Metarubric: Criteria for judging the quality of rubrics—a rubric for rubrics

Content/Coverage: What counts? What they see is what you'll get.

- Does it cover everything of importance?
- Does it leave out unimportant things?

Clarity: Does everyone understand the terms and criteria used?

- Are terms defined?
- Are various levels of quality defined?
- Are there samples of work to illustrate levels of quality?

Practicality: Is it easy for both teachers and students to use?

- Would students understand the terms and criteria used? Is there a student-friendly version?
- Could students use it to self-assess?
- Is the information provided useful for planning instruction?
- Is the rubric manageable?

Technical Quality/Fairness: Is it reliable and valid?

- Is it reliable? Would different raters give the same score?
- Is it valid? Do the ratings actually represent what students can do?
- Is it fair? Does the language adequately describe quality for all students? Are there racial, cultural, gender, or other biases?

This Metarubric was developed for classroom assessments, not large-scale assessments. Although many of the quality criteria are the same for both uses, some features required for large-scale assessments may be counterproductive in a rubric intended for classroom use.

Please note that we use the terms *rubric, scoring guide*, and *performance criteria* all to mean written statements of the characteristics and features that lead to a high-quality product or performance by students.

The descriptors under each trait in the Metarubric are not meant as a checklist. Rather, they are indicators to help the user focus on the correct level of quality of the rubric under consideration. A strong score ("Ready to Roll") doesn't necessarily mean that the rubric

is perfect; rather, it means that you would have very little work to do to get it ready to use. A middle-level score ("On Its Way, But Needs Revision") means that the rubric is strong in some ways but weak in others—it would take some work to make it usable, but it is probably worth the effort. A weak score ("Not Ready for Prime Time") means that the rubric needs so much work that it is probably *not* worth the effort—it is time to find another one or begin from scratch.

Please note that a middle score does not mean "average"; this is a criterion-reference scale, not a norm-referenced one. It is meant to describe levels of quality in a rubric, not to describe what is currently available. The "average" rubric currently available may, for example, in fact be closer to "needs revision" than to "on its way."

The scale could easily expand to a five-point scale. In this case, think of "4" as a balance of characteristics from the "5" and "3" levels, and a "2" as a balance of the "3" and "1" levels.

Trait 1: Content/Coverage

The content of a rubric defines what to look for in a student's product or performance to determine its quality. The rubric describes what will "count." Regardless of what is stated in content standards, curriculum frameworks, or instructional materials, the *content* of the rubric is what teachers and students will use to determine what they need to do in to succeed: what they see is what you'll get. Therefore, it is critical that the rubric cover all essential aspects that define quality in a product or performance and omit all things trivial.

Ask yourself these questions when evaluating a rubric for content: Can I explain why each item in the rubric is essential to a quality performance? Can I cite references that describe the best thinking in the field on the nature of quality performance? Can I describe what is left out and why? Do I ever find performances or products that are scored low (or high) that I really think are good (or bad)? (If so, it is time to reevaluate the content of the rubric.) Is this rubric worth the time devoted to it?

Ready to Roll

- Indicators of quality represent the best thinking in the field.

- The content has the "ring of truth"—your experience as a teacher confirms that the content is truly what you *do* look for when you evaluate the quality of student work or performance.

- If the number or amount of something is included as an indicator, such counts really *are* indicators of quality.

- Relevant emphasis on various features of performance is right—things that are more important are stressed more; things that are less important are stressed less.

- Each important feature that contributes to quality work appears in only one place in the rubric. The same feature is not repeated under different traits in the same rubric.

- Definitions of terms are precise, correct, and reflect current thinking in the field.

- The number of points used in the rating scale make sense. The level of precision is appropriate for the use.

- You have few questions about what was included or why.

- The rubric is insightful. It truly helps you organize your thinking about what it means to perform with quality. The content will help you assist students to understand the nature of quality.

Trait 1: Content/Coverage (Continued)

On Its Way, But Needs Revision

- The rubric is about halfway there on content. Much of the content is relevant, but you can easily identify important elements that have been left out or that have been given short shrift.
- The developer is beginning to develop the relevant aspects of performance. You can see where the rubric is headed, even though some features might seem of questionable value or out of balance with others.
- Although much of the rubric seems reasonable, some of it doesn't seem to represent current best thinking about what it means to perform well on the product or skill under consideration.
- Although the content seems fairly complete, the rubric sprawls—it is not organized as well as it could be.
- Although much of the rubric covers important content, it also contains several irrelevant features that might lead to an incorrect conclusion about the quality of students' performance.

