Using Rubrics for an Evaluation:
A National Research Council Pilot

Ghislaine Tremblay and Melissa Fraser

Office of Audit & Evaluation, National Research Council Canada

Frédéric Bertrand

Independent Senior Evaluation Consultant, Ber /

Abstract: *Rubrics are commonly used in the education sector to assess performance, products, or processes of student learning. Rubrics are gaining importance in organizational performance and program evaluation practice. According to several evaluation practitioners, rubrics can elucidate how excellence and value are defined and applied to evaluation questions or indicators in a given context. This practice note summarizes a pilot project of the National Research Council Canada (NRC) using evaluative rubrics for characterizing relevance and generating conclusions in an evaluation.*

Keywords: *performance, rubrics, science*

Résumé : *En pédagogie, on se sert souvent de rubriques pour évaluer la performance, les résultats ou la démarche d'apprentissage de l'étudiant. De plus en plus, les rubriques sont utilisées en analyse de la performance organisationnelle et dans la pratique évaluative. Selon certains évaluateurs, les rubriques peuvent contribuer à éclairer la façon dont l'excellence et le rendement sont définies et intégrées aux questions d'évaluation ou aux indicateurs dans un contexte donné. Cette note de pratique résume un projet pilote du Conseil national de recherches du Canada (CNRC) dans lequel on a utilisé des rubriques pour caractériser la pertinence d'un programme et générer les conclusions de l'évaluation.*

Mots clés : *performance, rubriques, science*

INTRODUCTION

Rubrics are a qualitative assessment tool commonly used in the education sector to measure student performance and provide descriptive feedback. They are also now gaining importance in organizational performance and program evaluation practice, where they have been found useful in clarifying and setting out the basis

Corresponding author: Melissa Fraser, National Research Council Canada, 1200 Montreal Road, Building M-58, Room S-107, Ottawa, Ontario, Canada K1A 0R6; Melissa.fraser@nrc-cnrc.gc.ca

on which "judgments about performance, quality, usefulness, and effectiveness are made" (King, McKegg, Oakden, & Wehipeihana, 2013, p. 12). In general terms, rubrics contain two key components: criteria and a scale of performance or quality (i.e., ratings). Evaluative rubrics are commonly represented in a cross-referencing table that describes the performance, value, and effectiveness of a program at different levels of performance based on evidence for a given criterion (Davidson, 2010). According to this approach, rubrics make "transparent how quality and value are defined and applied for each evaluation question or indicator given the context" (Davidson, 2010). When rubrics are well developed and implemented, they can substantially increase the use and credibility of evaluation (King et al., 2013).

The National Research Council (NRC) used rubrics in a recent program evaluation to:

- deliver a clear verdict on program performance based on well-defined performance levels and criteria; and
- create a shared understanding between program managers and the evaluators of the criteria for each performance level.

This practice note describes NRC's experience piloting evaluative rubrics. It first describes NRC's experience in developing rubrics followed by a discussion of lessons learned.

DEVELOPMENT OF RUBRICS

The steps followed by NRC to develop rubrics were adapted from Davidson, Wehipeihana, and McKegg (2011) and are provided in Table 1.

Using a collaborative approach

Rubrics were developed as part of the evaluation planning phase of one of NRC's research programs, in collaboration with program managers. The evaluation planning phase also included the development of a logic model and evaluation framework, including indicators, through a data and document review as well as key informant interviews. Following this, a 90-minute roundtable discussion was held with program managers to validate the evaluation framework and collaboratively develop rubrics for selected indicators.

The evaluation literature highlights the importance of using a collaborative process in developing rubrics. Collaborative rubric development facilitates stakeholder engagement, and sets expectations and standards that are realistic within the current organizational context (Adams, Nnawulezi, & Vandenberg, 2015; Davidson, 2010). Judy Oakden (2013) also noted that one of the benefits of collaboration is a shared understanding.

Consistent with the literature, NRC found that the roundtable discussion facilitated the program manager's buy-in to the evaluation and trust in the evaluation

Table 1. Steps for the Development of Evaluative Rubrics (Embedded in the Evaluation Planning and Design Process)

Use a collaborative approach

1. Develop logic model and identify evaluation questions through consultations with key program staff.
2. Develop evaluation framework.
3. Coordinate a roundtable session to validate the evaluation framework (evaluation questions and indicators) and develop rubrics for select indicators.
4. Provide roundtable discussion package to participants in advance of the session that includes an agenda, background information on rubrics, evaluation framework and logic model.

