Practical Assessment, Research & Evaluation

A peer-reviewed electronic journal.

Volume 19, Number 8, August 2014 ISSN 1531-7714

Revising an Engineering Design Rubric: A Case Study Illustrating Principles and Practices to Ensure Technical Quality of Rubrics

Gail Lynn Goldberg
Gail Goldberg Consulting

This article provides a detailed account of a rubric revision process to address seven common problems to which rubrics are prone: lack of consistency and parallelism; the presence of "orphan" and "widow" words and phrases; redundancy in descriptors; inconsistency in the focus of qualifiers; limited routes to partial credit; unevenness in incremental levels of performance; and inconsistencies across suites or sets of related rubrics. The author uses examples from both the draft stage precursor and the first revised (pilot) version of the Engineering Design Process Portfolio Scoring Rubric (EDPPSR), to illustrate the application of broadly relevant guidelines that can inform the creation of a new—or revision of an existing—rubric to achieve technical quality while preserving content integrity.

From elementary grades to post graduate studies, the evaluation of students' work relies at times on a rubric to assign a level of performance and sometimes a rating or grade. The task of creating rubrics, once the province primarily of assessment specialists, is today often assumed by classroom teachers and even by the students whose work will be subject to evaluation. This would suggest that creating a rubric is a relatively simple task, a notion supported by countless trade books, articles, and online rubric generators. Far less common are resources that identify technical characteristics of sound rubrics or guide review and revision to attain those characteristics (see, for example, Moskal, 2003; Tierney & Simon, 2004; Wiggins, 1998). Still missing, it appears, are any detailed accounts of a rubric revision process that illustrate the application of a broad set of rules or guidelines to achieve technical quality.

This article endeavors to fill that gap by describing a key stage in the evolution of the Engineering Design Process Portfolio Scoring Rubric (EDPPSR)[1], an instrument intended not only to guide valid and reliable score decisions on portfolio entries but to provide a blueprint for teaching and learning the engineering design process. It provides a case study of the application of various principles and practices in writing or revising any rubric to ensure technical quality and content integrity which may be applied by educators engaged in crafting a rubric "from scratch" as well as those who wish to be informed consumers of the plethora of rubrics available in print and online.

Background on Rubrics

A rubric is a scoring guide that outlines features of work at different levels of performance. It typically consists of a hierarchical score scale—numerical, descriptive, or both—and descriptors for each level. These descriptors may take the form of a paragraph or a list; either way, they should identify the characteristics indicative of each level. Rubrics can focus on a product, performance, or process and can be applied to a single artifact or an array. Although rubrics may differ widely in scope and structure, a strong case can be made, as did James Popham (1997), that the most useful rubrics are generic rather than task-specific—that is, they can be applied effectively to various assignments or tasks that

[1] The most current version of the EDPPSR (August 2011) can be accessed online on the Innovation Portal; see http://innovationportal.org

are intended to demonstrate the same or similar skills and understandings.

Although there are lots of wrong ways, there is no single right way to construct a rubric. The number of performance levels defined in a rubric depends upon its intended use or uses. Thus, a rubric used as a "gatekeeper" to make dichotomous decisions such as pass/fail, accepted/rejected, or credit/no credit need only have two levels: a student's performance either demonstrates or fails to demonstrate the criteria associated with success. In contrast, when rubric-based judgments are intended to inform instruction, teachers and students are likely to welcome many levels to differentiate performance along a continuum of growth and learning. Only then is "the rule of thumb…to have as many scale points as can be well defined and that adequately cover the range from very poor to excellent performance" pertinent (Perlman, 1994, p. 8). The use of "0" as a category may describe minimal or insufficient evidence in some instances but in others may be reserved to indicate that work is missing or completely incorrect. Regardless of how many levels are established, it is critical that criteria capture the essential attributes of work at each level. Otherwise attention may shift from consideration of the skills and understandings being assessed to simply sorting work from sub-par to stellar without connecting judgments to their implications for subsequent teaching and learning.

Early History of the EDPPSR

The idea of a rubric to evaluate evidence of the engineering design process in a portfolio of student work can trace its origin to discussion of the idea of an Advanced Placement (AP®) exam in engineering during the Strategies for Engineering Education K-16 (SEEK-16) Summit in 2005 (see Abts, 2011; Groves et. al., 2012; Groves et. al, 2014). With funding support from the National Science Foundation (NSF) and the Kern Family Foundation, under the leadership of Dr. Leigh Abts, University of Maryland, focus group and interview feedback from educators and engineering practitioners was integrated within and "layered-over" a draft rubric originally developed by Mark Schroll, Director of Strategic Initiatives for Project Lead the Way. When Jay McTighe, then a consultant to the project, recommended engagement of an assessment specialist with expertise in rubric design and development to lead review and revision of what was then referred to as the "Engineering Portfolio Grading Rubric,"[2] the principal project investigators sought the services of the author. The result was the first iteration of what is now known as the Engineering Design Process Portfolio Scoring Rubric, or EDPPSR.[3]