Not Ready for Prime Time

- The rubric does not include many readily identifiable important dimensions of a quality performance or product.
- The rubric contains irrelevant features. You find yourself asking, "Why assess this?" or "Why should this count?" or "Why should students do it *this* way?"
- Content is based on counting the number of something when quality is more important than quantity.
- The rubric seems "mixed up"—like items are placed in different categories. Items or criteria that are different are placed together.
- The rubric is out of balance—features of importance are weighted incorrectly.
- Some indicators of quality work are present in more than one trait.
- Definitions are incorrect—terms don't reflect current best thinking in the field.
- The rubric is a seemingly endless list of everything the developer thought might be even marginally important. There is no organization. The rubric looks like a brainstormed list.
- You are left with many questions about what was included and why.
- There are many features of the rubric that might lead a rater to an incorrect conclusion about the quality of students' performance.
- The rubric does not seem to align with the content standard it is supposed to assess.

Trait 2: Clarity

A rubric is clear to the extent that teachers, students, and others are likely to interpret the statements and terms in the rubric the same way. A rubric can be strong on the trait of content/coverage but weak on the trait of clarity—the rubric seems to cover the important dimensions of performance, but does not describe them well. Likewise, a rubric can be strong on the trait of clarity, but weak on the trait of content/coverage—the rubric's coverage is clear, but it is also trivial or focused on minutiae.

Ask yourself these questions when evaluating a rubric for clarity: Would two teachers give the same rating on a product or performance? Can I define each rubric statement so that students can understand what I mean? Could I find examples of student work or performances that illustrate each level of quality? Would I know what to say if a student asks, "Why did I get this score?"

Ready to Roll

- The rubric is so clear that different users (teachers or students) would give the same rating to the same product or performance.
- A single user could use the rubric to provide consistent ratings across assignments, time, and students.
- Terms are specific and accurate. It is easy to understand exactly what the rubric covers and why.
- Samples of student products or performance illustrate each score point. It is clear why each sample was scored the way it was.
- Definitions of terms are precise, correct, and reflect current thinking in the field.
- There is sufficient descriptive detail in the form of concrete indicators, adjectives, and descriptive phrases to allow you to match a student performance to the "right" score.
- There is not an overabundance of descriptive detail—the developer seems to have a sense of that which is most telling.
- The basis for assigning ratings or checkmarks is clear. Each score point is defined with indicators and descriptions.
- There are student-friendly versions.
- The rubric is so clear that a student doing poorly would know exactly what to do to improve.

Trait 2: Clarity (Continued)

On Its Way, But Needs Revision

- Major headings are defined, but there is little detail to assist the rater to choose the proper score points.

- Definitions of terms and descriptors are present, but are not sufficient.

- Teachers would agree on how to rate some elements in the rubric. Others are not well defined and would probably result in disagreements.

- A single teacher would probably have trouble being consistent in scoring across students or assignments.

- Student-friendly versions exist, but need more clarity.

Not Ready for Prime Time

- Language is vague. You find yourself saying such things as, "I'm confused," or "I don't have any idea what they mean by this."

- Definitions for terms used in the rubric are unhelpful, incorrect, or missing.

- The rubric is little more that a list of categories to rate followed by a rating scale. Nothing is defined. Few descriptors are given to define levels of performance.

- No sample student work is provided to illustrate the rubric's criteria.

- Teachers are unlikely to agree on ratings because descriptors are open to interpretation.

- Levels of quality are distinguished inadequately, using such terms as *extremely, very, some, little*, and *none*; or *completely, substantially, fairly well, little*, and *not at all*.

- There are no student-friendly versions; students are not likely to understand the rubric.

Trait 3: Practicality

Having clear criteria that cover the right "stuff" means nothing if the system is too cumbersome to use. *Practicality* refers to ease of use. Can teachers and students understand the rubric and use it easily? Does it give them the information they need for instructional decision making and tracking student progress toward important learning outcomes? Can the rubric become more than just a way to assess students—can teachers and students also use it to improve the very achievement being assessed?