Choose the type of rubric

5. Consider whether a holistic or analytical rubric addresses the objectives of the evaluation.

Define performance

6. Moderate the roundtable discussion to:
   a. Validate the evaluation framework.
   b. Discuss the needed or optimal number of performance levels.
   c. Brainstorm what distinguishes excellent, good/adequate, and poor performance (or with more levels as desired).
   d. Draw boundaries around the dimensions that should be included or excluded from the criteria for achieving each level of performance.
   e. Translate these differences and boundaries in distinct dimensions to be articulated in a set of detailed rubrics and ensure that mutually exclusive levels of performance are defined.
   f. If needed, discuss the relative importance or weight of performance dimensions and rubrics associated to each indicator.

Finalize rubric

7. Write the rubric considering the input of key staff members and key literature to help with specific performance concepts or wording, including statistical considerations if quantitative analysis of rubrics scoring is needed.
8. Share the rubrics with roundtable discussion participants for detailed feedback and further debate, as needed.
9. If possible, share the rubrics with a subject matter/industry expert to validate performance criteria. Such expert should be external to the organization and could be a member of an evaluation advisory committee.
10. Revise, adjust and finalize the rubrics.

team. The roundtable discussion was also highly effective in establishing a common language between the program and the evaluation team, providing contextual information that enhanced the evaluation framework and rubrics, as well as building consensus for the rubrics (e.g., what criteria define each level of performance).

Choosing the type of rubric

The NRC evaluation team had to choose the type of rubric it would use to meet the objectives of the evaluation. According to the literature, there are two types of rubrics commonly used to assess performance: holistic and analytic rubrics (Brophy, 2012; Danks & Allen, 2014). An analytic rubric presents a description of each level of achievement for each criterion, evaluates multiple criteria along a separate scale or continuum, and provides each criterion with a separate score. Conversely, a holistic rubric consists of a single scale with all criteria included in the evaluation being considered together. Compared to a holistic rubric, an analytic rubric provides more detailed feedback on performance and facilitates a program manager's ability to make improvements. Consequently, analytic rubrics take a significant amount of time to develop. And while holistic rubrics are quick to use, they are not suitable for complex issues (Brophy, 2012; Danks & Allen, 2014).

For the NRC evaluation, an analytic rubric was used and deemed more appropriate because it afforded additional detailed feedback on the level of performance. Table 2 illustrates two examples of analytic rubrics developed by NRC for relevance indicators.[1]

Defining performance

Following the choice of rubric type, the NRC evaluation team then defined performance. This included identifying the number of performance levels, the criteria that define each level of performance, and the way in which performance is scored. According to Tierney and Simon (2004), a series of six questions needs to be considered when creating rubrics. These questions, which informed NRC's approach, are presented in Table 3.

Identifying performance levels and criteria

The number of performance levels, which are typically divided into three- to six-point scales, can vary from project to project (Center for Institutional Effectiveness, 2011). The most commonly used scales, as found in the literature, are four- or five-point scales including an "insufficient evidence" category. A greater number of performance levels requires increased effort to develop and validate performance descriptors.

NRC used three levels of performance plus an insufficient evidence category. The choice to use a three-point scale was the preference of the NRC program managers and aligned with a three-point scale the program already used to monitor the health of projects. NRC's approach of developing rubrics in collaboration with the program allowed program managers to take ownership of the rubric development process and to incorporate considerations of organizational culture

Table 2. Sample Analytic Rubrics from the NRC Pilot Project for Two Relevance Indicators

Indicator	Dimension	Excellent	Good/ Adequate	Poor	Insufficient evidence
Evidence of uniqueness, complementarity, or overlapping of program value proposition (and capabilities) vis-à-vis other S&T organizations in related sectors	1	Program value proposition and competitive advantage are clearly articulated and effectively communicated to stakeholders (vis-à-vis other S&T organizations in related sectors).	Program value proposition and competitive advantage are somewhat articulated and communicated to stakeholders.	Program value proposition and competitive advantage are not articulated nor communicated to stakeholders.	Information on the level of uniqueness, complementarity, or overlapping of program value proposition (and capabilities) vis-à-vis other S&T organizations in related sectors tended to be unavailable or insufficient to determine performance.
	2	Program is proactive and forward looking by both managing the current needs and considering future directions/needs of stakeholders to identify new capabilities and facilities needed.	Program is focused on the present by managing the current needs in terms of capabilities and infrastructure, but is not considering future directions/ needs of stakeholders.	Program is reactive by not managing the current nor considering future needs in terms of capabilities and infrastructure.	
	3	Program is developing new capabilities (e.g., people) and unique facilities/infrastructures to address future needs.	Program is developing capabilities (e.g., people) and facilities/infrastructures to address current needs.	Program is not developing capabilities (e.g., people) and/or facilities/infrastructures.	
	4	Program's capabilities are viewed as quite unique vis-à-vis other S&T organizations in related sectors.	Program is viewed as a valued supplier of R&D services and for scientific excellence in technology development in program areas.	Program is viewed as one of the suppliers of R&D services and technology development.	

