Higher Education Quality Council of Ontario

An agency of the Government of Ontario

HEQCO's Guide to Developing Valid and Reliable Rubrics

Jess McKeown and Danielle Lenarcic Biss

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:

McKeown, J. & Lenarcic Biss, D. (2018). *HEQCO's Guide to Developing Valid and Reliable Rubrics.* Toronto: Higher Education Quality Council of Ontario.

Rubric Development Process

1

REVIEW EXAMPLE RUBRICS

- Identify the type of rubric needed
- Base the design on existing tools to increase validity

2

CREATE DRAFT RUBRIC

- Conceptualize the rubric's intended purpose, structural elements and construct descriptions
- Consider seeking insight from applicable experts

3

TEST RUBRIC'S VALIDITY

- Test relevant types of validity
- Engage experts and front-line rubric users in recursive rounds of rubric testing for feedback

4

TEST VALIDATED RUBRIC'S RELIABILITY

- Test relevant types of reliability
- Engage front-line rubric users in testing of rubric to calculate reliability scores

5

IMPLEMENT RUBRIC

- Assess potential for training front-line rubric users
- Provide anchor assignments
- Strip student identifiers to maintain confidentiality, validity and reliability

Introduction

In academic assessment, there are two general categories of tests to measure students' learning**:** objective testing, in which each test item has only one right answer (e.g., multiple-choice, matching, fill-in-the-blanks questions), and subjective testing, where there is no single right answer but rather constructed responses (e.g., short answer, essay, oral presentation). Although both forms of assessment are valuable for measuring students' learning, assessors' biases and inconsistencies make subjective tests significantly more difficult to assess, which diminishes reliability and sometimes validity.

Rubrics attempt to mitigate this problem; they are thought to bring a level of objectivity to grading subjective assessments. A rubric is a tool that seeks to both guide and assess students' work by clearly articulating the criteria to be measured (i.e., constructs or dimensions), which can include skills, knowledge, attitudes and/or behaviours. These criteria are then further described to align with increasing levels of achievement (i.e., performance levels), which might be assigned some numerical value to calculate a student's total grade on a given assignment. Without specific training in rubric development, it can be difficult for instructors to design rubrics that appropriately capture the criteria they are trying to measure. Fortunately, the Association of American Colleges & Universities (AAC&U) has created 16 generic "VALUE rubrics" for assessing essential learning outcomes in higher education. These rubrics are now considered a gold standard and have proven extremely useful for individuals unsure of where to start when developing rubrics for their specific context.

The purpose of this document is to give an overview of important considerations and provide a suggested order of process for those looking to develop and/or validate scoring rubrics for subjective assessment. This process is not the only valid approach to rubric development. The goal of this guide is to stimulate thinking and encourage robust approaches to assessment.

Table 1: Sample of a traditional rubric using generic performance levels

		Performance Levels *(Score with percentage, number or letter grade for each level)*					**Criteria Weight** *(examples)*	**Student Score**
		Advanced	**Proficient**	**Developing**	**Beginning**	**Inadequate**		
Criteria	**Construct 1**						**30%**	**Score x 0.3**
	Construct 2						**20%**	**Score x 0.2**
	Construct 3						**10%**	**Score x 0.1**
	Construct 4						**10%**	**Score x 0.1**
	Construct 5						**30%**	**Score x 0.3**
		↑ Descriptions of constructs at each performance level ↑					**100%**	**Total score**

Rubric Development Process

STEP 1: REVIEW EXAMPLE RUBRICS

An ideal first step is to begin with a thorough review of existing, validated rubric tools specific to the skill or subject area. If rubrics assessing the skill or subject area of interest are uncommon in the literature, or if the vision for the rubric structure is relatively unique, the development process may more likely be starting from scratch. On the other hand, pre-validated and/or generic rubrics, such as the VALUE rubrics, may provide an ideal benchmark for those looking to assess skills or subjects that are commonly evaluated. Depending on the resources available for the rubric development process and the context of use, existing rubrics may be suitable for either (a) direct implementation in their original form or (b) modification to foster greater alignment with a specific assignment or task. Although direct implementation of an existing rubric may seem less resource-intensive at the outset, rubric developers should consider the amount of training required for assessors, as generic rubrics tend to be less reliable than context-specific rubrics if training is insufficient.

