RUBRIC DEVELOPMENT AND INTER-RATER RELIABILITY ISSUES

In Assessing Learning Outcomes

JAMES A. NEWELL, KEVIN D. DAHM, AND HEIDI L. NEWELL
Rowan University • Glassboro, NJ 08028

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 other institutions,[3] Rowan University's 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 the education objectives specified in Engineering Criteria 2000 (Criterion 3, A-K) as well as the AIChE requirements and other department-specific goals. Regrettably, over-reliance on survey data often neglects those most qualified to assess student performance—the faculty themselves.

The faculty agreed that student portfolios would provide a valuable means of including faculty input into the process. The difficulty arose when the discussion turned to evaluating the portfolios. Paulson, *et al.*,[4] define portfolios as a "purposeful collection of student work that exhibits the students' efforts, progress, and achievement." As Rogers and Williams[5] noted, however, there is no single correct way to design a portfolio process. Essentially everyone agreed that a portfolio should contain representative samples of work gathered primarily from junior- and senior-year courses. 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 covered in Criterion 3 could be addressed by at least some part of the portfolio. At the same time, we were acutely aware that these portfolios would be evaluated every year and were understandably interested in minimizing the total amount of work collected. Ultimately, we selected the following items:

- 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[6-8] in which a student furnished an authentic response to a given assignment or test question. This format was selected over multiple choice selected response formats because it better represents realistic behavior.[9] The selected-response format presents alternative responses from which the student selects the correct answer; specific selected response formats include true-false, matching, or multiple choice exams, while constructed response formats include essay questions or mathematical

James Newell *is Associate Professor of Chemical Engineering at Rowan University. He is currently Secretary/Treasurer of the Chemical Engineering Division of ASEE. His research interests include high performance polymers, outcomes assessment and integrating communication skills through the curriculum.*

Kevin Dahm *is Assistant Professor of Chemical Engineering at Rowan University. He received his PhD in 1998 from Massachusetss Institute of Technology. Before joining the faculty of Rowan University, he served as Adjunct Professor of Chemical Engineering at North Carolina A&T State University.*

Heidi Newell *is the Assessment Consultant for the College of Engineering at Rowan University. She holds a PhD in Educational Leadership from the University of North Dakota, a MS in Industrial/Organizational Psychology from Clemson University, and a BA in Sociology from Bloomsburg University of Pennsylvania.*

problem solving.[10] Although the items contained in the portfolio provided a wide range of work samples, 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 student's 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's 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 performance level in each area would be required so that all faculty could understand the difference between a 4 and a 2. As Banta[11] stated, "The challenge for assessment specialists, faculty, and administrators is not collecting data but connecting them." The challenge became one of developing rubrics that would help map student classroom assignments to the educational objectives of the program. The four-point assessment rubric also followed the format developed by Olds and Miller[12] for evaluating unit operations laboratory reports at the Colorado School of Mines.

COURSE VS PROGRAMMATIC ASSESSMENT

Other chemical engineering departments are also developing rubrics for other purposes. In their exceptional (and Martin-Award winning) paper on developing rubrics for scoring reports in a unit operations lab, Young, *et al.*,[13] discuss the development of a criterion-based grading system to clarify expectations to students and to reduce inter-rater variability in grading, based on the ideas developed by Walvoord and Anderson.[14] This effort represents a significant step forward in course assessment. The goals of course assessment and program assessment are quite different, however.

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 arise exclusively from the program's educational objectives and fully encompass all of these objectives
- Tests and other graded assignments must completely capture these objectives
- Performance on exams or assignments must be a direct reflection of the student's abilities and not be influenced by test anxiety, poor test-taking skills, etc.