Application of Rubric "Rules" and the Evolution of the EDPPSR

Although from early stages the engineering design rubric was referred to in the singular as "the rubric", the prototype from which the EDPPSR evolved was—and the EDPPSR still is—better understood as a suite of rubrics, since a score scale and set of descriptive criteria exists for each of the various elements of the engineering design process (referred to hereafter as element rubrics). The draft stage document, as it existed in the autumn of 2010, consisted of thirteen "portfolio attributes"[4], subsequently renamed "elements" of the design process, organized under six broad categories (most but not all of which referenced steps in the engineering design process): identifying, articulating, and justifying a problem; generating an original solution; construction of a testable prototype or process; analyzing testing data; reflection and recommendations; and project presentation/representation quality.

The earliest and easiest revisions implemented were those that addressed global and cosmetic issues that pertained specifically to the EDPPSR and are only occasionally relevant to other rubrics. These included: 1) creating clear and consistent headings and subheadings and eliminating the "goal statements" that had been included for only some elements (and were no longer needed once the descriptors were thoroughly revised and the expectations for each element made clear); 2) reformatting so that all variations in font types, layout, and text features were either changed to become meaning-bearing or eliminated; and 3) reversing the order of score point descriptors, which originally went

[2] This early version has been more recently referred to as the "Design Process Rubric." See Abts, 2011

[3] In spite of the recommendation that "grading" be removed from the original name since the rubric would have many uses beyond the assignment of grades or scores, after all revisions to the 2010 draft were accepted, "grading" was transformed by project leadership to "scoring"—thus accounting for the "S" in the EDPPSR.

[4] Since the term "attribute" is already widely understood to refer in assessment literature to an essential characteristic of a performance criterion, to avoid confusion the term "element" will be the term used herein to refer to both original and revised rubrics that comprise the EDPPSR.

from low to high, to mirror the more customary structure—high to low—that characterizes rubrics used by large-scale assessments like the SAT, ACT, NAEP, PARCC, and SmarterBalanced, as well as those modeled for classroom assessment (see, for example, Arter & McTighe, 2001).

The far more demanding revisions were those required to ensure the technical soundness of the EDPPSR. Towards that end, each element rubric was screened for seven problems or design flaws identified by the author to which, based on her experience, rubrics across virtually all grade levels and disciplines are sometimes prone. Examples that follow from the 2010 draft and the revised version of that suite of rubrics highlight each of these problems and how they were addressed, thus serving as models for in-depth revision of other rubrics, whether newly created or adopted—with or without modification—from another source.

Problem 1: Lack of consistency and parallelism

A decade ago, based on their analysis of nearly two-dozen documents related to rubric design, Tierney and Simon (2004) concluded that design guidelines generally focused on the need for clarity but far less often on the need for consistency. They elaborated upon the concept of consistency by discussing consistency of attributes, performance criteria, and what they call "negative/positive consistency" (avoiding a shift from describing criteria in positive terms to negative ones). Beyond consistency, however, performance descriptors are improved through parallelism—not just in language (as called for by Wiggins, 1998), but in syntax as well. Simply put, across score point descriptions well-crafted rubrics will address the same attributes, in the same order. Rubrics that are effective and easy to follow tend to identify key features of a product or performance and then differentiate between and among score points through words or phrases that describe a variable such as quality, quantity, or frequency. Parallelism in language choices and in the arrangement of phrases, sentences, and the descriptors overall will permit users of the rubric to more easily match key features of a product or performance to attributes that accurately describe it, while a lack of parallelism will confound the scoring process. This is illustrated by comparing the original and revised versions of one element of the EDPPSR (see Table 1 below).

It is not easy to distill from the original rubric descriptors that the focus of Element C is supposed to be documentation and analysis of research into previous attempts to solve an identical or similar problem to the one featured in the portfolio. In the revised version of the rubric, parallel language choices and structure are incorporated to make the expectations for this element of the engineering design process clear.

Table 1: Two Adjacent Score Point Descriptors Before and After Revision for Parallelism of Language and Syntax

From the original rubric for Element C (*Analysis of current and past attempted solutions*):

2) Though some past and present solutions have been documented, either the research strategies employed to populate the list of possibilities were too narrow or the analysis of the results lacked any measurable details, technical understanding, or both.

3) Evidence of a thorough investigation of current and past solutions with sufficient technical explanations of the function and process of each. Analysis of this research produced a well-defined list of current and past solution attempt shortcomings relative to the problem statement.