STEP 2: CREATE A DRAFT RUBRIC

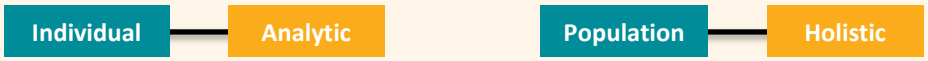
A) Identify the intended purpose of the rubric

- Clarify the portion of the student work being assessed: process versus product (or both).[1]
- The purpose of the assessment will drive what type of rubric is created: individual versus population. Rubrics designed to provide feedback to individual students will likely be more detailed and may lack technical quality for large-scale use, whereas rubrics designed to make inferences about a population may not capture the performance of individuals.

B) Consider various elements for rubric design

- Choose a structure: analytic versus holistic. Holistic rubrics require raters to make an overall judgement about the quality of a student's work, whereas analytic rubrics contain several dimensions for assessing student work.

Figure 1: Relationship between a rubric's level of interpretation and structure



Analytic rubrics are often more accurate when results are interpreted at the individual level, such as in classroom settings to identify students' strengths and learning needs. Holistic rubrics are thought to be easier and cheaper to implement for large-scale assessments that require population-level interpretations.

Source: Adapted from Jonsson & Svingby (2007); Brophy (2012)

[1] Process assessments examine the steps taken to complete a task, whereas product assessments examine the output or outcome of the task.

- Create scoring levels. Decide whether levels will refer to the quality of the work or the level of development. Select the number of levels (between three and five is most common[2]) and decide whether they will be paired with numerical values (see Tables 1 and 2).

- If translating to traditional grades, assign weights to scoring criteria. Having more levels allows for greater variability in students' grades, which may be preferable.

- Consider the minimum score/level for a student's work to be deemed passable. Otherwise, rubric-users may assign a passing grade to students who did not even attempt the exercise.

Table 2: Sample analytic rubric using generic descriptors for quality of work

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.

Source: Lesmond, McCahan & Beach (2017)

C) Compose clear and accurate construct descriptions

- Use clear terminology and ensure that all constructs are measurable.

- Create levels that are unidimensional (i.e., each row should only assess one construct, and new constructs should not be introduced as the levels increase).

- Consider aligning levels with learning outcome taxonomies such as Bloom's, as was done for the VALUE rubrics, or different levels of accuracy, originality, content coverage, effort or errors (AAC&U, 2017).

- Align language with the rubric's purpose, ensuring it is appropriate for both assessors and students. Some scholars recommend using subjective**,** qualitative language (e.g., sophistication, credibility, consistency and relevance of work). However, objective**,** quantitative language (e.g., all, most, none) may be useful for rubrics intended for population-level interpretations.

D) Consider seeking insight from applicable experts

- If time and project budget allow, consult subject-matter and/or assessment experts.

