Practical Assessment, Research, and Evaluation

Volume 9 *Volume 9, 2004* Article 2

2004

What's still wrong with rubrics: Focusing on the consistency of performance criteria across scale levels

Robin Tierney

Marielle Simon

Practical Assessment, Research & Evaluation

A peer-reviewed electronic journal.

Volume 9, Number 2, January, 2004 ISSN=1531-7714

What's still wrong with rubrics: Focusing on the consistency of performance criteria across scale levels

Robin Tierney & Marielle Simon
University of Ottawa

Scoring rubrics are currently used by students and teachers in classrooms from kindergarten to college across North America. They are popular because they can be created for or adapted to a variety of subjects and situations. Scoring rubrics are especially useful in assessment *for* learning because they contain qualitative descriptions of performance criteria that work well within the process of formative evaluation. In recent years, many educational researchers have noted the instructional benefits of scoring rubrics (for example, Arter & McTighe, 2001; Goodrich Andrade, 2000). Popham noted their potential as "instructional illuminators" in a 1997 article entitled *What's Wrong - and What's Right - with Rubrics*, but he also cautioned that "many rubrics now available to educators are not instructionally beneficial" (p.72). Unfortunately, many rubrics are still not instructionally useful because of inconsistencies in the descriptions of performance criteria across their scale levels. The most accessible rubrics, particularly those available on the Internet, contain design flaws that not only affect their instructional usefulness, but also the validity of their results. For scoring rubrics to fulfill their educational ideal, they must first be designed or modified to reflect greater consistency in their performance criteria descriptors.

This article examines the guidelines and principles in current educational literature that relate to performance criteria in scoring rubrics. The focus is on the consistency of the language that is used across the scale levels to describe performance criteria for learning and assessment. According to Stiggins (2001), "Our objective in devising sound performance criteria is to describe levels of quality, not merely judge them" (p. 299). What is valued in a classroom, in terms of performances or products, is communicated through descriptive language. As such, performance criteria descriptors are a critical component of rubric design that merit thorough consideration. The purpose of this article is twofold:

1. To contribute to the educational literature aimed at improving the design of classroom assessment rubrics.
2. To assist rubric developers in creating or adapting scoring rubrics with consistent performance criteria descriptors.

In the following sections, the components of a rubric will be identified and defined, existing principles for performance criteria descriptors will be discussed, and consistency will be examined closely as a design requirement for rubrics.

Anatomy of a Rubric for Learning and Assessment

Scoring rubrics can be adapted or created for a variety of purposes, from large-scale or high-stakes assessment to personal self-assessment, and each has its own design features. The most useful rubrics for promoting learning in the classroom have been called instructional rubrics (Goodrich Andrade, 2000), analytic-trait rubrics (Arter & McTighe, 2001;Wiggins, 1998), and skill-focused rubrics (Popham, 1999). This article is specifically concerned with the type of classroom rubrics that can be described as descriptive graphic rating scales which use generic traits as analytic performance criteria (See Table 1 as an example).

The performance criteria in a rubric identify the dimensions of the performance or product that is being taught and assessed. The rubric in Table 1 contains generic performance criteria to assess the mapping skills of elementary students. This rubric does not attempt to dichotomously measure specific geographic knowledge as being present/absent or right/wrong. Instead, it emphasizes the development of valuable skills on a continuum. This particular rubric evolved from the curriculum model used in Ontario, Canada, where state curriculum standards are generally referred to as expectations. Mertler (2001) offers a template for the development of such rubrics.

Table 1: Generic Scoring Rubric for Classroom Assessment of Basic Mapping Skills

Mapping Skills Rubric

Purpose: This rubric is designed to be used in a formative context to assess basic mapping skills as stated in the local curriculum.

Instructions: For each performance criterion, circle or highlight the level that best describes the observed performance. To aid in this decision, refer to exemplars of student work or the task indicator list that is provided with the assessment task.