From the revised version:

2. Documentation of existing attempts to solve the problem and/or related problems is drawn from a limited number of sources, some of which may not be clearly identified and/or credible; the analysis of past and current attempts to solve the problem—including strengths and shortcomings—is overly general and contains little detail and/or relevant supporting data

3. Documentation of existing attempts to solve the problem and/or related problems is drawn from several—but not necessarily varied—clearly identified and generally credible sources; the analysis of past and current attempts to solve the problem—including strengths and/or shortcomings—is generally clear and contains some detail and supporting data

Even when the same attributes are presented in the same order in a rubric, seemingly minor inconsistencies in language choices can still obscure the target and confound score decisions. Consider, for example, wording in the descriptors for the original Element E (*Design Process Thinking and Analysis*):

- solution possibilities (score points 0, 3 and 4); possible solutions (score points 2 and 5)

- given solution (score point 1); final solution (score point 4)
- final choice/plan of action (score point 4); design or plan of action (score point 5)

Identifying and eliminating even such minor inconsistencies was one focus of revision of the EDPPSR, and should be part of any rubric revision and refinement process as well.

Problem 2: The presence of "orphan" and "widow" words and phrases

The rubric revision process also focused upon the identification and elimination or correction of two specific flaws that are closely related to the problem of lack of consistency and parallelism: the presence among key words or terms of "orphans"—those that appear only at one score point level and nowhere else—and "widows"—those that are missing from one score point descriptor but appear in all others. For example, in the descriptors for the original Element H (*Sufficiency of prototyping*), one feature of each of the score points is a focus on "stated goals." Only at score point 4 was there a reference to "primary" stated goals, however, giving that the status of "orphan." In the original Element B (*Problem justification*), the absence in the descriptor for score point 5 of any reference to sources consulted, which was included at all other score points, can be regarded as a "widow." Revision to address such problematic uses of key words or phrases is advisable since, when readers apply evaluative criteria in order to identify the most appropriate score point level to describe a product or performance, the presence of "orphans" and "widows" needlessly complicates the process.

Problem 3: Redundancy in descriptors

Another potential source of confusion to which some rubrics are prone is the attribution of identical features to more than one score point level. Unless a rubric is structured so that features are clearly cumulative (e.g., a score point 1 includes feature A, score point 2 includes feature A plus B, and so on), any perceived redundancy may be frustrating and lead to scoring error. In several of the original rubrics comprising the EDPPSR, considerable effort was required to determine exactly what differences separated one point from another (and if, in fact, redundancy was symptomatic of a scale that was forced to be too broad to permit meaningful distinctions between levels). This is illustrated in the example in Table 2, below:

Table 2: An Excerpt from Original Element H (*Sufficiency of prototyping*) Before Revision[5]

4) The prototype or prototyping process submitted included or met at least the following criteria;

- The prototyping design went through an iterative process itself and was not a first design attempt;
- clear explanations were included about choices made as the prototyping design evolved;
- reflective statements about how the final iteration could be improved for testing purposes were explained;
- the final prototyping iteration submitted was explained and constructed with enough detail that some level of objective data relating to the value of the design at addressing EACH one of the stated goals could be determined through actual testing, mathematical modeling, or detailed expert reviews.* [6]
- ALL attributes of the unique solution that could be tested or modeled mathematically were addressed in the prototyping design:
- ALL attributes of the unique solution that could not be tested or modeled and would require the review and recommendation of an expert was explained with sufficient justification
- At least one portion, facet, or attribute of the prototyping process was so well designed and constructed that it could be tested definitively with respect to the ability of the solution design to address at least one of the primary stated goals of the design.

5) The prototype or prototyping process submitted included or met at least the following criteria;

- The prototyping design went through an iterative process itself and was not a first design attempt;
- clear explanations were included about choices made as the prototyping design evolved;

[5] Note that the text of the fall 2010 draft rubric here and elsewhere in this paper is presented exactly as it appeared in that document, without correction of errors in grammar and mechanics (e.g., agreement, punctuation).

[6] No explanation for this asterisk appears in the draft rubric.

- reflective statements about how the final iteration could be improved for testing purposes were explained;
- the final prototyping iteration submitted was explained and constructed with enough detail that some level of objective data relating to the value of the design at addressing EACH one of the stated goals could be determined through actual testing, mathematical modeling, or detailed expert reviews.*
- ALL attributes of the unique solution that could be tested or modeled mathematically were addressed in the prototyping design:
- ALL attributes of the unique solution that could not be tested or modeled and would require the review and recommendation of an expert was explained with sufficient justification
- The prototyping process was so well designed and constructed that it could be tested definitively with respect to the ability of the solution design to address most, if not all, of the stated goals of the design.

The first six bulleted descriptors for the two score points in this excerpt are identical; only the seventh bullet is slightly different, implying by omission a distinction between "at least one portion, facet, or attribute" of the prototyping process and more than one (without making clear whether that means most or all of them). At score point 5, "at least one" goal is changed to "most, if not all" goals. Although this is a feasible way to articulate one difference between these performance levels, any differentiation between them is confounded by the shift (previously identified under Problem 2, above) from "primary stated goals" to "stated goals," the wording that had been used for all previous score points under this attribute as well.