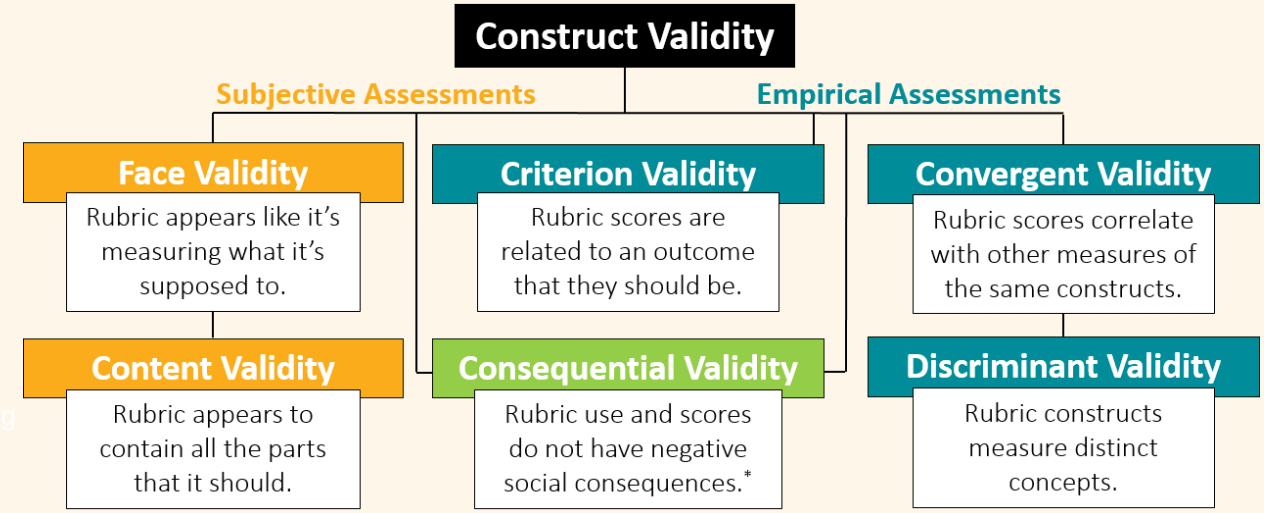
[2] Instructors may be accustomed to traditional five-letter grading (A, B, C, D and F), and thus prefer rubrics with five levels (Kapelus, Miyagi & Scovill, 2017).

STEP 3: TEST THE RUBRIC'S VALIDITY

A) Understand the relevant types of validity

- Ensure the rubric is measuring what it is supposed to be measuring (see Figure 2 and Table 3).

Figure 2: **Types of validity**



The overarching definition for construct validity is the extent to which the rubric measures the intended construct(s), and that these are accurately aligned with the appropriate theoretical framework. Validity can be assessed either subjectively, by reaching consensus between experts, or empirically, by analyzing scores. During rubric development, subjective judgements will be most important. It should be noted that there is overlap between these types of validity, as well as additional types of validity not included here.

Source: Adapted from Morling (2014, p. 138)

*****Slomp, Corrigan and Sugimoto (2014) propose several ways to measure consequential validity, including both subjective and empirical assessments.

Table 3: **Prompting questions for evaluating various types of validity**

Validity type	Questions to consider
Face validity	• Are all the important facets of the intended construct evaluated through the scoring criteria? • Are any of the evaluation criteria irrelevant to the construct of interest?
Content validity	• Do the evaluation criteria of the scoring rubric address all aspects of the intended content? • Do the evaluation criteria address any extraneous content?
Consequential validity	• What is the purpose of the rubric and how will the scores be used? • What stakeholders are important to help understand the consequences of the rubric? • What are the intended and unintended consequences based on the purpose/ use of the rubric?
Criterion validity	• How do the scoring criteria reflect competencies that would suggest success on future/related performances? • Are there any facets of the future/related performance that are not reflected in the scoring criteria?
Convergent validity	• Are there other well-established measures of this construct to which scores can be compared?
Discriminant Validity	• Is there any overlap in what is being measured between constructs? • Are scores from one construct correlated with an unrelated construct?

Source: Adapted from Moskal & Leydens (2000); Jonsson & Svingby (2007)

B) Test the rubric's validity by obtaining multiple rounds of recursive feedback

- Validate the rubric through subjective consensus of stakeholders, such as rubric developers, subject-matter experts, instructors, TAs and students.

Phase 1

- Test the rubric with front-line users (e.g., instructors or teaching assistants)[3] by grading multiple samples of pre-existing student work to compare differences in scores.[4]
- Host a feedback session to discuss scores that differ by two dimensions or more to grapple with constructs, identify unclear or awkward language, and assess criteria weighting.
- Prompt user feedback using questions like those in Table 3.

To Train or Not to Train?

Only train rubric users with the rubric before validity testing *if* there is intention to train all rubric users prior to implementation.

[3] The ideal number of student work samples and assessors is very context-dependent. Generally, the more front-line rubric users convened for testing, the fewer samples of student work required.

[4] Student work can be purposefully selected to have varying emphasis on as many rubric constructs as possible, and to represent a range of achievement. Consider whether assignments were completed on a voluntary basis or for grades, as this may have influenced the amount of effort put into the assignment, and thus could skew validity scores (Timmerman, Strickland, Johnson & Payne, 2011).

