Practical Assessment, Research & Evaluation

A peer-reviewed electronic journal.



Volume 7, Number 10, November, 2000 ISSN=1531-7714

# Scoring Rubric Development: Validity and Reliability

*Barbara M. Moskal* & *Jon A. Leydens*
Colorado School of Mines

In Moskal (2000), a framework for developing scoring rubrics was presented and the issues of validity and reliability were given cursory attention. Although many teachers have been exposed to the statistical definitions of the terms "validity" and "reliability" in teacher preparation courses, these courses often do not discuss how these concepts are related to classroom practices (Stiggins, 1999). One purpose of this article is to provide clear definitions of the terms "validity" and "reliability" and illustrate these definitions through examples. A second purpose is to clarify how these issues may be addressed in the development of scoring rubrics. Scoring rubrics are descriptive scoring schemes that are developed by teachers or other evaluators to guide the analysis of the products and/or processes of students' efforts (Brookhart, 1999; Moskal, 2000). The ideas presented here are applicable for anyone using scoring rubrics in the classroom, regardless of the discipline or grade level.

## Validity

Validation is the process of accumulating evidence that supports the appropriateness of the inferences that are made of student responses for specified assessment uses. Validity refers to the degree to which the evidence supports that these interpretations are correct and that the manner in which the interpretations are used is appropriate (American Educational Research Association, American Psychological Association & National Council on Measurement in Education, 1999). Three types of evidence are commonly examined to support the validity of an assessment instrument: content, construct, and criterion. This section begins by defining these types of evidence and is followed by a discussion of how evidence of validity should be considered in the development of scoring rubrics.

### *Content-Related Evidence*

Content-related evidence refers to the extent to which a student's responses to a given assessment instrument reflects that student's knowledge of the content area that is of interest. For example, a history exam in which the questions use complex sentence structures may unintentionally measure students' reading comprehension skills rather than their historical knowledge. A teacher who is interpreting a student's incorrect response may conclude that the student does not have the appropriate historical knowledge when actually that student does not understand the questions. The teacher has misinterpreted the evidence—rendering the interpretation invalid.

Content-related evidence is also concerned with the extent to which the assessment instrument adequately samples the content domain. A mathematics test that primarily includes addition problems would provide inadequate evidence of a student's ability to solve subtraction, multiplication and division problems. Correctly computing fifty addition problems and two multiplication problems does not provide convincing evidence that a student can subtract, multiply or divide.

Content-related evidence should also be considered when developing scoring rubrics. The task shown in Figure 1 was developed by the Quantitative Understanding: Amplifying Student Achievement and Reasoning Project (Lane, et. al, 1995) and requests that the student provide an explanation. The intended content of this task is decimal density. In developing a scoring rubric, a teacher could unintentionally emphasize the nonmathematical components of the task. For example, the resultant scoring criteria may emphasize sentence structure and/or spelling at the expense of the mathematical knowledge that the student displays. The student's score, which is interpreted as an indicator of the student's mathematical knowledge, would actually be a reflection of the student's grammatical skills. Based on this scoring system, the resultant score would be an inaccurate measure of the student's mathematical knowledge. This discussion does not suggest that sentence structure and/or spelling cannot be assessed through this task. If the assessment is intended to examine sentence structure, spelling, *and* mathematics, then the score categories should reflect all of these areas.

Figure 1. Decimal Density Task

| |
|---|
| *Dena tried to identify all the numbers between 3.4 and 3.5. Dena said, "3.41, 3.42, 3.43, 3.44, 3.45, 3.46, 3.47, 3.48 and 3.49. That's all the numbers that are between 3.4 and 3.5."* |

*Nakisha disagreed and said that there were more numbers between 3.4 and 3.5.*

*A. Which girl is correct?*

*Answer:*

*B. Why do you think she is correct?*

*Construct-Related Evidence*

Constructs are processes that are internal to an individual. An example of a construct is an individual's reasoning process. Although reasoning occurs inside a person, it may be partially displayed through results and explanations. An isolated correct answer, however, does not provide clear and convincing evidence of the nature of the individual's underlying reasoning process. Although an answer results from a student's reasoning process, a correct answer may be the outcome of incorrect reasoning. When the purpose of an assessment is to evaluate reasoning, both the product (i.e., the answer) and the process (i.e., the explanation) should be requested and examined.

