AC 2008-845: MEASURING STUDENT ABILITY TO WORK ON MULTI-DISCIPLINARY TEAMS: A METHOD FOR DETERMINING VALIDITY AND RELIABILITY OF A RUBRIC

Carolyn Plumb, Montana State University

Durward Sobek, Montana State University

Measuring Student Ability to Work on Multi-Disciplinary Teams: A Method for Determining Validity and Reliability of a Rubric

Engineering educators struggle to provide effective educational experiences for professional skills such as communication, cultural awareness, and ability to work on multi-disciplinary teams. As difficult as these skills are to teach, they are even more difficult to evaluate. Over the past year, we have introduced a new course at the junior-level, "Introduction to Engineering Design." The course focuses on the skills necessary to complete a project in a multi-disciplinary team, and it will eventually be required for all engineering students as a precursor to their department-specific capstone design courses. In a previous paper, we described our approach of using the engineering design process to determine the best solution to the problem of providing students with a multi-disciplinary educational experience in engineering at Montana State University.[1]

In order to determine if our new course improves student performance in this area, we developed a rubric for evaluating an individual's performance on a multi-disciplinary team. In previous work, we presented our rubric development process and preliminary results.[2] In this paper, we discuss an initial assessment of the validity and reliability of the rubric.

Broadly speaking, the validity of a measurement instrument refers to how well it measures what it is intended to measure, whereas reliability is the consistency of the results of using the instrument.

Validity

Creating a valid rubric begins early in the process. Below are brief descriptions of four different types of validity that should guide rubric development:[3]

> **Content Validity** refers to whether the rubric taps the knowledge and skills of the larger domain. The test for the rubric is whether a content expert would agree with the criteria chosen in the rubric.
>
> **Construct Validity** means that the rubric measures what it is supposed to measure. Do the observable and measurable criteria accurately address the knowledge and skills to be evaluated? For example, if the rubric is to assess a problem-solving outcome, does it have problem solving criteria and rating scales?
>
> **Criterion Validity** relates to the predictability of measurements in a real-world context. If a student receives a high score using the rubric as the measurement tool, can we reliably predict that the student will perform well on tasks requiring that skill or knowledge?

Face Validity does not refer to whether the rubric is valid (and measures what it is supposed to measure) in the true sense, but refers to whether the rubric *appears* to be valid to its users. Students have a stake in the quality of a rubric that measures their performance, so it's important that the rubric seems right to them. If content, construct, and criterion validity are strong, face validity is also likely to be strong.

In order to address **content validity**, we involved as many content experts as practically reasonable from the beginning of the rubric development process. The process involved developing an initial description of multi-disciplinary teamwork. This description was developed collaboratively with our Multi-Disciplinary Advisory Committee, which included a dozen faculty and professional staff from the college's five engineering degree programs and one research center. This advisory committee also helped us develop and hone a set of "key attributes" for effective performance on a multi-disciplinary team. The final set of key attributes selected is listed below:

Interpersonal Communication
Collaboration
Understanding and Communicating Disciplinary Tradeoffs
Empathy for Diverse Perspectives
Planning and Organization
Accountability and Reliability
Common Goals and Shared Outcomes
Conflict Management and Resolution
Willingness to Learn
Inclusive Decision Making

These ten attributes were then combined into five rubric areas of performance, as described in a previous paper.[4] The completed rubric includes levels of performance and accompanying descriptions on these key attributes. For example, the highest level of performance on Interpersonal Communication and Collaboration is described as "Stimulates team unity by advancing ideas of others, willingly filling in gaps of team performance, and by proactively and clearly communicating to facilitate progress toward team goals." These descriptions were vetted by the advisory committee.

In addition, several faculty who served as senior capstone design project advisors tested the rubric by using it to evaluate student team performance. A total of 47 rubrics (47 student team members) were completed prior to using the rubric in the new junior-level course. A survey of the faculty who used the rubric produced no major concerns about the instrument.

In regard to **construct validity**, early in the process of determining the best way to provide a multi-disciplinary educational experience for our students, we developed a set of multi-disciplinary objectives, which are listed below:

- View engineering projects from a systems perspective.

- Recognize and appreciate trade-offs across disciplinary perspectives.
- Communicate technical and other trade-offs, and negotiate satisfactory resolution.
- Generate creative, integrated, and effective solutions collaboratively.

When we developed our list of "key attributes" for the rubric, we matched those attributes back to these objectives in order to ensure that the rubric was addressing the skills and knowledge that we wanted to measure.[5]

Criterion validity is more difficult to determine until we can get feedback from alumni after they have taken the course and used the rubric for both self and peer assessment. However, considering the fact that several of the members of our Multi-Disciplinary Advisory Committee have industry experience, we are fairly confident that the rubric has the ability to measure skills students will need to exhibit in future performances and is satisfactorily predictive.

Face validity is important if we are to have valid measures of student performance. We have asked students for feedback about the rubric and its usefulness in the piloted course, both for formative and summative assessment purposes. Two student concerns have been raised to date.