Performance Criteria	Attribute	Level 1	Level 2	Level 3	Level 4
The map includes the expected conventions (e.g. title, legend, cardinal directions) and geographic elements (e.g. countries, cities, rivers).	Breadth	The map contains **few** of the expected map conventions and geographic elements.	The map contains **some** of the expected map conventions and geographic elements.	The map contains **most** of the expected map conventions and geographic elements.	The map contains **all** of the expected map conventions and geographic elements.
The map conventions are used correctly and the geographic elements are placed accurately.	Accuracy	The expected map conventions and the geographic elements are **seldom** accurate.	The expected map conventions and the geographic elements are **sometimes** accurate.	The expected map conventions and the geographic elements are **usually** accurate.	The expected map conventions and the geographic elements are **always** accurate.
The map conventions are used appropriately in relation to the purpose of the map (e.g. red dashed line indicating exit routes on map for school fire drills).	Relevance	Map conventions and geographic elements are **slightly** relevant.	Map conventions and geographic elements are **moderately** relevant.	Map conventions and geographic elements are **mainly** relevant.	Map conventions and geographic elements are **extremely** relevant.
The map clearly communicates the targeted geographic information (e.g. symbols are easy to interpret, legend is easy to read).	Clarity	Information on the map is **slightly** clear.	Information on the map is **moderately** clear.	Information on the map is **mainly** clear.	Information on the map is **extremely** clear.

The performance criteria in this type of rubric are designed to represent broad learning targets, rather than features of a particular task, and this increases the universality of the rubric's application. The trade-off for this benefit is that the rubric does not contain concrete or task-specific descriptions to guide interpretation. As Wiggins (1998) suggests, generic rubrics should always be accompanied by exemplars of student work or task indicator lists. The variability of student

and rater interpretation can be reduced significantly when generic terms are clarified with task-specific exemplars or indicators. For example, a descriptor such as *moderately clear* becomes more observable when it is accompanied by a list of possible indicators. Using the mapping skills example, the clarity of a student's product could be affected by the legibility of the labels, the border style, the background color, or the choice of font. However, these product-specific indicators should not be explicitly stated on the rubric itself, not only because they limit the application of the rubric, but also because they can be easily confused with the targeted criteria (Wiggins, 1998).

The attribute, or underlying characteristic of each performance criterion, on the other hand, should be explicitly stated within the rubric. This concept was illustrated in a rubric that Simon & Forgette-Giroux (2001) put forth for scoring post-secondary academic skills. In Table 1, the attribute is highlighted in a separate column. Each criterion statement is clearly articulated in the left-side column, and then modified four times to describe each level of the performance's attribute(s). The choice of words that describe the changing values of the attribute is another dimension that must be dealt with in rubric design. Verbal qualifiers, such as *few*, *some*, *most* and *all*, indicate what type of scale is being used for each performance criterion. Three measurement scales are commonly used: amount, frequency, and intensity (Aiken, 1996; Rohrmann, 2003). Table 1 includes an example of each: The attribute *breadth* varies in terms of amount or quantity, *accuracy* varies in terms of frequency, and the last two, *relevancy* and *clarity*, vary in terms of intensity.

Existing Principles for Performance Criteria Descriptors in Scoring Rubrics

Principles or guidelines for rubric design abound in current educational literature. This study analyzed 21 documents directly related to rubric design. Most of the principles reported in these documents specifically addressed the issue of performance criteria while many focused on the quality of the descriptors. Most frequently mentioned is the clarity of the descriptors, and the impact of clarity on the reliability of the interpretations made by both the students and the raters (Arter & McTighe, 2001; Harper, O'Connor & Simpson, 1999; Moskal, 2003; Popham, 1999; Stiggins, 2001; Wiggins, 2001). Several authors also stressed that the performance levels (or score points) should be clearly differentiated through description (Moskal, 2003; Wiggins, 1998). Others noted that a balance between generalized wording, which increases usability, and detailed description, which ensures greater reliability, must be achieved (Popham, 1997; Simon & Forgette-Giroux, 2001; Wiggins, 1998). Less frequently mentioned, but nonetheless a desirable quality of central concern, is the need for consistent wording to describe performance criteria across the levels of achievement (Harper et. al., 1999; Simon & Forgette-Giroux, 2003; Wiggins; 1998). This, in effect, is the heart of the discussion.