Consider the problem shown in Figure 1. Part A of this problem requests that the student indicate which girl is correct. Part B requests an explanation. The intention of combining these two questions into a single task is to elicit evidence of the students' reasoning process. If a scoring rubric is used to guide the evaluation of students' responses to this task, then that rubric should contain criteria that addresses both the product and the process. An example of a holistic scoring rubric that examines both the answer and the explanation for this task is shown in Figure 2.

Figure 2. Example Rubric for Decimal Density Task

| | |
|---|---|
| **Proficient:** | Answer to part A is Nakisha. Explanation clearly indicates that there are more numbers between the two given values. |
| **Partially Proficient:** | Answer to part A is Nakisha. Explanation indicates that there are a finite number of rational numbers between the two given values. |
| **Not Proficient:** | Answer to part A is Dana. Explanation indicates that all of the values between the two given values are listed. |

Note. This rubric is intended as an example and was developed by the authors. It is not the original QUASAR rubric, which employs a five-point scale.

Evaluation criteria within the rubric may also be established that measure factors that are unrelated to the construct of interest. This is similar to the earlier example in which spelling errors were being examined in a mathematics assessment. However, here the concern is whether the elements of the responses being evaluated are appropriate indicators of the underlying construct. If the construct to be examined is reasoning, then spelling errors in the student's explanation are irrelevant to the purpose of the assessment and should not be included in the evaluation criteria. On the other hand, if the purpose of the assessment is to examine spelling and reasoning, then both should be reflected in the evaluation criteria. Construct-related evidence is the evidence that supports that an assessment instrument is completely and only measuring the intended construct.

Reasoning is not the only construct that may be examined through classroom assessments. Problem solving, creativity, writing process, self-esteem, and attitudes are other constructs that a teacher may wish to examine. Regardless of the construct, an effort should be made to identify the facets of the construct that may be displayed and that would provide convincing evidence of the students' underlying processes. These facets should then be carefully considered in the development of the assessment instrument and in the establishment of scoring criteria.

*Criterion-Related Evidence*

The final type of evidence that will be discussed here is criterion-related evidence. This type of evidence supports the extent to which the results of an assessment correlate with a current or future event. Another way to think of criterion-related evidence is to consider the extent to which the students' performance on the given task may be generalized to other, more relevant activities (Rafilson, 1991).

A common practice in many engineering colleges is to develop a course that "mimics" the working environment of a practicing engineer (e.g., Sheppard, & Jeninson, 1997; King, Parker, Grover, Gosink, & Middleton, 1999). These courses are specifically designed to provide the students with experiences in "real" working environments. Evaluations of these courses, which sometimes include the use of scoring rubrics (Leydens & Thompson, 1997; Knecht, Moskal & Pavelich,

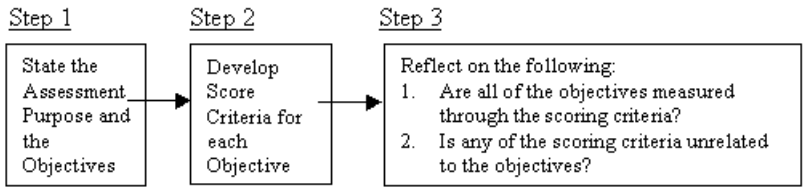
2000), are intended to examine how well prepared the students are to function as professional engineers. The quality of the assessment is dependent upon identifying the components of the current environment that will suggest successful performance in the professional environment. When a scoring rubric is used to evaluate performances within these courses, the scoring criteria should address the components of the assessment activity that are directly related to practices in the field. In other words, high scores on the assessment activity should suggest high performance outside the classroom or at the future work place.

*Validity Concerns in Rubric Development*

Concerns about the valid interpretation of assessment results should begin before the selection or development of a task or an assessment instrument. A well-designed scoring rubric cannot correct for a poorly designed assessment instrument. Since establishing validity is dependent on the purpose of the assessment, teachers should clearly state what they hope to learn about the responding students (i.e., the purpose) and how the students will display these proficiencies (i.e., the objectives). The teacher should use the stated purpose and objectives to guide the development of the scoring rubric.

In order to ensure that an assessment instrument elicits evidence that is appropriate to the desired purpose, Hanny (2000) recommended numbering the intended objectives of a given assessment and then writing the number of the appropriate objective next to the question that addresses that objective. In this manner, any objectives that have not been addressed through the assessment will become apparent. This method for examining an assessment instrument may be modified to evaluate the appropriateness of a scoring rubric. First, clearly state the purpose and objectives of the assessment. Next, develop scoring criteria that address each objective. If one of the objectives is not represented in the score categories, then the rubric is unlikely to provide the evidence necessary to examine the given objective. If some of the scoring criteria are not related to the objectives, then, once again, the appropriateness of the assessment and the rubric is in question. This process for developing a scoring rubric is illustrated in Figure 3.

