Using Rubrics to Increase the Reliability of Assessment in Health Classes

Lynette Silvestri, EdD; Jeffrey Oescher, EdD

The authors are Associate Professor in the College of Education and Human Development at the University of New Orleans. **Contact Authors:** Lynette Silvestri, University of New Orleans, Human Performance Center, Room 109, New Orleans, LA, 70148; phone: 504-280-6019; fax: 504-280-6018; email: **lsilvest@uno.edu.** Jeffrey Oescher, University of New Orleans, Education Building, Room 184, New Orleans, LA 70148; phone: 504-280-6649; fax: 504-280-6453; email: joescher@uno.edu.

Submitted June 8, 2005; Revised and Accepted March 4, 2006

Abstract

This study examined the use of rubrics in scoring a performance-based assessment. After receiving a health lesson of ways to have a healthy brain, fifth grade students were given an assignment to illustrate and write a booklet that demonstrated their knowledge of the topic. From students' responses the researchers constructed four sample papers based on their work that demonstrated varying levels of knowledge about the brain. A "true" score for each sample paper was the result of the researchers' agreement when the papers were scored independently using rubrics. Then, sixteen pre-service teachers assessed the four sample papers without using a rubric and a second time using a rubric developed by the researchers. Those 16 scores were compared to the "true" scores for each paper. Inferential statistical analyses indicated the scores produced without using rubrics significantly inflated students' scores for three of the four papers. Analyses of the scores produced when using rubrics indicated no significant differences across all four papers. Thus, the pre-service teachers' assessment for the four sample papers was closer to the "true" score when using rubrics. The results suggest the need to develop and use rubrics to ensure the reliability of assessments addressing critical thinking skills.

Key Words: *Rubrics, Assessment, Health Lesson*

Introduction

As the goals of schooling are being redefined to reflect national standards and the importance of students' abilities to write, create, and think critically, traditional assessment models are being challenged .[1-3] When outcomes are defined as critical thinking, complex performances, or products, traditional assessments such as paper-and-pencil tests do not provide reasonable information about a student's performance on such tasks. As a result, teachers are being encouraged to use alternative assessment methods such as performance assessments to collect relevant information upon which they can base their instructional or evaluative decisions. A performance assessment requires a student to perform a task in an observable way and can be assessed by several forms including oral, written and illustrated. However, all performance is based on knowledge and represents understanding of the information taught, observed and understood by the student. For example, a student may be asked to perform a laboratory experiment and write about the results. Given the performance-based nature of the outcomes important to most educational programs as well as the need for students to reason effectively to ensure their behaviors reflect healthy practices, abundant opportunities exist for such assessments.

Teachers trained in traditional approaches can be motivated, but apprehensive, about incorporating such assessments into their instruction. However, one major concern is related to the subjective nature of the assessment process and the concern about the potential lack of reliability associated with the results. Without clear, unambiguous criteria, two raters can easily evaluate an assignment differently.

Scoring rubrics address this concern by identifying specific criteria and scoring scales that "objectify" this process.[4] Performance can be assessed according to predetermined expectations and criteria that promote learning by offering clear performance targets to students.[5, 6] Rubrics also provide an important framework around which information related to what students know and think can be communicated effectively and efficiently to students, teachers, parents, and others.

The purpose of this study was to examine the effect of using a rubric on the reliability of scores with an alternative performance-based assessment related to students' knowledge of health-related issues.

Methods

The Lesson and the Assessment of Student's Critical Thinking Skills

In preparation for the study, 32 fifth grade students were taught a lesson that focused on maintaining a healthy brain in the larger context of a physical education unit covering general exercise and health issues. Students were told brains are healthy when they function properly because their physical needs are met. The lesson identified four specific needs of a healthy brain: fluid, energy, physical activity, and stress release. Students discussed these four needs and ways by which these needs could be met.

Students were given an assignment that involved their performance of a task. The task was assessed using performance-based assessment, which is defined as "a form of testing that requires students to perform a task rather than select an answer from a ready-made test".[7] Examples of performance-based tasks are creating a work of art, or performing using a musical instrument.