1. At least one student voiced concern about using the rubric for evaluation purposes, particularly if the scores are peer assessments. We have not yet used the peer rubric scores when considering final grades (although we have used the project advisor rubric scores). When the rubric was designed, we thought of it being most useful in two ways: (1) as a formative assessment and instructional tool, clearly spelling out the characteristics and attributes that were important to effective performance on a multi-disciplinary team and (2) as a college-level assessment tool to determine if our students are able to "function on a multi-disciplinary team" (for ABET). We do not have plans to use peer or self assessments via the rubric for final grades.
2. Another student commented that the scale of the rubric seemed unduly weighted to the positive end. The rubric includes 5 levels of performance for multiple criteria, with the typical (and acceptable) level of performance toward the middle. However, four of the five levels describe a positive team contribution. On one hand, we do not consider this comment to be a serious threat to the face validity of the rubric because the rubric is not meant to represent a normal, bell-shaped distribution of behaviors and an interval scale of 1 to 5 with 3 being "average." However, we are still considering eliminating the highest performance level on the rubric and inserting a level between the two lowest levels.

In summary, a large part of our method for ensuring rubric validity was built into the process of developing the rubric. We are still considering changes to the rubric to improve face validity, but we would like to get feedback from a larger number of students first, and we can get that feedback from students spring semester 2008, when nearly 90 students will take the course.

Reliability

We consider it important to get a measure of the inter-rater reliability of our rubric. Inter-rater reliability is "used to assess the degree to which different raters/observers give consistent estimates of the same phenomenon."[6] Inconsistent ratings from different raters affect the instrument's ability to produce a valid measurement.

We used data from spring and fall semesters, 2007, for our initial reliability analysis, which is described in detail below. Although the courses had low enrollment, we had usable data from a total of 6 teams, 4 teams with 4 students and 2 teams with 3 students. Each of these teams completed mid-term and final rubrics, assessing both themselves and their teammates.

Ideally, comparing at least two instructor sets of scores would be the best data for determining inter-rater reliability. Instructors are generally dedicated to providing fair and accurate measures and are more accustomed to rating or evaluating student performance. However, in this context, we did not—and probably will not in the future—have more than one instructor-level score. It simply is not feasible to use two instructors to advise the same set of teams of 3 or 4 students given the level of interaction or observation required to accurately assess teamwork skills. Thus, we had to settle for peer scores, knowing that they were likely to have less inter-rater reliability than would the scores of two experienced instructors.

Typically, rater training can increase inter-rater reliability; thus, each time the students used the rubric, we discussed the skill categories and levels of performance, and encouraged the students to provide thoughtful scores. We told students not to think of the 1 through 5 scores as a Likert-type scale, with "average" performance automatically deserving a "3" score; rather, we encouraged them to actually read the descriptions connected with each level of performance for the five areas or attributes and match their performance assessment scores of themselves and each other to the closest description.

We initially planned to use both mid-term and final rubrics to measure inter-rater reliability. However, an initial analysis of the rubric data found that mid-term rubric ratings showed somewhat more inter-rater reliability than final rubric ratings. Because there were differences in the inter-rater reliability between mid-term and final ratings, we did not combine them. Rather, we erred on the conservative side, **using only the final rubric scores for the reliability analysis**, thinking that students would have more information upon which to base their scores at the end of the semester, which would lend to more valid scores.

One way to determine or describe the inter-rater reliability of the rubric is to compare peer rubric scores for another peer on the rubric metrics or attributes. Ideally, if three team members rated the fourth on a particular skill, the ratings would be identical. Thus, a smaller **range** of ratings of an individual's performance along a given metric indicates higher inter-rater reliability for that metric. The authors could find no absolute level of rater consistency considered acceptable in the literature.

Table 1 reports the aggregate peer score ranges for each team on the following five rubric "attributes":

- Interpersonal Communication and Collaboration
- Understanding Disciplinary Tradeoffs & Empathy for Diverse Perspectives
- Planning/Organization and Accountability/Reliability
- Common Goals/Shared Outcomes & Conflict Management and Resolution
- Willingness to Learn and Inclusive Decision Making

Table 1: Comparison of the percent of ratings with various ranges, by team (includes only peer comparisons—not self or instructor). Each row is a team (4 teams from spring 2007 and 2 from fall 2007).

Team	% ranges = 3	% ranges = 2	% ranges = 1	% ranges = 0	% ranges ≤ 1
A: Spr 07	5	20	60	15	**75**
B: Spr 07	0	10	65	25	**90**
C: Spr 07	0	20	40	40	**80**
D: Spr 07	0	0	60	40	**100**
E: Fall 07	5	25	55	15	**70**
F: Fall 07	0	25	60	15	**75**

To illustrate how the ranges used for the percents in Table 1 were computed, the ratings for Team E (four members) on "Interpersonal Communication and Collaboration" were as follows (1 is the highest level, and 5 is the lowest level):

Ratings of Team Member a: 2, 1, 1 (range = 1)
Ratings of Team Member b: 3, 4, 5 (range = 2)
Ratings of Team Member c: 1, 1, 1 (range = 0)
Ratings of Team Member d: 2, 2, 2 (range = 0)

This process was repeated for the other four metrics, which yielded a total of 20 ranges for this team (4 range values for each of 5 metrics). A three-person group would have a total of 15 ranges (3 range values for each of 5 metrics). Thus, the percents in Table one are based on 20 or 15 ranges, depending on the number of students on the team.