Figure 3. Evaluating the Appropriateness of Scoring Categories to a Stated Purpose




Reflecting on the purpose and the objectives of the assessment will also suggest which forms of evidence—content, construct, and/or criterion—should be given consideration. If the intention of an assessment instrument is to elicit evidence of an individual's knowledge within a given content area, such as historical facts, then the appropriateness of the content-related evidence should be considered. If the assessment instrument is designed to measure reasoning, problem solving or other processes that are internal to the individual and, therefore, require more indirect examination, then the appropriateness of the construct-related evidence should be examined. If the purpose of the assessment instrument is to elicit evidence of how a student will perform outside of school or in a different situation, criterion-related evidence should be considered.

Being aware of the different types of evidence that support validity throughout the rubric development process is likely to improve the appropriateness of the interpretations when the scoring rubric is used. Validity evidence may also be examined after a preliminary rubric has been established. Table 1 displays a list of questions that may be useful in evaluating the appropriateness of a given scoring rubric with respect to the stated purpose. This table is divided according to the type of evidence being considered.

Table 1: Questions to Examine Each Type of Validity Evidence

| Content | Construct | Criterion |
|---|---|---|
| 1. Do the evaluation criteria address any extraneous content?<br>2. Do the evaluation criteria of the scoring rubric address all aspects of the intended content?<br>3. Is there any content addressed in the task that should be evaluated through the rubric, but is not? | 1. Are all of the important facets of the intended construct evaluated through the scoring criteria?<br>2. Is any of the evaluation criteria irrelevant to the construct of interest? | 1. How do the scoring criteria reflect competencies that would suggest success on future or related performances?<br>2. What are the important components of the future or related performance that may be evaluated through the use of the assessment instrument?<br>3. How do the scoring criteria measure the important components of the future or related |

| | | |
|---|---|---|
| | | performance?<br>4. Are there any facets of the future or related performance that are not reflected in the scoring criteria? |

Many assessments serve multiple purposes. For example, the problem displayed in Figure 1 was designed to measure both students' knowledge of decimal density and the reasoning process that students used to solve the problem. When multiple purposes are served by a given assessment, more than one form of evidence may need to be considered.

Another form of validity evidence that is often discussed is "consequential evidence". Consequential evidence refers to examining the consequences or uses of the assessment results. For example, a teacher may find that the application of the scoring rubric to the evaluation of male and female performances on a given task consistently results in lower evaluations for the male students. The interpretation of this result may be the male students are not as proficient within the area that is being investigated as the female students. It is possible that the identified difference is actually the result of a factor that is unrelated to the purpose of the assessment. In other words, the completion of the task may require knowledge of content or constructs that were not consistent with the original purposes. Consequential evidence refers to examining the outcomes of an assessment and using these outcomes to identify possible alternative interpretations of the assessment results (American Educational Research Association, American Psychological Association & National Council on Measurement in Education, 1999).

## Reliability

Reliability refers to the consistency of assessment scores. For example, on a reliable test, a student would expect to attain the same score regardless of when the student completed the assessment, when the response was scored, and who scored the response. On an unreliable examination, a student's score may vary based on factors that are not related to the purpose of the assessment.

Many teachers are probably familiar with the terms "test/retest reliability," "equivalent-forms reliability," "split half reliability" and "rational equivalence reliability" (Gay, 1987). Each of these terms refers to statistical methods that are used to establish consistency of student performances within a given test or across more than one test. These types of reliability are of more concern on standardized or high stakes testing than they are in classroom assessment. In a classroom, students' knowledge is repeatedly assessed and this allows the teacher to adjust as new insights are acquired.

The two forms of reliability that typically are considered in classroom assessment and in rubric development involve rater (or scorer) reliability. Rater reliability generally refers to the consistency of scores that are assigned by two independent raters and that are assigned by the same rater at different points in time. The former is referred to as "interrater reliability" while the latter is referred to as "intrarater reliability."