In the case of the EDPPSR, redundancy was eliminated during revision by standardizing the format across element rubrics so that none of them contained identical descriptors across score points. One way for creators and consumers of other rubrics to avoid redundancy is to steer clear altogether of the cumulative approach to differentiating score points which is evident in some rubrics circulating in print and online; alternatively, they must also engage in revision to eliminate any needless repetition or superfluous language in that approach. The superfluous use of underlined text in the excerpt from the original Element H rubric (Table 2) may also serve as a reminder that the judicious and meaningful use of text features like underlining, capitalization, boldface or italics to highlight particular words and phrases can help differentiate among levels of performance in score point descriptors. Redundancy in use of text features, however, like redundancy in language, will defeat that purpose.

Problem 4: Inconsistency in focus of qualifiers

Rubrics typically differentiate levels of performance by describing gradations of various sorts and distinguishing between the degrees to which various types of evidence are present. Score scale descriptors often use qualitative words or phrases to capture the frequency of a particular observed behavior (e.g., consistently, generally, sometimes, rarely, never). Sometimes categories of performance are described in terms of scope (e.g., well-substantiated, generally substantiated, partially substantiated, minimally substantiated, unsubstantiated). Other rubrics use numerical criteria (e.g., one source for one point, two sources for two points, etc.); however, if considering this approach to distinguishing among score points (whether in a newly crafted rubric or one adopted from another source), it would be wise to ask, as Arter and McTighe (2001, p. 46) suggest, "If counting the number of something (such as the number of references at the end of a research report) is included as an indicator, such counts really *are* indicators of quality."

Since the original versions of the design portfolio rubrics were not characterized by that approach, it was unnecessary to caution its authors that "you shouldn't score by counting on your fingers" (Goldberg, 1995). However, the original draft stage element rubrics exhibited some tendency to shift focus, going from one basis for differentiating score point levels to another—for example shifting from suitability to frequency or frequency to quality, when describing a given attribute at different levels of performance. The original score point descriptors for Element A (*Identification and definition of the problem*) in Table 3 illustrate this flaw in addition to other consistency issues.

Table 3: **Inconsistencies in Focus**

0) Evidence of the process for identifying and defining the problem was not present in this submission. One or both were missing from the submission.
1) The information presenting describing the problem or the problem statement itself (or both) were so general in their articulation that it would be difficult or impossible to gauge the effectiveness of any of the project's results.
2) Both the nature and background of the problem and a problem statement were submitted but together the objective purpose of the project was left to some interpretation.
3) Sufficient information was presented to explain the nature and background of the problem in an objective fashion and an equally objective problem statement was presented.
4) Both the problem background information and the problem statement are clear, objective, and focused. The problem statement defines a measurable cause and effect relationship.
5) The level of detail and depth of both the explanation of the problem and the problem statement are measurably objective, well researched and presented no area for subjective interpretation of purpose.

When descriptors are written in this way, the essential characteristics that distinguish levels of performance shift from score point to score point. Focus swings from the degree of development to clarity, and then to objectivity, instead of addressing the degree or extent to which one or more of these variables characterize each of the different levels of performance. It is far better to determine the essential characteristics of the desired product or performance, and then to select words and phrases to describe how much, how often, and/or how consistently those characteristics are evident. It was <u>that</u> approach that informed the revision of this attribute (See Table 4 for the revised Element A rubric).

Table 4: Score point descriptors for revised EDPPSR Element A

5 The problem is clearly and objectively identified and defined with considerable depth, and it is well elaborated with specific detail; the justification of the problem highlights the concerns of many primary stakeholders and is based on comprehensive, timely, and consistently credible sources; it offers consistently objective detail from which multiple measurable design requirements can be determined.

4 The problem is clearly and objectively identified and defined with some depth, and it is generally elaborated with specific detail; the justification of the problem highlights the concerns of some primary stakeholders and is based on various timely and generally credible sources; it offers generally objective detail from which multiple measurable design requirements can be determined.

3 The problem is somewhat clearly and objectively identified and defined with adequate depth, and it is sometimes elaborated with specific detail, although some information intended as elaboration may be imprecise or general; the justification of the problem highlights the concerns of at least a few primary stakeholders and is based on at least a few sources which are timely and credible; although not all information included may be objective, the justification of the problem offers enough objective detail to allow at least a few measurable design requirements to be determined.

2 The problem is identified only somewhat clearly and objectively and defined in a manner that is somewhat superficial and/or minimally elaborated with specific detail; the justification of the problem highlights the concerns of only one or two primary stakeholders and/or may be based on insufficient sources or ones that are outdated or of dubious credibility; although little information included is objective, the justification of the problem offers enough objective detail to allow at least a few design requirements to be determined; however, these may not be ones that are measurable.