Phase 2

- Revise the rubric based on first feedback session.
- Host another feedback session with students to discuss terminology, clarity of criteria and their overall perceived utility of the rubric for guiding and assessing their work.
- Contact any other applicable stakeholders (e.g., employers) for their input on the rubric.

Phase 3

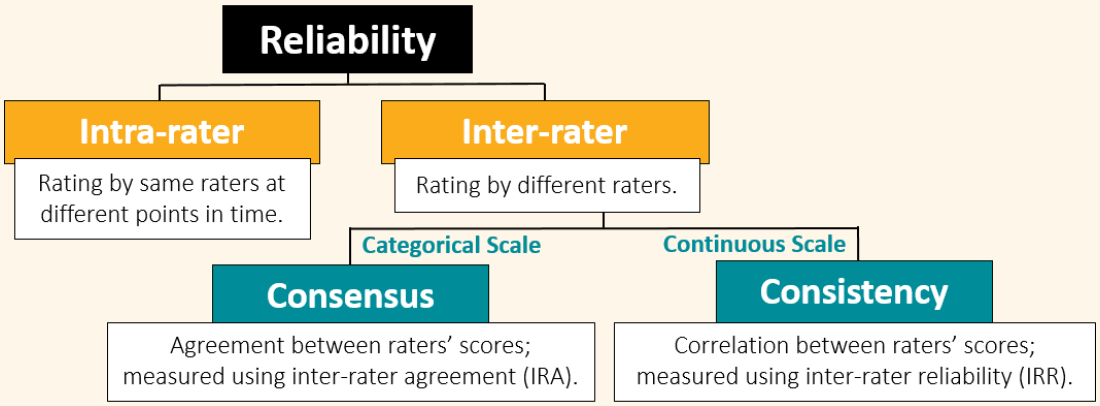
- Revise the rubric based on suggestions from the additional feedback session(s).
- Test the rubric a second time with a larger group of front-line rubric users, again by grading multiple samples of pre-existing student work.
- Ensure stakeholder feedback has been appropriately incorporated into the rubrics, and that phases 1-3 are repeated until no further revisions are needed.

STEP 4: TEST THE VALIDATED RUBRIC'S RELIABILITY

A) Understand the relevant types of reliability

- Ensure the rubric can be trusted to accurately and consistently assess student work (see Figure 3).[5]

Figure 3: Types of reliability



The most relevant element of reliability in this context is the consistency in assessment scores resulting from variation in raters' judgements. This can occur for an individual rater over time (i.e., intra-rater reliability), or across raters (i.e., inter-rater reliability). When comparing judgements across raters, measures will differ according to the grading scale (i.e., categorical vs. continuous), which will determine whether to examine the consensus of raters or the consistency of raters.

[5] Rezaei and Lovorn (2010) emphasize the importance of only using rubrics that are reliable, as they find that improper use of an unreliable tool is sometimes worse than not having used the tool at all.

B) Test and interpret the rubric's reliability

Phase 1

- Test the rubric with front-line users by reviewing multiple samples of pre-existing student work. Student work should be different from the samples used for validity testing if the testing group of front-line rubric users remains the same.

Phase 2

- Calculate reliability results (see Figure 3). The most common statistical measure is inter-rater reliability (IRR), which calculates the correlation between different raters' scores; alternatively, inter-rater agreement (IRA) calculates the exact agreement of scores between raters.[6]

Phase 3

- Analyze reliability measures in relation to their specific context, considering the purpose and scope of the rubric.[7]
- If reliability scores are deemed appropriate and no further changes are suggested by front-line rubric users, the rubric is ready for implementation.
- If reliability scores below 0.7 are not acceptable for the circumstances, repeat validity testing to understand inconsistencies with marking and/or repeat phase 1 of reliability testing using larger samples of student work and/or additional assessors until reliability coefficients are above 0.7.