Students in this study were given an assignment requiring them to identify the four needs of the brain, provide solutions, examples of solutions, as well as recommended amounts of the solution examples to meet the needs. Students' responses could take the form of a brief narrative or pictures representing the needs, solutions, examples and amounts. The purpose of the assignment was two-fold. First, the students were given a performance-based task to give them an opportunity to express their knowledge in an alternate form. Second, the assignment was developed for researchers to determine what misconceptions fifth-grade students had about the brain. With this information the researchers were able to develop four papers that represented the varying levels of understanding the students expressed in their papers. An example of required contents of the student papers is shown in Table 1.

The Development of the Instrument and the Scoring Rubric

The Instrument. The responses of all 32 students were examined to identify specific misconceptions related to the health unit being taught. From this analysis, four papers that reflect the student work were developed by the researchers. Each paper contained a response related to each of the four needs. Each paper was constructed to represent common misconceptions exemplified by the students' work as well as differing levels of

correctness for each of the responses. All four papers used responses that included pictures as well as narrative discussions. Researchers only constructed four sample papers, to give examples of the way students may respond, so pre-service teachers could focus more on doing an accurate assessment and not be burdened by assessing a large number of responses.

A rubric is defined as a guide, usually presented as a chart, which identifies and describes various levels of performance on any given assignment.[8] Levels of performance on a 3-point scale have different ratings, e.g. 3= excellent, 2= good, 1= needs improvement. Rubrics were used in a study conducted by Andrade[9] using 242 eighth-grade subjects. Statistical significance (p. <.05) was found in the number of points scored by subjects from the first to the second essay written. The researcher felt that rubrics played a key role in providing helpful feedback to students.

The rubric discussed below was used to develop the "true" score for each paper. Each researcher independently assessed the 4 sample papers using the rubrics and showed identical results. Their agreement, by using rubrics, was the basis for the "true" score for the assignment. Responses ranged from a high score of 5 to a low score of 1. The first paper was the most accurate with three of the four responses earning a perfect score of 5; this paper earned a score of 19 of the 20 possible points. The second paper contained obvious flaws across all four of the responses; it earned a score of only 8 of the 20 possible points. The third and fourth papers reflected varying degrees of accuracy and earned 12 and 15 points respectively. Thus, the "true" scores for the 4 sample papers was 19 for the first, 8 for the second, 12 for the third, and 15 for the fourth paper. In summary, the process involved three steps. First, researchers established the rubrics, then created the four sample papers, and finally assessed the papers to determine the "true" score for the assignment.

The Rubric. The criteria used to score each response reflected the inclusion of the four issues related to the lesson content. That is, the criteria focused on the identification of a need, a general solution to the need, example of a solution, and the required amount of the solution. Table 2 summarizes the five-point scoring scale developed for each specific need. Given a five-point scale being used across four responses, a score of 20 represented a perfect paper.

Suppose a student responded to the need for fluids by identifying the need itself, solution of drinking fluids, and an example of drinking water with a quantity of four glasses. This response would be scored as a 5. If a student's response indicated only that a person needed to drink something to satisfy their need for fluids, the response would be scored as a 3. An example of a score of 1 would reflect a response that suggested a person should drink water but neither the need for fluids or the solution to drink beverages were mentioned. Thus, the score for a paper represents the accuracy of the illustration and written information to describe the needs, solution, example and amount required for a healthy brain.

Participants

The participants were convenient sample of 16 pre-service teachers enrolled in an undergraduate education methods class. They were familiarized with the lesson, its objectives, and the instruction provided to the fifth grade students.

Procedures

Copies of the four sample papers developed by the researchers were distributed to the participants two times. On the first occasion the only directions the participants received were those related to the total points for each response and the means by which they were to calculate a total score. They were specifically told each response was to be rated on a scale of 1 to 5 with the latter score representing the highest possible score. They were also asked to sum the scores across all four responses to create a total score that could range from 4 to 20. All four papers and scoring sheets were collected when the participants completed their work.