1) The identification and/or definition of the problem is unclear, is unelaborated, and/or is clearly subjective; any intended justification of the problem does not highlight the concerns of any primary stakeholders and/or is based on sources that are overly general, outdated, and/or of dubious credibility; information included is insufficient to allow for the determination of any measurable design requirements.

0 The identification and/or definition of the problem are missing OR cannot be inferred from information included. A justification of the problem is missing, cannot be inferred from information included as evidence, OR is essentially only the opinion of the researcher.

Problem 5: Limited routes to partial credit

It is relatively easy to identify criteria for the highest score point on a scale—students will have done everything called for fully and exceptionally well. It is similarly easy to identify criteria for the lowest score point—students will have done little or nothing to demonstrate proficiency. It is much harder to identify criteria for the remaining levels of performance. That is primarily because there is often more than one way to demonstrate partial or overly general understanding and emerging but not yet mastered levels of skill. Nevertheless, examination of the literature on rubric development failed to find articulation of various acceptable routes to a given score point ever mentioned as a key principle of design, other than one adage of the author's that there is (or should be) more than a single way to earn partial credit (Goldberg, 1994).

Instead of "all or nothing," well-crafted score point descriptors need to articulate the multiple routes to the range of scores that can be assigned to entries that are not exemplary. This was accomplished during revision of the EDPPSR by identifying some features as ones that might be evident in an entry at a particular score point level (but need not characterize that entry). With application of a focused holistic approach to scoring, raters—whether students, teachers, or trained readers—are to ask themselves which descriptor is the "best fit." An entry that is more like a 3 than a 2 should receive a 3; if more like a 4 than a 5, it should receive a 4.

Judicious use of the conjunction "and/or" (although frowned upon by some grammarians), is another way to make clear that one or more features of an attribute may in evidence (and are not all required although all may be present to some degree). Thus, for example, in the original version of Element F, every performance level from 0-5 described the degree to which the entry provided evidence that the proposed design solution was supported "with math, science, and engineering principles related to the design constraints, project goals, and design criteria" (underlines are the author's). However, entries characteristic of novice or developing levels of performance are as likely to provide strong support from only one discipline—mathematics, for example—as they are to provide only some evidence of support from all three disciplines. The introduction of and/or before the series of disciplines and before the list of design concerns (constraints, goals, and criteria) opened up opportunities to reach a particular performance level.

Score point descriptors may even identify two or more different pathways rather than, or in addition to, this "mix and match" approach. For the EDPPSR, this was limited to descriptors for score point 0, as in the example from the 2011 version of Element I (*Testing, data collection and analysis*) below:

> Any test(s) for requirement(s) or attempts at physical or mathematical modeling fail to demonstrate even minimal understanding of testing procedure, including the gathering and analysis of resultant data; OR there is no evidence of testing or physical or mathematical modeling to address any requirements.

Particularly if there is no plan to use condition codes—designations for non-scorable responses such as ones that are missing (M), or off-topic/off-task (OT)—laying out such alternative options makes sense. However, in virtually any instance in which descriptors address more than one attribute (typically in holistic rather than analytic rubrics), it is critical that creators and consumers of rubrics recognize that students often demonstrate related skills and understandings to different degrees—something that should be captured in a well-crafted rubric.

Problem 6: Unevenness in incremental levels of performance

The literature on rubric development includes the recommendation that distinctions between score levels be clear (Dornisch and McLoughlin, 2006; Moskal, 2003); however, this recommendation has not specifically addressed the need for evenness of increments between those levels. In workshops on rubric development, the author often compares a rubric score scale to a staircase. Ascending and descending, each step is generally evenly spaced, in order to move easily without having to make adjustments. We can navigate stairs even in the dark, since we assume that each step will be exactly the same distance from the one above and below. Imagine what would happen, however, if we encountered a set of stairs that varied in height! Although the consequences of uneven increments in a rubric are arguably not as serious as those in a set of stairs, they still warrant attention.

This weakness in rubric design is illustrated in the descriptors for the original Element F (see Table 5 below), score points 2 and 3 seem quite close to each other, with a much greater "step" between score points

3 and 4. That unevenness is made by clear by comparing the distinctions between score point 4 and 5 as well.

Table 5: Score point descriptors exhibiting variable increments between levels

2) Math, science and design principles relative to the design constraints, project goals, and design criteria have been submitted to document technical understanding of the problem and to justify that the design has merit as a possible solution to the problem stated. Each functional claim of the proposed solution is backed up with sound and detailed evidence from this perspective. However, at least one of the functional or beneficial claims of the design was missing this support or the information presented was incorrect.
3) Math, science and design principles relative to the design constraints, project goals, and design criteria have been submitted to document technical understanding of the problem and to justify that the design has merit as a possible solution to the problem stated. Each functional claim of the proposed solution is backed up with sound and detailed evidence from this perspective.
4) Math, science and design principles relative to the design constraints, project goals, and design criteria have been submitted to document technical understanding of the problem and to justify that the design has merit as a possible solution to the problem stated. Each functional claim of the proposed solution is backed up with sound and detailed evidence from this perspective. The information has been reviewed and verified by a qualified consultant or project mentor. The reviewer's comments concerning each piece of information have been submitted with this section.
5) Math, science and design principles relative to the design constraints, project goals, and design criteria have been submitted to document technical understanding of the problem and to justify that the design has merit as a possible solution to the problem stated. Each functional claim of the proposed solution is backed up with sound and detailed evidence from this perspective. The information has been reviewed and verified by more than one qualified consultant or project mentor. Each reviewer's comments concerning each piece of information have been submitted with this section.

