Rubric Development and Inter-Rater Reliability Issues in Assessing Learning Outcomes

[1]James A. Newell, [1]Kevin D. Dahm, and [2]Heidi L. Newell
[1]Department of Chemical Engineering/[2]College of Engineering
Rowan University, Glassboro, NJ 08028

Abstract
This paper describes the development of rubrics that help evaluate student performance and relate that performance directly to the educational objectives of the program. Issues in accounting for different constituencies, selecting items for evaluation, and minimizing time required for data analysis are discussed. Aspects of testing the rubrics for consistency between different faculty raters are presented, as well as a specific example of how inconsistencies were addressed. Finally, a consideration of the difference between course and programmatic assessment and the applicability of rubric development to each type is discussed.

Introduction

With the increased emphasis placed by ABET (1) on assessing learning outcomes, many faculty struggle to develop meaningful assessment instruments. In developing these instruments, the faculty members in the Chemical Engineering Department at Rowan University wanted to ensure that each instrument addressed the three fundamental program tasks as specified by Diamond (2):

- The basic competencies for all students must be stated in terms that are measurable and demonstrable
- A comprehensive plan must be developed to ensure that basic competencies are learned and reinforced throughout the time the students are enrolled in the institution
- Each discipline must specify learning outcomes congruent with the required competencies

Like many institutions (3), the Rowan University Chemical Engineering Department chose to use items that address multiple constituencies including alumni, industry, and the students themselves. Assessment data from these groups were obtained through alumni surveys, student peer-reviews, and employer surveys. These instruments were fairly straightforward to design and could be mapped directly to ABET A-K as well as the AIChE requirements and other department specific goals.

The difficulty arose when the discussion turned to student portfolios. As Rogers (4) observes, there is no one correct way to design a portfolio process. Essentially everyone agreed that a portfolio should contain representative samples of student work gathered primarily from courses taken in the junior and senior years. The ABET educational objectives are summative rather than formative in nature, so the faculty decided to focus on work generated near the end of the student's undergraduate career. A variety of assignments would be required to ensure that all of the diverse criteria required by

Criteria 2000 could be addressed by at least parts of the portfolio. At the same time, the faculty were acutely aware that these portfolios would be evaluated every year and were understandably interested in minimizing the total amount of work collected. Ultimately, the following items were selected for inclusion:

- A report from a year-long, industrially sponsored research project through the Junior/Senior Clinics
- The Senior Plant Design final report
- A hazardous operations (haz-op) report
- One final examination from a junior level chemical engineering class (Reaction Engineering or Heat Transfer)
- One laboratory report from the senior level Unit Operations Laboratory Course)

These items were all "constructed-response formats" (5-7) in which a student furnished an "authentic" response to a given assignment or test question. These were selected over multiple choice "selected response" formats because they better represented realistic behavior (8).

Although the items contained in the portfolio provided a wide range of samples of student work, they could not be as neatly mapped to the ABET criteria. There was simply no way to look at a laboratory report and assign a number evaluating the students ability to apply math, science and engineering. The immediate question that arose from the faculty was "Compared to whom?" A numerical ranking comparing Rowan University Chemical Engineering students to undergraduates from other schools may be very different than one comparing students to previous classes. It became clear that specific descriptions of the level of performance in each area would be required so that all faculty could understand the difference between a "4" and a "2." As Banta (9) stated, "The challenge for assessment specialists, faculty, and administrators is not collecting data but connecting them." The challenge became developing rubrics that would help map student classroom assignments to the educational objectives of the program.

Course vs Programmatic Assessment

Other chemical engineering departments are developing rubrics for other purposes as well. In their exceptional (and Martin Award-Winning) paper on developing rubrics for scoring reports in a unit operations lab, Young *et al*. (10) discuss the development of a criterion-based grading system to clarify expectations to students and to reduce inter-rater variability in grading. This effort represents a significant step forward in course assessment. However, the goals of course assessment and program assessment are quite different.

For graded assignments to capture the programmatic objectives a daunting set of conditions would have to be met. Specifically,

- every faculty member must set proper course objectives that spring exclusively from the program's educational objectives and fully encompass all of these objectives
- his or her tests and other graded assignments must completely capture these objectives

- student performance on exams or assignments must be a direct reflection of their abilities and not be influenced by test anxiety, poor test taking skills, etc.