Indicator	Dimension	Excellent	Good/ Adequate	Poor	Insufficient evidence
Level of complementarity and/or overlapping role and responsibilities of the program vis-à-vis other federal organizations	1	The program is collaborating with all the main federal departments/agencies with mandates that are related to those of the program's areas of activity (as there is no obligation for collaboration). Opportunities for further collaboration are identified and considered.	The program is collaborating with most of the main federal departments/ agencies with mandates that are related to those of the program's areas of activity (as there is no obligation for collaboration). Opportunities for collaboration are identified, but not yet considered.	Other federal departments/agencies are not (or avoiding) collaborating with the program by working with other service providers (for services/expertise offered by the program). Opportunities for collaboration are not identified nor considered.	Information on the collaboration activities of the program with federal departments/ agencies (and on their activities in the program's areas) tended to be unavailable or insufficient to determine performance.
	2	The program is viewed by other federal departments/ agencies as "the place to go" (e.g., resources, expertise, problem solving, and participation in committees).	The program is viewed by other federal departments/agencies as a valued supplier of R&D services and for scientific excellence in technology development in the program areas.	The program is viewed as one of the federal departments/agencies working in support of the industry sectors related to the program areas.	
	3	The program does not compete with other federal departments/agencies.	The program competes with some activities or services provided by other federal departments/ agencies.	The program competes with several other federal departments/ agencies.	

Table 3. Considerations When Creating Rubrics

1. How many performance levels should be used, considering current organizational practices and complexity of the evaluation framework?
2. What distinguishes excellent, good/adequate, and poor performance?
3. Are all the performance dimensions explicitly stated?
4. Are the attributes explicitly stated for each performance dimension and are the underlying characteristics of the performance dimension known and measurable?
5. Are the dimensions equally important (or weighted) for each indicator?
6. Are the boundaries around the dimensions for each criterion, for achieving each level of performance, clear and mutually exclusive?

Adapted from Tierney and Simon (2004)

and processes (i.e., use of a rating scale NRC program managers were familiar and comfortable with). As King et al. (2013) highlight, rubrics are a flexible and adaptable tool, and like good evaluative criteria, tend to be specifically tailored to the evaluation context and its use.

Scoring rubrics

There are several ways to score performance in rubrics. It is possible to assign a score for each criterion used to define an indicator. This approach allows evaluators and program managers to identify areas where the program is doing well and areas that need improvement.

It is also possible to compile an aggregated score for the indicator by taking an average of the scores assigned to each criterion in that indicator. One potential issue with this approach is the representativeness of the score for the indicator (i.e., taking an average of all criteria is not representative of the individual scores). In this regard, the validity and reliability of rubrics may be affected (Moskal & Leydens, 2000). If taking an average score for the indicator, consideration needs to be given to whether each criterion in the indicator is weighted equally. Weighted rubrics are used to reflect that some criteria may be of more significance than others. A common way to weight criteria in the education literature is to assign a percentage to each of them.

In the NRC evaluation, the analytic rubrics were not designed to quantify or attribute a score to each criterion. While the original intent was to weight the different criterion of each indicator, the evaluators decided against it because of the complexity involved in the program (i.e., multiple projects, infrastructure/equipment).

LESSONS LEARNED

Following the NRC pilot project, the evaluation team determined that the development and use of rubrics is not as straightforward as originally expected. In the words of King et al. (2013), "Rubrics, while useful, are not a panacea and they

are not as easy as they may appear. Like any skill, you can learn the theory, but it takes time and experience to become a skilled practitioner" (p. 13). As a result of the pilot project, the NRC evaluation team identified several lessons learned. Contributing to the lessons learned were findings from postproject consultations with the program managers involved in the development and use of rubrics. The key advantages and disadvantages of using rubrics stemming from this pilot are summarized in Table 4. The two major lessons learned, related to performance criteria and using rubrics in report writing, are discussed below.

Performance criteria

It was generally agreed upon that more time should be allocated in the future for the development of performance criteria to fully benefit from rubrics. While the NRC pilot project made use of additional resources (i.e., approximately two full-time equivalents above what is normally invested in the planning stage of an evaluation), insufficient time affected the quantity (number) of rubrics that could be developed and the quality of performance criteria that were developed. The NRC evaluation team reflected that, in the future, an additional three to four weeks of time would be needed if rubrics were to be effectively developed and used in the analysis and report writing stage.