With correction, each score point descriptor defined a level of performance distinguished to an equal degree from those above and below it—an aspect of technical quality for which all rubrics should be checked.

Problem 7: Inconsistencies across suite of rubrics

The various problems to which rubrics are prone are compounded when a rubric is part of a suite or set, as in the case of the EDPPSR. Beyond ensuring consistency of language and format wherever appropriate, the revision of the original set of attribute-based score scales and descriptors addressed a more critical concern—that there be consistency in the meaning of each of the score point levels. In the original version, inconsistency was most evident at the 0 level. Across attributes, descriptors for this score point sometimes referred to "little or no evidence" while other times indicated that the key evidence was missing. It was difficult to determine the meaning of score point 0 across the suite of rubrics. In some draft stage attribute rubrics, overlap between score points 0 and 1 added to confusion as to the traits of a response at the lowest level of performance. In one instance, for example (see Table 6 below), criteria are identical except for those in the third and fifth bullets. However, the stem (in boldface font) is identical for these score points, making it is impossible to know what score to assign an entry missing any one of the other bulleted criteria.

Table 6: Confounding of Criteria in Original Rubric for Element I (*Sufficiency of testing*)

***0) The testing procedure or set of procedures submitted was missing at least one of the following criteria or insufficient in detail to meet one of the following criteria*:**

- A testing procedure or process that targeted most of the stated design goals;
- An clear and logical explanation of how the testing procedure would yield objective data regarding the effectiveness of the design was submitted;;
- Some portion of the testing process was either attempted in an effort to evaluate the effectiveness of the design;

- The results and description of the testing procedure or process was explained with generous and appropriate use of pictures, graphs and charts;
- ALL attributes that and would require the review and recommendation of an expert was explained with sufficient justification;
- A detailed suggestion for improvement of the testing procedure was submitted;

1) The testing procedure or set of procedures submitted was missing at least one of the following criteria or insufficient in detail to meet one of the following criteria:

- A testing procedure or process that targeted most of the stated design goals;
- An clear and logical explanation of how the testing procedure would yield objective data regarding the effectiveness of the design was submitted;;
- At least one part of the procedure or process submitted involving actual testing or mathematical modeling was attempted;
- The results and description of the testing procedure or process was explained with generous and appropriate use of pictures, graphs and charts;
- ALL attributes that and would require the review and recommendation of an expert was explained with sufficient justification and the results of at least one of those reviews were submitted;
- A detailed suggestion for improvement of the testing procedure was submitted;

This confounding of criteria was evident, although less often, at other score point levels as well in several of the draft-stage rubrics. To facilitate standardization of meaning for each score point on the six-point scale for the EDPPSR, a generic scoring scale was developed which served as a template for each of the element rubrics (see Table 7 below). Through subsequent revision, a consistent relationship between the generic scale and specific performance criteria for each EDPPSR element was ensured.

Table 7: EDPPSR Generic Scoring Scale and Descriptors

5	**Exemplary**: Demonstrates thorough and penetrating understanding of key concepts; exhibits copious evidence of attainment of skills
4	**Advanced**: Demonstrates considerable understanding of key concepts; exhibits considerable (substantial) evidence of attainment of skills
3	**Proficient**: Demonstrates general/adequate understanding of key concepts; exhibits adequate evidence of attainment of skills
2	**Developing**: Demonstrates a partial understanding of key concepts; exhibits some evidence of attainment of skills
1	**Novice**: Demonstrates a lack of/little understanding of key concepts; exhibits minimal evidence of attainment of skills
0	**No evidence** (No evidence of engagement, pre-engagement): Demonstrates no understanding of key concepts; exhibits no evidence of attainment of skills

Next Steps Towards an Operational Rubric

Without piloting and subsequent refinement, the necessity for which becomes evident as a result, any rubric must be regarded as only provisional. That applied to the EDPPSR as well—even more so, perhaps, given that although the EDPPSR reflected input over several years from a wide array of educators and other professionals involved in engineering, there had to date been no formal trial of the rubric. Supported by funding from the Kern Family Foundation, that situation was rectified in June 2011.The primary purpose of that scoring pilot was to address a series of questions that ought to underlie any investigation of the efficacy of scoring rubric, which are generalized below to apply to any rubric or suite of rubrics:

- Does review of sample student work reveal that the rubric has not yet captured any elements critical to the construct being assessed?
- What, if anything, is missing from any score point descriptors that might assist rater(s) in reaching a score decision?
- Does the rubric contain any evidence of redundancy?