If all of these conditions are met every time, then there should be a direct correlation between student performance in courses and the overall learning of the students. Moreover, much of the pedagogical research warns of the numerous pitfalls associated with using evaluative instruments (grades on exams, papers etc.) within courses as the primary basis for program assessment (11).

One of the immediate difficulties is that many criteria are blended into the "grade". A student with terrific math skills could handle the partial differential equations of transport phenomena but might never understand how to apply the model to practical physical situations. Another student might understand the physical situation perfectly, but struggles with the math. In each case, the student might wind up with a C on an exam, but for very different reasons. This is not a problem from the perspective of evaluation. Both students deserve a "C." However, from an assessment standpoint, the grade does not provide enough data to indicate areas for programmatic improvement.

Moreover, if exams or course grades are used as the primary assessment tool, the impact of the entire learning experience on the student is ignored entirely (12). Community activities, field trips, service projects, speakers, and campus activities all help shape the diverse, well-rounded professional with leadership skills that industry seeks. The influence of these non-classroom factors cannot be measured by course grades alone.

The goal of our rubrics was to map student work directly to the individual learning outcomes. This also put us in a position to more directly compare our assessment of student work with the assessment of performance provided by student peer reviews, employers and alumni.

Rubric Development

The first step was to take each educational objective and develop indicators, which are measurable examples of an outcome through phrases that could be answered with "yes" or "no." A specific educational objective and indicator is shown below.

> *Goal 1, Objective 1: The Chemical Engineering Program at Rowan University will produce graduates who demonstrate an ability to apply knowledge of mathematics, science, and engineering (ABET - A).*
>
> *Indicators:* *1. Formulates appropriate solution strategies*
>
> *2. Identifies relevant principles, equations, and data*
>
> *3. Systematically executes the solution strategy*
>
> *4. Applies engineering judgment to evaluate answers*

Once the indicators for each objective were developed, the next task involved defining the levels of student achievement. Clearly, the lowest level should be that a novice would demonstrate when confronted with a problem. The highest level should show metacognition (12), which is students' awareness of their own learning skills,

performance, and habits. To achieve the highest level, a student would not only have to approach the problem correctly, he or she would also need to demonstrate an understanding of his or her problem solving strategies and limitations. The intermediate scores would represent steps between a metacognitive expert and a novice. It is important to note that the numbers are ordinal rather than cardinal. A score of four does not imply "twice as good" as a score of two.

All of the other assessment instruments used by the Chemical Engineering Department at Rowan University used a 5-point Likert scale, so a faculty team set out to develop meaningful scoring rubrics using a 5-point scoring system. Initially, the scores contained labels (5 = excellent, 4 = very good, 3 = good, 2 = marginal, 1 = poor), but the qualitative nature of the descriptive labels led to confusion in scoring. Some professors have different distinctions between excellent and very good and tended to use these distinctions more than the descriptive phrases that define the difference between levels for each indicator. More importantly, if the rubrics are well designed, the descriptive phrases should stand alone, without the need for additional clarifiers. Ultimately, it was decided to eliminate all labels.

It became readily apparent that a four-point scale allowed for more meaningful distinctions in developing the scoring rubrics for the portfolios. Providing four options instead of five, eliminates the default "neutral" answer and forces the evaluator to choose a more positive or negative ranking. The four-option scale also made it easier to write descriptive phrases that were meaningfully different from the levels above and below. In developing these phrases, the following heuristic was used. For the four-point phrases, the writer attempted to describe what a meta-cognitive expert would demonstrate. For the three-point phrases, the target was what a skilled problem solver who lacked meta-cognition would display. For the two point words, the writers attempted to characterize a student with some skills, but who failed to display the level of performance required for an engineering graduate. Finally, the one-point value captured the performance of a novice problem solver.

To evaluate a given indicator, the professor would read the left-most description. If it did not accurately describe the performance of the student, they would continue to the next block to the right until the work was properly described. A sample rubric is shown below.