Also, the NRC evaluation team felt that it would be beneficial to hold two 90-minute roundtable sessions (as opposed to one 90-minute session) to develop and validate the rubrics with program managers. Using this two-stage approach, the first session would address the evaluation framework and the second would be fully dedicated to the development of rubrics. Having one roundtable session focused on the evaluation framework is important because the discussion around language, contextual factors, and performance targets for each indicator of the evaluation framework serves as the foundation for rubrics.

Quantity of performance criteria

As a result of insufficient time in the pilot project, performance criteria could not be developed for all indicators. As such, the NRC pilot project was only able to develop rubrics for select relevance indicators, and select efficiency and economy indicators. To develop rubrics for performance-related indicators, additional time would have been necessary to review the available data and define performance targets.

Quality of performance criteria

Insufficient time also contributed to several challenges with the quality of performance criteria that were developed. The NRC evaluation team reflected that each rubric contained too many criteria that had to be met to receive a given performance level. For example, to achieve an excellent rating in one case, three criteria needed to be met. However, in cases where one or two of the criteria were met, but not the second or third, it was difficult to assign a rating. The evaluators were left questioning whether meeting one or two criteria warranted an excellent rating as opposed to a lower rating. As such, some degree of subjectivity was

Table 4. Summary of Key Advantages and Disadvantages of Using Rubrics

Section	Advantages	Disadvantages
Types of rubrics	• help to determine information to be collected (based on criteria) • • • *Analytic* • more detailed feedback on performance • shared understanding between evaluators and clients • easier for management to make improvements *Holistic* • quick to use • useful for single dimension criteria	*Analytic* • time consuming to develop *Holistic* • not suitable for complex issues • no detailed feedback • difficult to assign overall score
Application	• criteria are more focused and concise • ease of use • many different uses • a flexible tool • collaboration is needed for the development • more credible results • more useful for the client	• demands a lot of effort and experience • collaboration is needed for the development • evaluation becomes limited by rubrics criteria
Performance levels	• tailored to context	• more levels used, the more effort needed in development • difficult developing score
Scoring	• possibility of an average indicator (weighting sections) • a numerical value can be assigned to each level • provides consistency in scoring	• difficulties in assigning overall score criteria
Collabora-tion	• shared understanding •	• time-consuming process • large time investment from client

introduced. Postproject consultations with program managers also revealed that multidimensional criteria made it difficult for the program to validate their overall performance score. The difficulty with the rating criteria appears to be common, as Stone-Jovicich (2015) reported similar challenges. In the future, criteria used to define each level of performance should be distinct. As clarity and distinction are important to client comprehension, the criteria should not contain too many ideas.

The program managers also noted that it would have been a good practice to have external experts in their field consulted to ensure that the definitions of the performance criteria were appropriate and accurately represented the standard expected for the level of performance. This step may help ensure the quality of the performance criteria.

Report writing

One of the reasons that the NRC evaluation team first sought to pilot evaluation rubrics was for the oft-cited benefit of rubrics simplifying and expediting report writing. This, however, was not the experience of the NRC evaluators during the pilot and is a lesson learned. The NRC evaluation team found that in practice, rubrics made report writing harder because findings were bound by predetermined criteria. More specifically, rubrics hindered overall findings as they were not able to capture how different aspects interacted and what emerged as a result of those interactions. To some extent, the development of evaluation findings was constrained by the language of the rubrics. The challenge was reporting on the findings according to the rubrics while still being able to integrate additional findings. Integrating rubrics into the report also proved difficult, as not all key aspects of performance explored in the evaluation were outlined in the rubrics.

As a result of these challenges, NRC did not include the rubrics in the final evaluation report. The challenge posed by rubrics in facilitating emergent findings was also highlighted by Stone-Jovicich (2015). Ultimately, in the NRC pilot project rubrics ended up being an analytical tool that helped the evaluators identify data needs, collect and organize data, and make judgements about performance; they were not a reporting tool. Future use of rubrics will benefit from realistic expectations about their role and where they can add value to the evaluation process.

CONCLUSION

As a result of this pilot project, the NRC evaluation team reflected that it is unsure about the feasibility and practicality of using rubrics in NRC evaluations given the challenges experienced. That said, there were many benefits, particularly for stakeholders, as a result of the rubrics (e.g., client's understanding of what was being assessed was enhanced, evaluators' understanding of the program was improved). Due to these benefits, NRC will employ rubrics in a subsequent evaluation, incorporating the lessons learned from this first pilot project, before rendering a judgement as to their viability as an effective evaluation tool.