- Does the rubric contain any instances of ambiguity?
- Are there any instances in which adherence to the language of a score point descriptor leads to cognitive dissonance (the perception that the assigned score does not "fit")?
- What evidence, if any, is there to support an expansion or reduction in score scale?

After all key questions were answered and engineering design content corrected or confirmed, additional refinement of the EDPPSR took place. Throughout this post-pilot revision process, extreme care was taken—as it must be when revising any rubric—to ensure that the technical quality of the EDPPSR so carefully addressed during revision of the original draft was maintained. Otherwise, it would have been far too easy to disrupt parallelism, to introduce new "orphan" and "widow" words and terms, and to create new inconsistencies in gradation, focus, and/or incremental distinctions between score point levels—to "go back to square one" so to speak.

The initial, systematic effort to revise the draft rubrics was—and has always been acknowledged to be—only the first of what must necessarily be a series of enterprises needed to refine and finalize the EDPPSR. There remain many possible revisions documented as a result of the scoring pilot (Goldberg, 2011), but these are being held in abeyance until hands-on experience by experts in engineering design education and practice yields further evidence supporting additional changes. Such evidence has been marshaled through scoring workshops conducted through an NSF Promoting Research and Innovation in Methodologies for Evaluation (PRIME) award (National Science Foundation, 2011) as part of a three-year investigation into the validity and reliability of the EDPPSR before it can be used for such high-stakes purposes program admission, course assignment, or advanced placement credit, the last a goal harkening back to 2006.

Meanwhile, at present, the post-pilot (2011) version of the EDPPSR is the one authorized for dissemination and use. It serves as the portfolio template on the Innovation Portal (www.InnovationPortal.org), an open-source online platform for engineering design process e-portfolios. The full text of the rubric, along with a growing body of scored sample entries annotated using the language of the EDPPSR, can be accessed on that site and is being used by an ever-increasing number of high school and college educators and their students as an instructional and formative assessment tool. At the same time, conversation with the College Board is ongoing about the development and implementation of an AP® in engineering design (Groves, 2012; Robelin, 2013) in which the EDPPSR plays a central part. What changes—if any—to the rubric that may be deemed necessary or advisable in order that it become a framework for that exam, while retaining its integrity as a tool for instruction and for other assessment purposes, remain to be seen. It is to be hoped that continued regard for principles and practices to ensure technical quality and content integrity will inform those changes. Attention to those same principles and practices can enhance the work of educators in other disciplines than engineering design, as they also endeavor to create or revise rubrics to evaluated students' work products or performances.

Implications for the Development and Revision of Other Rubrics

Resources on rubric design make clear that beyond identifying criteria and defining levels of performance, draft rubrics must be subject to piloting and subsequent revision and refinement. Rubric revision should be regarded not as a linear process — one among a series of steps in rubric development — but as a recursive one, informed by students, teachers, and other end-users. The revision "rules" outlined in this paper are not intended solely to inform initial rubric development and revision. Any changes considered post-pilot (and even beyond) to ensure clarity and ease of use, ought to be subject to repeated scrutiny based on these rules to ensure technical quality.

A quick perusal of the literature (of which references for this paper are representative) makes clear that interest in rubrics was high at the beginning of the millennium and has ebbed somewhat since then. That phenomenon may be explained by the interest in, and attention to, performance-based learning and assessment in the decade or so leading up to No Child Left Behind—during which time, in the words of one authority on rubrics, they were "becoming increasingly popular with educators moving toward more authentic, performance-based assessments" (Andrade, 1997), and the subsequent reduction in performance assessment in favor of multiple-choice tests. Although a commitment to the creation of products and performances to demonstrate learning has survived in many classrooms, with the result that rubrics continue to be created and used, attention to issues of design and technical quality

seems to have diminished. That attention is very likely to revive, however, (and indeed needs to) with the implementation of the Common Core State Standards for English language arts and mathematics and the introduction of the Next Generation Science Standards, the assessment of at least some of which will require the development and use of scoring rubrics. Indeed, signs of this are already evident with updated online rubric generators and rubric banks (see, for example, http://www.schrockguide.net/assessment-and-rubrics.html). Furthermore, as Linda Darling-Hammond has stressed (Darling-Hammond, 2014), these new assessment systems—while including open-ended tasks that are likely to require scoring rubrics—will not address everything that students should know and be able to do. School districts and states are recognizing and responding to the need to supplement, and provide multiple measures through, performance tasks and portfolio assessment. Those instruments will require rubrics, making this a critical time to marshal and apply all that we know about ensuring their technical quality. Consideration of the revision "rules" illustrated in this paper may contribute towards that end.