Interpreting Reliability Coefficients

- > 0.9 = very strong
- 0.8 to 0.89 = strong
- 0.7 to 0.79 = acceptable
- < 0.7 = weak

Source: Morrison, Ross, Kemp & Kalman (2010)

STEP 5: IMPLEMENT RUBRIC

A) Train rubric users if implementing at a large scale

- Clearly articulate to assessors the purpose of the rubric, how many reviews will comprise a score for each piece of student work, how differences between raters will be adjudicated and what to do with scores.
- Ideally, complement the rubric with anchor assignments.[8]

[6] Reliability will inevitably be higher for tests that measure students' performance on the same test or task than for unique assignments that students can modify to align with their interests (Jonsson & Svingby, 2007).

[7] Rubrics used solely for individual-level interpretations may still be deemed reliable even when scores are below 0.7, whereas rubrics used for interpretations at the population level usually necessitate reliability scores above 0.7 (Jonsson & Svingby, 2007).

[8] Anchor assignments are actual work samples that illustrate various levels of attainment on the rubric (Jonsson & Svingby, 2007; Brophy, 2012).

B) Maintain rubric's validity and reliability

- Preserve students' anonymity by stripping student work of all identifiers, assigning each an identification code and standardizing the format (e.g., font, margins, line spacing).
- Total students' scores appropriately:
    - Be thoughtful when adding the ratings from separate constructs (rows) on an analytic rubric to provide a total score, as some constructs may be more important to course outcomes than others.
    - Refrain from averaging scores across constructs on analytic rubrics, as each score is applicable only to the dimension it is assigned.

Conclusion

This document gives an overview of important considerations for those looking to develop and/or validate rubrics. Rather than advocating this process as an exact science, we hope to encourage any process undertaken to create valid and reliable assessment tools for subjective assessment. Besides their obvious benefit as an assessment tool for instructors, rubrics should also be valued for their potential to influence the creation of better prompts and/or assignments, guide students' work and provide students with improvement-oriented feedback.

It's important to remember that validity and reliability are not fixed points to be reached in the final stages of rubric development. The dynamic educational contexts in which such assessments are implemented should evolve as the needs of students and society continue to change. As a result, testing the validity and reliability of such assessment tools and/or training assessors for their use should be an ongoing process.

References

Association of American Colleges and Universities (AAC&U). (2017). *On solid ground: Value report 2017.* https://www.luminafoundation.org/files/resources/on-solid-ground.pdf

Brophy, T. S. (2012). *Writing effective rubrics.* https://assessment.aa.ufl.edu/media/assessmentaaufledu/academic-assessment/writing_effective_rubrics_guide_v2.pdf

Jonsson, A. & Svingby, G. (2007). The use of scoring rubrics: Reliability, validity and educational consequences. *Educational Research Review*, *2*(2), 130–144.

Kapelus, G., Miyagi, N. & Scovill, V. (2017). *Building capacity to measure Essential Employability Skills: A focus on critical thinking.* Toronto: Higher Education Quality Council of Ontario.

Lesmond, G., McCahan, S. & Beach, D. (2017). *Development of analytic rubrics for competency assessment*. Toronto: Higher Education Quality Council of Ontario.

Morling, B. (2014). *Research methods in psychology: Evaluating a world of information.* New York: WW Norton & Company.

Morrison, G. R., Ross, S. M., Kemp, J. E. & Kalman, H. (2010). *Designing effective instruction*. Hoboken: John Wiley & Sons.

Moskal, B. M. & Leydens, J. A. (2000). Scoring rubric development: Validity and reliability. *Practical Assessment, Research & Evaluation*, *7*(10), 71–81.

Rezaei, A. R. & Lovorn, M. (2010). Reliability and validity of rubrics for assessment through writing. *Assessing Writing*, *15*(1), 18–39.

Slomp, D. H., Corrigan, J. A. & Sugimoto, T. (2014). A framework for using consequential validity evidence in evaluating large-scale writing assessments: A Canadian study. *Research in the Teaching of English*, *48*(3), 276–302.

Timmerman, B. E. C., Strickland, D. C., Johnson, R. L. & Payne, J. R. (2011). Development of a 'universal' rubric for assessing undergraduates' scientific reasoning skills using scientific writing. *Assessment & Evaluation in Higher Education*, *36*(5), 509–547.

Higher Education Quality Council of Ontario

An agency of the Government of Ontario