*Interrater Reliability*

Interrater reliability refers to the concern that a student's score may vary from rater to rater. Students often criticize exams in which their score appears to be based on the subjective judgment of their instructor. For example, one manner in which to analyze an essay exam is to read through the students' responses and make judgments as to the quality of the students' written products. Without set criteria to guide the rating process, two independent raters may not assign the same score to a given response. Each rater has his or her own evaluation criteria. Scoring rubrics respond to this concern by formalizing the criteria at each score level. The descriptions of the score levels are used to guide the evaluation process. Although scoring rubrics do not completely eliminate variations between raters, a well-designed scoring rubric can reduce the occurrence of these discrepancies.

*Intrarater Reliability*

Factors that are external to the purpose of the assessment can impact the manner in which a given rater scores student responses. For example, a rater may become fatigued with the scoring process and devote less attention to the analysis over time. Certain responses may receive different scores than they would have had they been scored earlier in the evaluation. A rater's mood on the given day or knowing who a respondent is may also impact the scoring process. A correct response from a failing student may be more critically analyzed than an identical response from a student who is known to perform well. Intrarater reliability refers to each of these situations in which the scoring process of a given rater changes over time. The inconsistencies in the scoring process result from influences that are internal to the rater rather than true differences in student performances. Well-designed scoring rubrics respond to the concern of intrarater reliability by establishing a description of the scoring criteria in advance. Throughout the scoring process, the rater should revisit the established criteria in order to ensure that consistency is maintained.

*Reliability Concerns in Rubric Development*

Clarifying the scoring rubric is likely to improve both interrater and intrarater reliability. A scoring rubric with well-defined score categories should assist in maintaining consistent scoring regardless of who the rater is or when the rating

is completed. The following questions may be used to evaluate the clarity of a given rubric: 1) Are the scoring categories well defined? 2) Are the differences between the score categories clear? And 3) Would two independent raters arrive at the same score for a given response based on the scoring rubric? If the answer to any of these questions is "no", then the unclear score categories should be revised.

One method of further clarifying a scoring rubric is through the use of anchor papers. Anchor papers are a set of scored responses that illustrate the nuances of the scoring rubric. A given rater may refer to the anchor papers throughout the scoring process to illuminate the differences between the score levels.

After every effort has been made to clarify the scoring categories, other teachers may be asked to use the rubric and the anchor papers to evaluate a sample set of responses. Any discrepancies between the scores that are assigned by the teachers will suggest which components of the scoring rubric require further explanation. Any differences in interpretation should be discussed and appropriate adjustments to the scoring rubric should be negotiated. Although this negotiation process can be time consuming, it can also greatly enhance reliability (Yancey, 1999).

Another reliability concern is the appropriateness of the given scoring rubric to the population of responding students. A scoring rubric that consistently measures the performances of one set of students may not consistently measure the performances of a different set of students. For example, if a task is embedded within a context, one population of students may be familiar with that context and the other population may be unfamiliar with that context. The students who are unfamiliar with the given context may achieve a lower score based on their lack of knowledge of the context. If these same students had completed a different task that covered the same material that was embedded in a familiar context, their scores may have been higher. When the cause of variation in performance and the resulting scores is unrelated to the purpose of the assessment, the scores are unreliable.

Sometimes during the scoring process, teachers realize that they hold implicit criteria that are not stated in the scoring rubric. Whenever possible, the scoring rubric should be shared with the students in advance in order to allow students the opportunity to construct the response with the intention of providing convincing evidence that they have met the criteria. If the scoring rubric is shared with the students prior to the evaluation, students should not be held accountable for the unstated criteria. Identifying implicit criteria can help the teacher refine the scoring rubric for future assessments.

## Concluding Remarks

Establishing reliability is a prerequisite for establishing validity (Gay, 1987). Although a valid assessment is by necessity reliable, the contrary is not true. A reliable assessment is not necessarily valid. A scoring rubric is likely to result in invalid interpretations, for example, when the scoring criteria are focused on an element of the response that is not related to the purpose of the assessment. The score criteria may be so well stated that any given response would receive the same score regardless of who the rater is or when the response is scored.

A final word of caution is necessary concerning the development of scoring rubrics. Scoring rubrics describe general, synthesized criteria that are witnessed across individual performances and therefore, cannot possibly account for the unique characteristics of every performance (Delandshere & Petrosky, 1998; Haswell & Wyche-Smith, 1994). Teachers who depend solely upon the scoring criteria during the evaluation process may be less likely to recognize inconsistencies that emerge between the observed performances and the resultant score. For example, a reliable scoring rubric may be developed and used to evaluate the performances of pre-service teachers while those individuals are providing instruction. The existence of scoring criteria may shift the rater's focus from the *interpretation* of an individual teacher's performances to the mere *recognition* of traits that appear on the rubric (Delandshere & Petrosky, 1998). A pre-service teacher who has a unique, but effective style, may acquire an invalid, low score based on the traits of the performance.

