate evidence from a *range* of different sources" to more clearly reflect its intended meaning (i.e., the extent of the type of sources used). Footnotes and examples were also added to clarify terms likely to be misinterpreted along with a comprehensive rubric guide for instructors and TAs (see Appendix C).

- Performance levels lacked sufficient differentiation: Assessors often found it difficult to easily distinguish between proximate levels of the rubric, for example between "fails" and "below" or between "meets" and "exceeds" ("And for 'meets' and 'exceeds', I could not really tell the

difference"). Furthermore, they told us that "exceeds" was often defined using overly subjective language which did not adequately distinguish it from "meets" ("I found that also the difference between 'meets' and 'exceeds' wasn't clear…to say something is sophisticated is very subjective, is not very specific"). To improve the distinction between performance levels, the research team conducted an extensive review of the rubrics and, with input from the entire team, added more distinct qualifiers to more clearly separate performance categories.

Similarly, assessors also indicated the importance of flexibility when selecting performance levels:

> I oftentimes have difficulty to decide between one category, for example between "below" and "meets". Sometimes the description is clear and sometimes not. I had this in my mind when I used to do marking […] Let's say you have a line and here's a fail and here you have exceed expectation, maybe you can have an arrow of where to go, so you can change, so you can actually have something between categories.

The rubrics were designed so that instructors could, for example, add spaces for comments or items to allow assessors the ability to not only select within, but also between performance levels. The researchers thus emphasize the importance of instructor customization.

- Rubrics omitted important criteria: Assessors provided examples of indicators that they believed were necessary for assessment, but not included in the rubric. These included, for example, writing mechanics and professionalism in communication, solution independence in design, and the distribution of workload among team members in teamwork. In many cases, indicators believed to be missing were included in the rubric (though not in the assessment for a particular focus group session for practical purposes). In cases where an important indicator was indeed missing from a competency, it was ultimately added to the rubric with its accompanying descriptors. For example, following feedback from assessors regarding missing content from the communication rubric, the researchers added several indicators to specifically assess the abstract or executive summary of a report and grammatical accuracy and clarity.

The aforementioned feedback provided by assessors led to significant improvements to the rubrics. Based on these key findings (and the experience of conducting the project), the research team has identified key lessons that might be useful to others interested in developing and implementing universal tools for learning outcomes assessment. The following list describes the lessons learned in developing the rubrics for design, communication and teamwork, and in finalizing the list of outcomes and indicators for problem analysis and investigation.

- The Delphi method can be an efficient alternative to one-on-one consultation: The Delphi method was used to identify and refine a comprehensive list of outcomes and indicators that would inform the rubric criteria for problem analysis and investigation. In the case of design, communication and teamwork, the Delphi method was not used and expert consultation was mainly conducted through

in-person, one-on-one sessions. Although this approach provided valuable data that contributed to significant rubric modification, it meant that only a few experts were able to participate in the process. For example, expert consultation for communication took place over a six-month period, but resulted in meetings with only five experts. In contrast, the Delphi study occurred over a similar period with 10 experts for investigation and 12 for problem analysis. Thus, the Delphi method allowed us to consult with a larger number of experts over a reasonably short period of time.

- Assessor training is critical: As described above, the focus group discussions revealed many misconceptions and points of confusion regarding rubric terminology. Participants' tendency to misinterpret rubric terms suggests that the rubric, by itself, is insufficient. More specifically, we suggest that benchmarking sessions be conducted to ensure that the rubric is interpreted in the same way by everyone assessing student work. Training will help assessors better understand the assignment instructions, the objective of the rubric, and specific terms in the rubric, thus resulting in more consistent grading. To address this, a draft training manual has been developed.

- Teamwork is difficult to assess because of the complexity of human behaviour: Teamwork presented tremendous challenges to rubric development. As previously mentioned, teamwork is a process-oriented competency rooted in social interaction, which makes it difficult to observe by those outside of the team. In fact, many of the indicators included in our own rubric are almost impossible to evaluate by external assessors alone. An example of this is the teamwork indicator, "Provide assistance to team members as needed or required." To address this issue, we recommend that teamwork be assessed using multiple sources, in particular, students' own reflections, reflections from their teammates, and evaluations from external observers (Association of American Colleges and Universities, 2013). A triangulated assessment will allow for each data source to contribute different, unique insights into the teamwork process.