	4	3	2	1
Formulates appropriate solution strategies	Can easily convert word problems to equations. Sees what must be done	Forms workable strategies, but may not be optimal. Occasional reliance on brute force	Has difficulty in planning an approach. Tends to leave some problems unsolved	Has difficulty getting beyond the given unless directly instructed

Identifies relevant principles, equations, and data	Consistently uses relevant items with little or no extraneous efforts	Ultimately identifies relevant items but may start with extraneous info	Identifies some principles but seems to have difficulty in distinguishing what is needed.	Cannot identify and assemble relevant information
Systematic-ally executes the solution strategy	Consistently implements strategy. Gets correct answers	Implements well. Occasional minor errors may occur	Has some difficulty in solving the problem when data are assembled. Frequent errors.	Often is unable to solve a problem, even when all data are given
Applies engineering judgment to evaluate answers	Has no unrecognized implausible answers	Has no more than one if any unrecognized implausible answers. If any it is minor and obscure	Attempts to evaluate answers but has difficulty. Recognizes that numbers have meaning but cannot fully relate.	Makes little if any effort to interpret results. Numbers appear to have little meaning

Rubric Testing and Inter-Rater Reliability

Once the lengthy process of developing scoring rubrics for each objective was completed, the rubrics needed testing. C. Robert Pace (13) stated the challenge of accurate assessment succinctly, saying "The difficulty in using faculty for the assessment of student outcomes lies in the fact that different professors have different criteria for judging students' performance." The intent of the rubrics was to create specific and uniform assessment criteria, such that the role of subjective opinions would be minimized. The ideal result would be that all faculty members using the rubrics would assign the same scores to a given piece of student work every time.

To evaluate whether the rubrics were successful in this respect, six samples of student work (four exams and two engineering clinic reports) were copied and distributed to the entire chemical engineering department, which consisted of seven faculty at that time. All faculty members assigned a score of 1, 2, 3, 4 or "not applicable" to every student assignment for every indicator. This produced some 160 distinct score sets (not including those that were all "not applicable") that were examined for inter-rater reliablity.

The results, in general, were excellent. Every faculty member in the department scored the items with one level of each other in 93% of the items. In 47% of the score sets (75 of 160) agreement was perfect- all faculty members assigned exactly the same score. In another 46% of the score sets, all assigned scores were within 1. Rubrics for which this level of agreement was not routinely achieved were examined more closely for possible modification. After all of the scoring sheets had been compared, the faculty met to discuss discrepancies in their evaluations. In essence, this applied the continuous

improvement process to the instruments used to facilitate the continuous improvement process.

The primary example of a rubric that required modification is shown below. "Solutions based on chemical engineering principles are reasonable," in the originally developed scheme, was an indicator that applied to a number of different educational objectives. This was the only rubric for which scores were not routinely consistent. One heat transfer exam received a range of scores that included multiple occurrences of both "4" and "1."

	4	3	2	1
Solutions based upon chemical engineering principles are reasonable	Has no unrecognized implausible answers	Has no more than one if any unrecognized implausible answer. If any it is minor and obscure.	Attempts to evaluate answers but has difficulty. Recognizes that numbers have meaning but cannot fully relate.	Makes little if any effort to interpret results. Numbers appear to have little meaning.

In the ensuing discussion, it was found that the difficulty with this exam was that nothing recognizable as a final answer was presented for any question. The student formulated a solution strategy and progressed through some work but never finished solving the equations. Some faculty interpreted the rubric wording and chose to give an assignment of "4". This interpretation is understandable. Because no answer was given, there was certainly no "unrecognized implausible answer." By the letter of the criteria, the student earned a four. However, some faculty interpreted the criteria differently, resulting in the assignment of "1". This interpretation is also reasonable- since there were no results, there was certainly no attempt to interpret the results. The rubric was simply re-written to specify that a rating of N/A be given if no recognizable "final answer" was provided and the discrepancies in scoring were not present in subsequent evaluations.

In addition to pointing out necessary revisions, this testing provided a good measure of inter-rater reliability. Having every faculty member review every item in an annual assessment portfolio would be a laborious task. Consequently, the results of this test were examined to determine what level of accuracy could be expected when a group of 3 faculty reviewed an item. For example, in the following score set:

2 2 2 2 1 3 2

The mean score assigned by the faculty was 2, and the mean of a 3 score subset could be 1.67, 2 or 2.33. This means that *any* panel of three faculty members would have assessed this sample of work with a score *within 0.5* of that assigned by the entire faculty. It was found (after one rubric was revised as described above) that 95% (153 of 160) of the score sets showed this level of consistency. Thus, it was concluded that when using the rubrics, a randomly constituted panel of three faculty members would be reasonably representative of the department. Detailed rubrics are available through the web at http://engineering.eng.rowan.edu/~newell/rubrics.