Consistency of the Attributes in Performance Criteria Descriptors

Given the fact that consistency has not been discussed extensively in relation to rubric design, it is not widely understood by rubric developers as a technical requirement. The variety of terms that have been used to date in the literature on performance criteria may also have confused matters. One notion of consistency suggests that "parallel" language should be used (Harper et al, 1999; Wiggins, 1998). Parallel language is helpful when the attribute is clear, but this is regrettably not always the case. The performance criteria attributes in many of the rubrics that are found on the Internet are implied rather than explicitly stated, and their nature shifts from level to level. In a list of technical requirements, Wiggins addresses this problem and identifies as *coherent* rubrics those with consistent descriptor attributes:

> Although the descriptor for each scale point is different from the ones before and after, the changes concern the variance of quality for the (fixed) criteria, not language that explicitly or implicitly introduces new criteria or shifts the importance of the various criteria. (1998, p.185)

Simon & Forgette-Giroux (2003) also discuss consistency in performance criteria. They suggest that the descriptors for each level should deal with the same performance criteria and attributes in order for the progressive scale to be continuous and consistent from one level to the other.

Although the language that has been used in educational literature to discuss the consistency of performance criteria varies somewhat, the idea is essentially the same. Consistency in performance criteria can basically be viewed as the reference to the same attributes in the descriptors across the levels of achievement. In Table 1, the attribute, or underlying characteristic, of each criterion is consistently present across the scale, and it is the degree of the attribute that changes (e.g. level 4 reflects more accuracy than level 1). In another example, a rubric used in an intermediate history class might contain a performance criterion such as: student demonstrates an accurate and thorough understanding of the causes of the rebellion. The attributes of this criterion would be the accuracy and the depth of the student's understanding. In this case, accuracy and depth should be explicitly stated in the criterion statement, and they should also be present in each of the qualitative descriptors for that criterion across the levels of achievement.

Improving the Consistency of Performance Criteria Descriptors

Describing performance criteria can be a challenging aspect of rubric construction, which is in itself a task that many teachers find time-consuming. As an alternative to developing rubrics from scratch, teachers may adapt ready-made versions for use in their classrooms. A quick investigation using any popular search engine reveals that there are numerous sources for an endless variety of rubrics. When adapting a scoring rubric, it is important to realize that the

original purpose of the assessment may have resulted in design features that are not suitable for the adapted use. Many of the rubrics that are accessible online were created by teachers for specific tasks, and others were originally designed as holistic rubrics for large scale assessment, where the goal is to create an overall portrait of the performance. The latter are not necessarily intended to describe a continuum of learning as it is assessed in classrooms. The following examples were created to illustrate how some of the consistency problems found in accessible rubrics can be corrected for classroom use. In both examples, the problems are highlighted in the first row, and the modified versions are presented in the following rows (see Tables 2 and 3).

Example One: Basic Consistency

Many ready-made rubrics have basic consistency problems, meaning that the attribute or the performance criterion itself changes from level to level. Table 2 presents a task-specific rubric for assessing a science journal. The product, a science journal, is listed as if it is a performance criterion. This provides very little guidance for students who are learning to write a science journal. The attributes are implicit, and they change from level to level. At the *Novice* level, the descriptors stress accuracy of spelling, organization and breadth. Organization is dropped at the *Apprentice* level, but breadth and accuracy of spelling remain. At the *Master* level, only breadth remains of the original attributes, but clarity is added. And, finally, at the *Expert* level, neatness is further added, along with clarity and a vague requirement for creativity. In the modified version, an effort was made to stay true to the implied intent of the original criteria. The changes involve stating the performance criteria and the attributes clearly, as well as describing the qualitative degrees of performance more consistently from level to level. The modifications make the task, criteria, and attributes clearer for students, and they broaden the possibilities for the rubric's use. Accompanied by exemplars of student work or product-specific indicators, this rubric could be used by teachers and students to assess journal writing in any content-area class. It could also be used to assess the same skills in either a formative or a summative context with respective instructions. The corrections for this example deal specifically with the performance criteria. To complete the rubric, a title, a statement of purpose, and instructions for using the rubric should also be added.