6. Conclusion

Learning outcomes assessment has become an increasingly integral part of educational policy at the postsecondary level. Its appeal lies in its ability to provide concrete information on the quality and effectiveness of higher education. In this regard, the current project sought to develop a universal rubric bank to assess student learning in five key areas, namely, design, communication, teamwork, problem analysis and investigation. The process of rubric development began with a review of the literature on learning outcomes assessment to identify a list of outcomes and indicators for each competency. This was followed by rubric construction where experts were consulted, descriptors for design, communication and teamwork were drafted, and the outcomes and indicators for problem analysis and investigation were refined. In the third and final phase, assessors and instructors provided their feedback on the rubrics through focus group discussions, and "quasi-implementation" in the case of teamwork. Ultimately, the work of the project resulted in a set of validated, analytic rubrics to assess design, communication and teamwork (see Appendix A), a list of validated outcomes and indicators for problem analysis and investigation (see

Appendix B) and a series of papers to help users in developing their own customized rubrics using the rubric bank (see Appendix C).

A key takeaway from our work is that rubrics are only as good as the training that accompanies them. Therefore, we stress that assessors undergo training on how to use the rubric to effectively assess student work. For teams of instructors evaluating work in the same course, or for program level outcomes, benchmarking is critical to inter-rater reliability. In benchmarking, groups of assessors evaluate student work samples, some of which clearly represent each level of the rubric and, and others which are not as straightforward. Assessors work through the samples together. This allows them to not only develop a shared understanding of rubric terminology but also the quality of work that exemplifies each performance level, thus facilitating greater consistency in grading. In addition to benchmarking with assessors, we recommend that the rubric bank be integrated into existing learning management engines (LME) or other educational technology tools that are used to manage assessment and feedback. The educational technology integration will ensure that the institution — not only the students — will receive detailed information on student learning and performance and, thus, help make transparent the value added elements of higher education.

The most up-to-date versions of the rubrics are posted on a publically available website: (https://sites.google.com/site/uoftlearningoutcomesproject/). As further work is done to refine these instruments, updated versions will be posted. For a list of publications related to this study, please see Appendix C.

References

Adelman, C., Ewell, P., Gaston, P., & Schneider, C. G. (2011). The Degree Qualifications Profile. Retrieved from http://degreeprofile.org/advantage/publication/The_Degree_Qualifications_Profile.pdf

Association of American Colleges & Universities. (2013). *VALUE: Valid Assessment of Learning In Undergraduate Education*. Retrieved from http://aacu.org/value/rubrics/index.cfm

Bulgarelli, A., Lettmayr, C., & Menéndez-Valdés, J. (2009). *The Shift to Learning Outcomes: Policies and Practices in Europe.* European Centre for the Development of Vocational Training.

Canadian Engineering Accreditation Board. (2008, September). *Accreditation Criteria and Procedures 2008.* Retrieved from http://www.engineerscanada.ca/e/files/report_ceab_08_txt_only.pdf

Canadian Information Centre for International Credentials. (2012). *Pan-Canadian Quality Standards in International Academic Credential Assessment – Phase II: Final Report.* Council of Ministers of Education.

Council of Europe. (1999). Convention on the Recognition of Qualifications Concerning Higher Education in the European Region.

Council of Ontario Universities. (2011). *Ensuring the Value of University Degrees in Ontario.* Council of Ontario Universities.

Deek, F. P., Hiltz, S. R., Kimmel, H., & Rotter, N. (1999). Cognitive Assessment of Students' Problem Solving and Program Development Skills. *Journal of Engineering Education , 88*(3), 317–326.

Division of Higher Education. (2005). *Guidelines for Quality Provision in Cross-Border Higher Education.* United Nations Educational, Scientific and Cultural Organization.

Linstone, H. A., & Turoff, M. (Eds.). (1975). *The Delphi Method: Techniques and Applications* (Vol. 29). Reading, MA: Addison-Wesley.

Ontario Universities Council on Quality Assurance. (2012). *Quality Assurance Framework Technical Report.* Council of Ontario Universities.

Ralph, P., & Wand, Y. (2009). A proposal for a formal definition of the design concept. In K. Lyytinen, P. Loucopoulos, J. Mylopoulos, & W. Robinson (Eds.), *Design Requirements Workshop* (pp. 103–136). Berlin: Springer-Verlag.

Woods, D. R., Kourti, T., Wood, P. E., Sheardown, H., Crowe, C. M., & Dickson, J. M. (2001). Assessing Problem-Solving Skills–Part 1: The Context for Assessment. *Chemical Engineering Education , Fall*, 300–307.

Woods, D. R., Kourti, T., Wood, P. E., Sheardown, H., Crowe, C. M., & Dickson, J. M. (2002). Assessing Problem-Solving Skills–Part 2: Assessing the Process of Problem Solving. *Chemical Engineering Education, Winter*, 60–66.

Higher Education Quality Council of Ontario

An agency of the Government of Ontario