The Next Level

The next goal is to use the rubrics to help guide the selection of course objectives across the curriculum. With detailed educational objectives in place and rubrics to assist in their assessment, the hope is that improved course objectives will be developed that more directly link classroom activities and evaluations with the goals set by the program.

Conclusions

A complete set of rubrics was developed and tested that maps student performance on a variety of junior/senior levels assignments directly to program educational objectives. These rubrics were tested for inter-rater reliability and were shown to yield the same mean (within 0.5), regardless of which set of three faculty evaluate the material. These results, in conjunction with input from alumni, employers, and the students themselves serve as a basis for assessment of the chemical engineering program.

Bibliographic Information

1. ABET Engineering Criteria 2000.
2. R. M. Diamond, *Designing and Assessing Courses and Curricula: A Practical Guide*," Jossey-Bass Inc., San Francisco (1998).
3. J. A. Newell, H. L. Newell, T. C. Owens, J. Erjavec, R. Hasan, and S. P. K. Sternberg, "Issues in Developing and Implementing an Assessment Plan in Chemical Engineering Departments," *Chemical Engineering Education* **34** (3), 268-271 (2000).
4. G. M. Rogers and J. W. Williams. "Asynchronous Assessment: Using Electronic Portfolios to Assess Student Outcomes," *Proceeding of the 1999 ASEE National Meeting*, Session 2330, Charlotte, NC (1999).
5. L. L. Morris, C. T. Fitz-Gibbon, and E. Lindheim, *How to Measure Performance and Use Tests*, Sage Publishers, Newberry Park, CA (1987).
6. G. H. Roid and T. M. Haladyna, *A Technology for Test-Item Writing*, Academic Press, San Diego (1982).
7. G. J. Robertson, "Classic Measurement Work Revised: An Interview with Editor Robert L. Linn," *The Score*, p.1 (1989)
8. R. Fitzpatrick and E. J. Morrison, "Performance and Product Evaluation," in *Educational Measurement*, R. Thorndike ed., American Council of Education, Washington D.C. (1989).
9. T. W. Banta, J. P. Lund, K. E. Black, and F. W. Oblander, *Assessment in Practice*, Jossey-Bass Inc., San Francisco (1996).
10. V. L. Young, D. Ridgway, M.E. Prudich, D.J. Goetz, B.J. Stuart, "Criterion-based Grading for Learning and Assessment in the Unit Operations Laboratory," *Proceedings of the 2001 ASEE National Meeting*, Albuquerue, (2001).
11. P. T. Terzini and E. T. Pascarella, *How College Affects Students: Findings and Insights from Twenty Years of Research*, Jossey-Bass, Inc., San Francisco (1991).
12. T. A. Angelo and K. P. Cross, *Classroom Assessment Techniques: A Handbook for College Teachers*, 2nd edition, Jossey-Bass Inc., San Francisco (1993).
13. C. R. Pace, "Perspectives and Problems in Student Outcomes Research," in *Assessing Educational Outcomes*, Peter Ewell ed., Jossey-Bass Inc., San Francisco (1985).

Biographical Information

JAMES NEWELL

Jim Newell is an Associate Professor of Chemical Engineering at Rowan University. He currently serves as Secretary/Treasurer of the Chemical Engineering Division of ASEE and has won the Ray Fahien award from ASEE for contributions to engineering education and a Dow Outstanding New Faculty Award. His research interests include high performance polymers, outcomes assessment and integrating communication skills through the curriculum.

KEVIN DAHM

Kevin Dahm is an Assistant Professor of Chemical Engineering at Rowan University. He received his Ph.D. in 1998 from Massachusetts Institute of Technology. Prior to joining the faculty of Rowan University, he served as an Adjunct Professor of Chemical Engineering at North Carolina A&T State University. He also served for one year as a Postdoctoral Researcher at the University of California at Berkeley, where he assisted in the development of ModelLA, a process simulation software package for use in the undergraduate chemical engineering curriculum.

HEIDI NEWELL

Heidi Newell is currently the assessment/accreditation coordinator for the College of Engineering at Rowan University. She previously served as the assessment consultant for the University of North Dakota. She hold a Ph.D. in Educational Leadership from the University of North Dakota, a M.S. in Industrial and Organizational Psychology from Clemson, and a B. A. in Sociology from Bloomsburg University.