Based on the NRC evaluation teams experience, key actions for success include working with invested evaluation clients, having distinct and clear criteria for the rubrics, and defining a scale that is understood by all parties involved. In addition, realistic expectations about the role rubrics can play in the evaluation process will facilitate the successful use of rubrics in evaluations (e.g., as a data collection and analytical tool as opposed to a reporting tool).

ACKNOWLEDGEMENTS

NOTES

1   Readers interested in the full set of rubrics developed for the NRC pilot project are asked to contact the lead author for a copy.

REFERENCES

Adams, A. E., Nnawulezi, N. A., & Vandenberg, L. (2015). "Expectations to Change" (E2C): A participatory method for facilitating stakeholder engagement with evaluation findings. *American Journal of Evaluation*, *36*(2), 243–255. https://doi.org/10.1177/1098214014553787

Brophy, T. S. (2012). Writing effective rubrics (Institutional Assessment, University of Florida, Office of the Provost, Institutional Assessment, Continuous Quality Enhancement Series). Retrieved from http://assessment.aa.ufl.edu/Data/Sites/22/media/slo/writing_effective_rubrics_guide_v2.pdf

Center for Institutional Effectiveness. (2011). Developing rubrics for assessment. Ewing, NJ: Office of Assessment, The College of New Jersey. Retrieved from http://assess.pages.tcnj.edu/files/2011/06/Developing-Rubrics.pdf

Danks, S., & Allen, J. (2014). Performance-based rubrics for measuring organizational strategy and program implementation. *Performance Improvement Quarterly*, *27*(1), 33–49. https://doi.org/10.1002/piq.21166

Davidson, E. J. (2010, August 25). E. Jane Davidson on Evaluative Rubrics. AEA365| A Tip-a-Day by and for Evaluators [Web log comment]. Retrieved from http://aea365.org/blog/e-jane-davidson-on-evaluative-rubrics/

Davidson, J., Wehipeihana, N., & McKegg, K. (2011, September). The rubric revolution [PowerPoint slides]. Retrieved from http://www.betterevaluation.org/sites/default/files/AES-2011-Rubric-Revolution-Davidson-Wehipeihana-McKegg-xx.pdf

King, J., McKegg, K., Oakden, J., & Wehipeihana, N. (2013). Evaluative rubrics: Method for surfacing values and improving the credibility of evaluation. *Journal of Multidisciplinary Evaluation*, *9*(21), 11–20.

Moskal, B. M., & Leydens, J. A. (2000). Scoring rubric development: Validity and reliability. *Practical Assessment, Research & Evaluation*, *7*(10). Retrieved from http://PARE-online.net/getvn.asp?v=7&n=10

Oakden, J. (2013) Evaluation rubrics: how to ensure transparent and clear assessment that respects diverse lines of evidence (prepared for Better Evaluation. With permission from the Ministry of Education, New Zealand). Retrieved from http://betterevaluation.org/sites/default/files/Evaluation%20rubrics.pdf

Stone-Jovicich, S. (2015, August) To rubrics or not to rubrics? An experience using rubrics for monitoring, evaluating and learning in a complex project (Practice Note Series, Managing for Impact, The Food Systems Innovation (FSI) Initiative). Retrieved from http://foodsystemsinnovation.org.au/sites/default/files/pn_rubrics_ssj_final_0.pdf

Tierney, R., & Simon, M. (2004). What's still wrong with rubrics: Focusing on the consistency of performance criteria across scale levels. *Practical Assessment, Research & Evaluation, 9*(2). Retrieved from http://PAREonline.net/getvn.asp?v=9&n=2

AUTHOR INFORMATION

Ghislaine Tremblay is Director, Evaluation at the National Research Council. She previously held the position of Director of Evaluation at the Canada Foundation for Innovation. She has significant experience in conducting science and technology evaluations and in the management of the evaluation function. Ghislaine holds a Graduate Diploma in Program Evaluation and is a Credentialed Evaluator.

Frédéric Bertrand is an independent senior evaluation consultant (Ber /). Over the last 15 years, he has collaborated with more than 60 science and technology organizations, universities, foundations, and private firms by supporting managers and internal evaluators, designing and managing applied studies, and leading evaluations. Frédéric is a Credentialed Evaluator, holds a Masters in Environmental Sciences, and regularly supports the Canadian Evaluation Society (CES) for sponsorship activities and the student case competition.

Melissa Fraser is a Senior Evaluation Officer at the National Research Council. Melissa has ten years of experience evaluating science and technology programs. Methodologically, Melissa has significant experience with both qualitative and quantitative research methods, and holds a specialization in inferential statistics.