Table 2: Example of Inconsistent Performance Criteria and Correction for Science Journal

Performance Criteria	Attribute	Novice	Apprentice	Master	Expert
		Problem Criterion			
Science Journal	(not stated)	Writing is messy and entries contain spelling errors. Pages are out of order or missing.	Entries are incomplete. There may be some spelling or grammar errors.	Entries contain most of the required elements and are clearly written.	Entries are creatively written. Procedures and results are clearly explained. Journal is well organized presented in a duotang.
		Suggested Correction			
The required elements are present for each journal entries (e.g. Lab Summary, Materials, Procedure, Results, Conclusion).	Breadth	**Few** of the required elements are present in each journal entry.	**Some** of the required elements are present in each journal entry.	**Most** of the required elements are present in each journal entry.	**All** the required elements are present in each journal entry.
The entries are	Clarity	Journal entries	Journal entries	Journal entries	Journal entries

clearly written (e.g. style, grammar enhance understanding).		are **slightly** clear.	are **moderately** clear.	are **mainly** clear.	are **extremely** clear.
The journal is organized (e.g. visible titles, ordered pages, etc.)	Organization	The journal is **slightly** organized.	The journal is **moderately** organized.	The journal is **mainly** organized.	The journal is **extremely** organized.

Example Two: Negative/Positive Consistency

Many rubrics, such as the problematic examples presented in Tables 2 and 3, describe the lower levels of performance criteria in purely negative terms, which creates a dichotomous (negative/positive) tone in the rubric. For young learners who are progressing along a continuum, this format sends the wrong message. Students who find themselves on the lower part of the scoring rubric may not be motivated to progress with this type of feedback. The performance criteria in a classroom rubric should reflect a positive learning continuum, and should not suggest that progression from Level 2 to 3 is a leap from failure to success. This does not mean that words, such as *none*, *not* or *seldom*, should always be avoided in rubric design, but that their use should represent one end of a continuous and consistent scale without undue negativity. However, when rubrics are not modified to reflect a positive continuum, they may perpetuate low expectations for certain students rather than promote learning.

In Table 3, autonomy, attention and enthusiasm are implicitly used as indications of silent reading ability. Essentially, such a complex and high-referenced skill is not one that can be adequately assessed with abstract attributes and a single criterion. The suggested corrections highlight the limitations of the rubric as a tool for assessing performance criteria that rely highly on inference rather than direct observation. As shown in Table 3, it is possible to measure these attributes with frequency and amount scales, but it is questionable whether the rubric would provide an accurate assessment of a student's reading ability. The process of articulation helps ensure that rubric designers are aware of the attributes that are actually involved, and forces them to question the validity of the performances being assessed in relation to the targeted construct. This example also illustrates that it is possible to include more than one attribute for each performance criterion without compromising the statement's clarity.

Table 3: Example of Inconsistent Performance Criteria for the Assessment of Silent Reading Skills.

Performance Criteria	Attribute	Emerging	Developing	Achieving	Extending
Problem Criterion					
Silent Reading	(not stated)	Off task and disruptive during sustained silent reading period.	Has difficulty choosing books for sustained silent reading.	Reads independently during sustained silent reading.	Chooses books with enthusiasm and reads independently during sustained silent reading.
Suggested Correction:					
1. If reading ability is the target, rethink the criterion to ensure that the attribute is meaningful.					

2. If learning behaviors are being measured, and autonomy and attention are the desired attributes, reword the descriptors as shown below.

Student reads independently and stays on task during a silent reading period.	Autonomy and Attention	Student **seldom** reads independently and stays on task for **little** of the time during a period of silent reading.	Student **sometimes** reads independently and stays on task **some** of the time during a period of silent reading.	Student **usually** reads independently and stays on task **most** of the time during a silent reading period.	Student **always** reads independently and stays on task **all** of the time during a silent reading period.

Guiding Questions to Ask in the Rubric Construction Process

The following questions are provided to further guide the process of creating consistent criteria descriptors while constructing or adapting scoring rubrics, particularly in an assessment for learning context:

1. **Are all the performance criteria explicitly stated?** Are the performance criteria present in the rubric those intended? Is there anything that is implicitly expected in the students' products or performances that is not stated in the rubric?
2. **Are the attributes explicitly stated for each performance criterion?** Are the underlying characteristics of the performance criteria known? Are these attributes clearly articulated within the rubric?
3. **Are the attributes consistently addressed from one level to the next on the progression scale?** Is the rubric addressing the same attributes for each student's product or performance across the levels? Does the value of the attribute vary in each level descriptor, while the attribute itself remains consistent across the scale levels?