The analysis indicates that overall inter-reliability of the rubric is fairly good. Some 80% of ratings agreed within one performance level. However, the analysis also indicates a good deal of variation among the teams in the level of consistency of the peer scores: Team D, for example, had higher inter-rater reliability than Team E. This variation could stem from at least three sources:

- The level of commitment and/or expertise present in the team members in regard to using the scoring rubric. Variation due to this source can be somewhat minimized by training in using the rubric as well as "selling" the value of the rubric scores to the students, not as individual evaluations but as a formative

assessment measure for the college objectives related to multi-disciplinary teamwork. We have used training, but perhaps this training could be improved.

- The level of functionality of the team. Variation due to this source is difficult to eliminate. One peer's assessment of another's "accountability," for example, may differ from another peer's assessment simply because they had two different experiences with the peer being assessed. For this reason, the peer rubric scores will always likely have less inter-rater reliability than would two instructor scores.
- The validity of the rubric. Variation in the data could be due to poorly defined metrics or performance levels that lead to differences in interpretation.

To gain some insight into this last item, we analyzed the ranges of the rubric scores by metric rather than by team. Table 2 below shows the ranges for each rubric attribute.

Table 2. Percent of Ranges at Each Level (Across Teams) for Each Rubric Attribute from Spring 2007 and Fall 2007 Teams.

	No. of Ranges = 0	No. of Ranges = 1	No. of Ranges = 2	No. of Ranges = 3	No. of Ranges = 4	% Ranges ≤ 1
Interpersonal Communication & Collaboration	7	11	4	0	0	81.8
Understanding Disciplinary Tradeoffs & Empathy for Diverse Perspectives	5	13	4	0	0	81.8
Planning/Organization & Accountability/Reliability	7	13	2	0	0	90.9
Common Goals/Shared Outcomes & Conflict Management and Resolution	5	10	6	1	0	68.2
Willingness to Learn & Inclusive Decision Making	2	16	3	1	0	81.8

As the table shows, the score ranges for four of the rubric attributes are acceptably and consistently high, but the ranges for the "Common Goals/Conflict Management" attribute is lower than the others.

Conclusion and Future Activities

We have reported our efforts to ensure the validity and determine the inter-rater reliability of a rubric developed to measure the performance of an individual on a multi-disciplinary team. The bulk of our efforts in regard to validity were integrated into our

collaborative process to develop the rubric. We do have some concerns about "face validity," that is, whether the rubric appears to be valid to the users. We will be collecting a much larger sample of data from spring semester 2008 and hope to get more information from students.

We used peer assessments from the rubric to obtain an initial measure of inter-rater reliability. Rubric scores from more than one instructor would likely provide a better measure, but it is not feasible for more than one instructor to advise a group of 3 or 4 students.

Our method involved obtaining the ranges of peer scores for another peer on each of the 5 teamwork attributes described on the rubric. Larger ranges indicate less consistency in the scores, and thus worse inter-rater reliability.

From this initial analysis, the inter-rater reliability looks satisfactory. There is some variability in the reliability across teams, and we are not sure what that means yet. The data also show some variability across the rubric "attributes," which may mean that one of the attributes needs some revision. The larger data sample that we will have available from the spring 2008 semester will help us confirm and gain insight into these findings.

Rubrics are becoming more common for assessment and evaluation purposes in engineering education; however, it is important that faculty understand how to design valid and reliable rubrics, and how to collect data that informs us about the validity and reliability of a rubric. We hope that the method presented in this paper will be useful to others in this process.

References

[1] Sobek, D., and Plumb, C., Using the Engineering Design Process to Re-Envision Multi-Disciplinary Educational Experiences for Engineering Students, *Proceedings of the 2007 American Society for Engineering Education Annual Conference and Exposition.* June 2007.

[2] Plumb, C., and Sobek, D., Measuring Student Ability to Work on Multi-Disciplinary Teams: Building and Testing a Rubric, *Proceedings of the 2007 American Society for Engineering Education Annual Conference and Exposition.* June 2007.

[3] University of Colorado Center for Development, Creating a Rubric: An Online Tutorial for Faculty, retrieved 1/07/08 from http://citt.cudenver.edu/elearning/Assessment_Rubrics/course/4_quality/4_validity.htm

[4] Plumb, C., and Sobek, D., Measuring Student Ability to Work on Multi-Disciplinary Teams: Building and Testing a Rubric, *Proceedings of the 2007 American Society for Engineering Education Annual Conference and Exposition.* June 2007.

[5] Moskal, B. M. and Leydens, J.A., Scoring Rubric Development: Validity and Reliability, *Practical Assessment, Research, & Evaluation 7(10)*, retrieved 9/19/06 from http://PAREonline.net/getvn.asp?v=7&n=10.

[6] Web Center for Social Research Methods, retrieved 1/12/08 from http://www.socialresearchmethods.net.