The purpose of this article was to define the concepts of validity and reliability and to explain how these concepts are related to scoring rubric development. The reader may have noticed that the different types of scoring rubrics—analytic, holistic, task specific, and general—were not discussed here (for more on these, see Moskal, 2000). Neither validity nor reliability is dependent upon the type of rubric. Carefully designed analytic, holistic, task specific, and general scoring rubrics have the potential to produce valid and reliable results.

REFERENCES


American Educational Research Association, American Psychological Association & National Council on Measurement in Education (1999). *Standards for Educational and Psychological Testing*. Washington, DC: American Educational Research Association.

Brookhart, S. M. (1999). *The Art and Science of Classroom Assessment: The Missing Part of Pedagogy*. ASHE-ERIC Higher Education Report (Vol. 27, No.1). Washington, DC: The George Washington University, Graduate School of Education and Human Development.

Delandshere, G. & Petrosky, A. (1998) "Assessment of complex performances: Limitations of key measurement assumptions." *Educational Researcher, 27* (2), 14-25.

Gay, L.R. (1987). "Selection of measurement instruments." In *Educational Research: Competencies for Analysis and Application* (3rd ed.). New York: Macmillan.

Hanny, R. J. (2000). *Assessing the SOL in classrooms*. College of William and Mary. [Available online: http://www.wm.edu/education/SURN/solass.html].

Haswell, R., & Wyche-Smith, S. (1994) "Adventuring into writing assessment." *College Composition and Communication, 45,* 220-236.

King, R.H., Parker, T.E., Grover, T.P., Gosink, J.P. & Middleton, N.T. (1999). "A multidisciplinary engineering laboratory course." *Journal of Engineering Education, 88* (3) 311- 316.

Knecht, R., Moskal, B. & Pavelich, M. (2000). *The design report rubric: Measuring and tracking growth through success*. Proceedings of the Annual Meeting American Society for Engineering Education, St. Louis, Missouri.

Lane, S., Silver, E.A., Ankenmann, R.D., Cai, J., Finseth, C., Liu, M., Magone, M.E., Meel, D., Moskal, B., Parke, C.S., Stone, C.A., Wang, N., & Zhu, Y. (1995). *QUASAR Cognitive Assessment Instrument (QCAI).* Pittsburgh, PA: University of Pittsburgh, Learning Research and Development Center.

Leydens, J. & Thompson, D. (1997, August). *Writing rubrics design (EPICS) I*, Internal Communication, Design (EPICS) Program, Colorado School of Mines.

Moskal, B. M. (2000). "Scoring rubrics: What, when and how?" *Practical Assessment, Research & Evaluation, 7* (3) [Available Online: http://pareonline.net/getvn.asp?v=7&n=3].

Rafilson, F. (1991). "The case for validity generalization." *Practical Assessment, Research & Evaluation, 2* (13). [Available online: http://pareonline.net/getvn.asp?v=2&n=13].

Sheppard, S. & Jeninson, R. (1997). "Freshman engineering design experiences and organizational framework." *International Journal of Engineering Education*, *13* (3), 190-197.

Stiggins, R. J. (1999). "Evaluating classroom assessment training in teacher education programs." *Educational Measurement: Issues and Practice, 18* (1), 23-27.

Yancey, K.B. (1999). "Looking back as we look forward: Historicizing writing assessment." *College Composition and Communication, 50,* 483-503.

Authors

Barbara M. Moskal
Associate Director of the Center for Engineering Education
Assistant Professor of Mathematical and Computer Sciences
Colorado School of Mines
1500 Illinois St.
Golden, Colorado 80401

Jon A. Leydens
Writing Program Administrator
Division of Liberal Arts and International Studies
Colorado School of Mines
1500 Illinois St.
Golden, Colorado 80401

**Descriptors:** *Rubrics; Scoring; *Student Evaluation; *Test Construction; *Evaluation Methods; Grades; Grading; *Scoring; Reliability; Validity

**Citation:** Moskal, Barbara M. & Jon A. Leydens (2000). Scoring rubric development: validity and reliability. *Practical Assessment, Research & Evaluation*, 7(10). Available online: http://PAREonline.net/getvn.asp?v=7&n=10.