Concluding Remarks

Rubrics that are used for classroom assessment must present clear and consistent performance criteria in order to live up to their educational ideal. When the attributes of each performance criterion shift from level to level across the scale, through variations either in presence or in tone, rubrics are less effective as learning tools. Students do learn from rubrics with inconsistent performance criteria, but *what* they learn may not be the intended learning goal. Rubric development can be challenging, and a rubric's design must be thoughtfully matched to its purpose. Consistency is an important technical requirement that should be considered carefully for all scoring rubrics designed or adapted for classroom use.

The most challenging aspect of designing rubrics for the classroom is in the language used. Although indicators and exemplars can help operationalize the attributes and performance criteria in rubrics, the choice or wording is still critical. The verbal qualifiers of the attributes used in rubrics, and their underlying scales, have not been standardized to the degree that they are universally understood, and fuzziness is associated with the interpretations. The precision of language in rubrics, and the development of common scales, are areas that would benefit from further research.

This article examines principles and provides suggestions for improving the consistency of performance criteria across rubric scale levels. By making a contribution to the educational literature on advancing the design of rubrics, this article strives to improve current classroom assessment practices. As Stiggins noted, "constructive classroom assessment [involves] defining the achievement targets" (2001, p. 3). To provide students and teachers with a clear and common understanding of these targets, rubrics must be accompanied by exemplars or clear indicators, and they should contain consistent descriptions of performance criteria as well as explicitly stated attributes. Within a formative context, students who use these rubrics then have an opportunity to build on their initial performance and adjust their learning accordingly. Rubrics do benefit instruction and they do become ideal tools in the assessment for learning process when they are designed with consistency in mind.

References

Aiken, L.R. (1996). *Rating scales and checklists: Evaluating behavior, personality, and attitudes*. New York, NY: John Wiley & Sons, Inc.

Arter, J. & McTighe, J. (2001). *Scoring rubrics in the classroom: Using performance criteria for assessing and improving student performance*. Thousand Oaks, CA: Corwin Press/Sage Publications.

Goodrich Andrade, H. (2000). Using rubrics to promote thinking and learning. *Educational Leadership, 57*, 13-18.

Harper, M., O'Connor, K. & Simpson, M. (1999). *Quality assessment: Fitting the pieces together*. Toronto, ON: Ontario Secondary School Teachers Federation.

Mertler, C.A. (2001). Designing scoring rubrics for your classroom. *Practical Assessment, Research & Evaluation, 7*(25). Retrieved May 12, 2003 from http://pareonline.net/getvn.asp?v=7&n=25

Moskal, B.M. (2003). Recommendations for developing classroom performance assessments and scoring rubrics. *Practical Assessment Research & Evaluation,* 8 (14). Retrieved on May 30, 2003 from http://pareonline.net/getvn.asp?v=8&n=14

Popham, W.J. (1997). What's wrong - and what's right - with rubrics. *Educational Leadership, 55*, 72-75.

Popham, W.J. (1999). *Classroom assessment: What teachers need to know* (2nd Ed.). Needham Heights, MA: Allyn & Bacon.

Rohrmann, B. (2002). *Verbal qualifiers for rating scales: Sociolinguistic considerations and psychometric data.* Retrieved October 7, 2003, from University of Melbourne Web site: http://www.psych.unimelb.edu.au/staff/br/vqs-report.pdf

Simon, M. & Forgette-Giroux, R. (2001). A rubric for scoring postsecondary academic skills.*Practical Assessment, Research & Evaluation, 7*(18). Retrieved December 23, 2003 from http://pareonline.net/getvn.asp?v=7&n=18

Simon, M. & Forgette-Giroux, R. (2003). Étude critique des composantes actuelles des grilles d'évaluation du rendement du curriculum de l'Ontario. Report submitted to the Ontario Ministry of Education. Toronto, ON.

Stiggins, R.J. (2001). *Student-involved classroom assessment* (3rd ed.). Upper Saddle River, NJ: Prentice-Hall, Inc.

Wiggins, G. (1998). *Educative assessment: Designing assessments to inform and improve student performance*. San Francisco, CA: Jossey-Bass Publishers.

Descriptors: Scoring rubrics; rating scales; performance criteria; consistency; classroom assessment; assessment for learning; student evaluation.

Citation: Tierney, Robin & Marielle Simon (2004). What's still wrong with rubrics: focusing on the consistency of performance criteria across scale levels. *Practical Assessment, Research & Evaluation*, 9(2). Available online: http://PAREonline.net/getvn.asp?v=9&n=2.