References

Abts, L. (2011). "Analysis of the barriers, constraints and issues for dual credit and/or an Advanced Placement ® pathway for introduction to engineering/design. Conference paper, American Society of Engineering Educators. Retrieved July 1, 2013 from www.**asee**.org/public/conferences/1/author_index/66429

Andrade, H. G. (December 1996-January 1997). "Understanding rubrics." retrieved on May 4, 2014 from https://learnweb.harvard.edu/alps/thinking/docs/rubricar.htm. Also available under author's maiden name (Goodrich, H) in *Educational Leadership 54* (4), 14-19.

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

Darling-Hammond, L. (January-February 2014). "Testing to, and beyond, the Common Core." *Principal*. Retrieved June 28, 2014 from http://www.naesp.org/principal-januaryfebruary-2014-assessments-evaluations-and-data/testing-and-beyond-common-core

Dornisch, M.M. and McLoughlin, A.S. (2006). "Limitations of web-based rubric resources: Addressing the challenges." *Practical Assessment, Research & Evaluation*, *11* (3). Retrieved June 25, 2014 from http://pareonline.net/pdf/v11n3.pdf.

Goldberg, G. (2011). *Engineering Design Process Portfolio Scoring Rubric (EDPPSR): Scoring Pilot Final Report*. Unpublished report. College Park, MD: University of Maryland.

Goldberg, G. (1995). "Gail's axioms of scoring performance assessment." In *Implementing Performance Assessments: A Guide to Classroom, School and System Reform* (p. 30), (Monty Neill, et. al., eds.). Cambridge, MA: FairTest.

Goldberg, G. (1994). "Learning the score: What teachers discover from scoring performance assessment tasks." *Teaching Thinking and Problem Solving, 16* (1), 1, 3-6.

Groves, J., R. Reshetar, M. Schroll, L. Abts. (2012). "The development and implementation of a potential AP for engineering design using a rubric based e-portfolio. Conference presentation at 2012 ASEE. Retrieved July 1, 2013 from *http://www.asee.org/conferences-and-events/conferences/edi/2012/program-schedule/Leigh_R._Abts_EDI_2012_Presentation.pdf*

Groves, J., L. Abts, G. Goldberg. (2014). "Using an Engineering Design Process Portfolio Scoring Rubric to Structure Online High School Engineering Education." Conference presentation at 2014 ASEE. Retrieved June 24, 2014 from http://www.asee.org/public/conferences/32/papers/10738/view

Moskal, B. M. (2003). "Recommendations for developing classroom performance assessments and scoring rubrics." *Practical Assessment, Research & Evaluation*, *8* (14). Retrieved July 10 2013 from http://pareonline.net/getvn.asp?v=8&n=14

National Science Foundation Research Spending & Results Award Detail. (2011). Retrieved July 22, 2013 from http://www.research.gov/research-portal/appmanager/base/desktop;jsessionid=z5RlPCnWg4XSdWR1TxldhhL2syr12QfWwjGzxypypZWzLhtdhQyy!811304842!1301015568?_nfpb=true&_windowLabel=awardStatistics_1&_urlType=action&awardStatistics_1_action=viewAwardDetailEvent_rsr&awardStatistics_1_fedAwrdId=1118755

Perlman, C. "An introduction to performance assessment scoring rubrics," in *Understanding Scoring Rubrics: A Guide for Teachers* (Boston, C. ed). University of Maryland College Park, 2002: Clearinghouse on Assessment and Evaluation.

Popham, J. W. (1997). What's wrong—and what's right—with rubrics." *Educational Leadership*, *55* (2), 72-75.

Robelen, E. "AP Engineering May Be on the Horizon." (2013). *Education Week*, March 29, 2013. Retrieved July 1, 2013 from http://blogs.edweek.org/edweek/curriculum/2013/03/ap_engineering_may_be_on_the_horizon.html

Tierney, R. & M. Simon. (2004). "What's still wrong with rubrics: Focusing on the consistency of performance criteria across scale levels." *Practical Assessment, Research,*

& Evaluation, 9 (2). Retrieved April 27, 2014 from http://PAREonline.net/getvn.asp?v=9&n=2.

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

Note:

The views, opinions, and positions expressed in this paper are those of the author and do not necessarily reflect those of other individuals referred to therein or those of the Kern Family Foundation or National Science Foundation which have provided funds in support of work on the EDPPSR.

Acknowledgement:

The author wishes to thank Jay McTighe for his helpful feedback on an earlier draft of this paper.

Citation:

Goldberg, Gail Lynn (2014). Revising an Engineering Design Rubric: A Case Study Illustrating Principles and Practices to Ensure Technical Quality of Rubrics. *Practical Assessment, Research & Evaluation*, 19(8). Available online: http://pareonline.net/getvn.asp?v=19&n=8

Author:

Gail Lynn Goldberg
Goldberg Consulting
2766 Westminster Road
Ellicott City, MD 21043
Email: gailgoldbergconsulting [at] gmail.com