Ready to Roll

- The rubric is manageable—there are not too many things to remember; teachers and students can easily internalize them.

- It is clear how to translate results into instruction. For example, if students appear to be weak in writing, is it clear what should be taught to improve performance?

- The rubric's focus is usually analytical trait rather than holistic when the product or skill is complex.

- For reasoning, skills, and product targets, the rubric is general rather than task specific.

- If task-specific and/or holistic rubrics are used, their justification is clear and appropriate. Justifications may include (1) the complexity of the skill being assessed—a "big" skill would require an analytical trait rubric while a "small skill" might only need a single "holistic" rubric; or (2) the nature of the skill being assessed—understanding a concept might require a task-specific rubric while demonstrating a skill (such as an oral presentation) might warrant a general rubric.

- Students easily may use the rubric to revise their own work, plan their own learning, and track their own progress. The author provides assistance on how to use the rubric in this fashion.

- The rubric is visually appealing to students; it draws them into its use.

- The rubric is divided into easily understandable chunks (traits) that help students grasp the essential aspects of a complex performance.

- The language used in the rubric is developmental—low scores do not imply "bad" or "failure."

Trait 3: Practicality (Continued)

On Its Way, But Needs Revision

- The rubric provides useful information, but it is not always easy to use.
- The rubric might be generic, yet holistic (when to be of maximal use, an analytical trait rubric would be better).
- The rubric has potential for teacher use, but needs some "tweaking." This could include combining long lists of attributes into traits or adjusting the language so criteria of quality are clear.
- The rubric has potential for student self-use, but needs some "tweaking." This could include wording changes, streamlining, or making the format more appealing.
- Students could accurately rate their own work or performances, but it might not be clear to them what to do to improve.
- Although there are some problems, it would be easier to try and fix the rubric than look elsewhere.

Not Ready for Prime Time

- There is no justification given for the type of rubric used—holistic or analytical trait; task specific or generic. You get the feeling that the developer did not know the options available and just did what seemed a good idea at the time.
- The author did not seem to consider how the rubric might be useful to teachers. The intent seems to be only large-scale assessment efficiency.
- The rubric is not manageable—there is an overabundance of things to rate—it would take forever; or, everything is presented all at once and might overwhelm the user.
- It is not clear how to translate results into instruction.

Trait 4: Technical Quality/Fairness

It is important to have "hard" evidence that the performance criteria adequately measure the goal being assessed, that they can be applied consistently, and that there is reason to believe that the ratings actually do represent what students can do. Although this might seem beyond the scope of individual classroom teachers, we all still have the responsibility to ask the tough questions when we adopt or develop a rubric.

Ready to Roll

- Technical information associated with the rubric describes rater agreement rates and the conditions under which they can be obtained. These rates are at least 65 percent exact agreement, and 98 percent within one point.

- The language used in the rubric is appropriate for the diversity of students found in typical classrooms. The language avoids stereotypic thinking, appeals to various learning styles, and uses language that English Language Learner students would understand.

- There have been formal bias reviews of rubric content; studies of ratings under the various conditions in which ratings will occur, and studies that such aspects as students' handwriting, gender, or race will not affect judgment.

- Wording is supportive of students—it describes learning or achievement status and does not make judgments of student worth.

On Its Way, But Needs Revision

- There is technical information associated with the rubric that describes rater agreement rates and the conditions under which such agreement rates can be obtained. These agreement rates are not at the levels described under "Ready to Roll," but this might be due to less than adequate training of raters rather than to the scale itself.

- The language used in the rubric is inconsistent in its appropriateness for the diversity of students found in typical classrooms. However, these problems may be easily corrected.

- The authors present hard data on the technical soundness of the rubric that is suggestive but not sufficient.

- Wording is inconsistently supportive of students, but may be easily corrected.

Trait 4: Technical Quality/Fairness (Continued)

Not Ready for Prime Time

- There is no technical information associated with the rubric.

- There have been no studies on the rubric to show that it assesses what it intends.

- The language used in the rubric is not appropriate for the diversity of students found in typical classrooms. The language might include stereotypes, appeal to some learning styles over others, and might put ELL students at a disadvantage. These problems are not easily corrected.

- The language used in the rubric might be hurtful to students. For example, the low end of the rating scale uses terms such as *clueless* or *has no idea how to proceed*. These problems would not be easy to correct.