After a short break, the participants were shown the rubric. The criteria were explained as well as the descriptions of each point on the scoring scale. Participants were then given copies of the four sample papers and were asked to re-score each response using the rubric to guide their efforts. They were instructed to create the total score that reflected the sum of the scores for all four responses. In addition, each participant was asked to reflect on the following questions and provide a concise written response to each one. The first question stated, "In your opinion, what was the difference between scoring papers the first time without rubrics and the second time with rubrics?" The second question asked, "Which of the two ways would you prefer to use to grade student papers and why do you feel that way?" All papers were collected as the participants completed their work.

Results

Means and standard deviations for the scores for each paper when scored without rubrics and with rubrics are presented in Table 3. In addition, the

"true" scores based on the assessment of each paper by the researchers are presented.

Scoring Without Rubrics

An examination of the data indicates the scoring performed without rubrics tended to inflate the scores associated with all of the papers. In the case of Paper 1, there was very little difference from the true score. This is likely due to the fact that very few mistakes were present in this paper. The scores for Papers 2-4 tended to be quite different from their respective true scores. In the most extreme case of Paper 3, the difference was almost 8 points. Differences for Papers 2 and 4 were about four-and-two-thirds and three-and-one-half points respectively. An inferential analysis of the comparison of the scores for each paper to the respective true score using a one-sample t-test indicated a non-significance difference for Paper 1 ($t_{15} = 1.43$, p = .173) and significant differences for Papers 2-4 ($t_{15} = 11.09$, $p = .000$; $t_{15} = 12.00$, $p = .000$; and $t_{15} = 4.34$, $p = .001$). An examination of the mean scores indicates students earned approximately 97%, 83%, 80%, and 88% of the possible points for papers 1-4 respectively. This is a range of about 14%. On a ten point grading scale (e.g., 60-69, 70-79, 80-89, 90-100) all papers reflected at least "B" level work with one reflecting a high "A" level. All papers were scored higher than the true scores. The smallest difference was that for paper 1 (e.g., 2%), while the largest was for paper 3 (e.g., 40%). Papers 2 and 4 were inflated by 23% and 14 % respectively.

Scoring With Rubrics

An examination of the data indicates the scoring performed with rubrics tended to be relatively close to the true scores for all papers. Scores for two of the four papers were lower than the true score and two were higher, although these differences were typically less than one-third of a point. An inferential analysis of the comparison of the scores performed with rubrics for each paper to their respective true scores indicated no significant differences for papers 1-4 ($t_{15} = -0.89$, $p = .386$; $t_{15} = -0.47$, $p = .643$; $t_{15} = 1.98$, $p = .067$; and $t_{15} = 1.15$, $p = .270$).

An examination of the mean scores indicates students earned approximately 94%, 59%, 48% and 74% of the total possible points for papers 1-4 respectively. The range of percentages was 46%, more than three times that of the range of the papers when scored without rubrics. On a ten point scale, grades would range from A-F. The scores for papers1 and 2 were lower than the true score; those for papers 3 and 4 were higher. Differences in the percentages between the scores and the true scores were highest for paper 3 (8%) and lowest for paper 2 (1%). Scores for papers 1 and 4 differed from the true score by 2% and 4% respectively. These differences reflect substantial decreases from those found when all papers were scored without rubrics.

Discussion

Results of this study confirm the need to address issues of score reliability, particularly when using alternative assessments that require significant levels of subjectivity during the scoring process. It is evident that when pre-service teachers used the rubric, scores tended to be very close to what could be considered the "true" score. Researchers determined the "true" score by independently using rubrics to assess the 4 sample papers and reaching agreement on the score for each of the papers. When these same participants scored the assignments using only a very rough scoring guide (e.g., 20 total points), scores tended to differ greatly from the "true" scores. In all cases these scores were inflated; in some cases they were grossly inflated.

Beyond the fact that more reliable scores resulted from using the rubric, two interesting issues surfaced through the responses to the two semi-structured questions asked of each subject upon completion of the scoring procedures. First, a majority of the participants reported the rubrics facilitated the assessment process by providing specific guidelines to follow when grading. This is exactly the purpose of a rubric. According to Goodrich:

> "rubrics reduce the time teachers spend grading work and make it easier for teachers to explain to students why they received the grade they did and what they can do to improve".[10]

Second, when asked whether they preferred using rubrics or not, the majority of participants indicated the use of a rubric resulted in consistent grading that was far less subjective in nature. Only two participants preferred assessing without using rubrics, and their concerns focused on the fact that although students did not state their answers correctly, the assessors wanted to give the students credit for having the right idea. While this perspective is potentially admirable, it does not reflect sound assessment practice.

While the results of this study demonstrate the enhancement of reliability when using rubrics, it is necessary to note the difficulty often encountered when developing such rubrics. Both of the researchers in this study discussed at great length the criteria around which the rubric was developed, and considerable time was spent developing the actual

scoring scale for this criteria. In addition, it was necessary to "field test" the rubric to ensure its clarity, comprehensiveness, and communication. All of these are time consuming and somewhat difficult. This extra effort, however, is more than compensated by the resulting clarity of the assessment target and scoring process for both teachers and students.

References

1. Stiggins R. *Student Involved Classroom Assessment*. Upper Saddle River, NJ: Prentice Hall; 2000.

2. Airasian PW. *Classroom assessment: concepts and applications*. New York, NY: McGraw Hill; 2000.

3. McMillan JH. *Classroom Assessment: Principles and Practice for Effective Instruction*. Boston, Mass: Allyn and Bacon; 1997.

4. Boston C. *Understanding scoring rubrics: a guide for teachers.* College Park, MD: Clearing House on Assessment and Evaluation; 2002.

5. Taggart GL, Phifer SJ, Nixon JA, & Wood M. *Rubrics.* Lancaster, Pa: Technomic; 1998.

6. Marzano RJ, Pickering DJ, & McTighe J. *Assessing student outcomes: performance assessment using the dimensions of learning model.* Aurora, Colo: McRel Institute;1993.

7. Sweet D. Performance Assessment. *Education Research Consumer Guide.* [serial online].1993. Available at http://www.ed.gov/pubs/OR/ConsumerGuides/perfasse.html. Assessed 22 February, 2006.

8. Teaching Today. *Rubrics Defined* [serial online]. 2000.Available at http://www.glencoe.com/sec/teachingtoday/weeklytips.phtml/23. Assessed 21 February, 2006.

9. Goodrich Andrade H. The effects of instructional rubrics on learning to write. *Curr Is Ed* [serial online]. 2001. 4(4). Available at http://cie.ed.asu.edu/volume4/number4/.html. Assessed 12 January, 2006.

10. Goodrich H. *Just what is a rubric?* Reforming Middle Schools and School Systems 1997. 1(2). Available at http://middleweb/CSLB2rubric.html. Assessed 15 December 2005.

Table 1. *Example of Required Contents of Student Papers*

NEEDS	SOLUTION	EXAMPLE	AMOUNT
Fluid	Beverage	Water	4 glasses/day
Energy	Carbohydrates	Potatoes	6 servings /day
Physical activity	Exercise	Jogging	3 times/week
Stress release	Exercise	Jump rope	As needed to reduce stress

Table 2. *Scoring Scale for each Response*

SCORE	RESPONSE DESCRIPTION
5	The need, solution, example, and the required amount of the specific amount are identified.
4	The need, solution and either the example or required amount of the example are identified correctly.
3	The need and the solution are identified correctly. The example is inaccurate and the amount of the example is incorrect.
2	Both the solution and the required amount of the example are identified. The example is inaccurate and the amount of the example is incorrect.
1	Only the example or the required amount of the example is identified correctly. The need and the solution are not identified or they are incorrect.

Table 3. *Summary of the Scores for Papers 1-4*

Paper		Without Rubric		With Rubric		True Score
	N	Mean	SD	Mean	SD	
1	16	19.31	0.87	18.69	1.40	19.00
2	16	16.63	1.67	11.74	2.11	12.00
3	16	15.94	2.64	9.56	3.16	8.00
4	16	17.50	3.22	14.75	2.62	14.00