If all of these conditions are met, there should be a direct correlation between student performance in courses and the student's overall learning. Moreover, much of the pedagogical research warns of numerous pitfalls associated with using evaluative instruments (grades on exams, papers, etc.) within courses as the primary basis for program assessment.[15]

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 struggle with the math. In each case, the student could wind up with a C on an exam, but for very different reasons. This is not a problem from the perspective of the evaluation; both students deserve a C. But, 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 entirely ignored[16]. 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 nonclassroom 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 what a novice demonstrates when confronted with a problem. The highest level should show metacognition,[16] the students' awareness of their own learning skills, performance, and habits. To achieve the highest level, students not only have to approach the problem correctly, but they must also demonstrate an understanding of their problem-solving strategies and limitations. The intermediate scores 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 had a five-point Likert scale, so a faculty team set out to develop meaningful scoring rubrics using a five-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 descrip-

tive phrases should stand alone, without the need for additional clarifiers. Ultimately, it was decided to eliminate all labels.

It became 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 definitive 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 metacognitive expert would demonstrate; for the three-point phrases, the target was what a skilled problem solver who lacked metacognition 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; the one-point value captured the performance of a novice problem solver.

To evaluate a given indicator, professors would read the leftmost 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 in Table 1.

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[17] succinctly stated the challenge of accurate assessment, 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 so 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 every time to a given piece of student work.

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

The results, in general, were excellent. Every faculty member scored the items within 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%, all assigned scores were within 1. Rubrics for which this level of agreement was not 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.

The primary example of a rubric that required modification is shown in Table 2. "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

TABLE 1

	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 information	Indentifies some principles but seems to have difficulty in distinguishing what is needed	Cannot identify and assemble relevant information
Systematically 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 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

TABLE 2

	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 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

which scores were not routinely consistent. One heat-transfer exam received a range of scores that included multiple occurrences of both 4 and 1.

In the ensuing discussion, we 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. Interpreting the rubric wording in one way, some faculty chose to assign 4. This interpretation is understandable because no answer was given, and there was no "unrecognized implausible answer." By the letter of the criteria, the student earned a 4. Some faculty interpreted the criteria differently, however, resulting in the assignment of 1. This interpretation is also reasonable—since there were no results, there was 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 three faculty reviewed an item. For example, in the score set 2, 2, 2, 2, 1, 3, 2; the mean score assigned by the faculty was 2, and the mean of a three-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. We found (after one rubric was revised as described above) that 95% (153 of 160) of the score sets showed this level of consistency. Thus, we 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>

CLOSING THE LOOP

Ultimately, the purpose of gathering detailed assessment data is to improve student learning. Once each year, we review the data in a two-day assessment meeting[3] where we discuss all aspects of the program, including the data from each tool. We identify strengths and areas for improvement and make decisions affecting curriculum and policies. Specific changes resulting from these meetings have included a decision to introduce product engineering and economics earlier in the curriculum and to adjust topical coverage in thermodynamics.

THE NEXT LEVEL

The next goal is to use the rubrics to help guide selection of course objectives across the curriculum. With detailed educational objectives in place and rubrics to assist in their assessment, we hope improved course objectives will be developed that more directly link classroom activities and evaluations with the program goals. The rubrics described in this paper should provide the basis for a more in-depth, formative assessment. Although the ABET criteria are summative, the educational process itself centers around formative changes, incrementally enhancing a student's knowledge, skill set, and problem-solving capabilities.

CONCLUSIONS

A complete set of rubrics was developed and tested that maps student performance of a variety of junior/senior-level 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 members evaluated 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.

REFERENCES

1. Engineering Accreditation Commission, *Engineering Criteria 2000*, Accreditation Board for Engineering and Technology, Inc., Baltimore (1998)
2. Diamond, R.M., *Designing and Assessing Courses and Curricula: A Practical Guide,"* Jossey-Bass Inc., San Francisco (1998)
3. Newell, J.A., 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," *Chem. Eng. Ed.*, **34**(3), p. 268 (2000)
4. Paulson, L.F., P.R. Paulson, and C. Meyer, "What Makes a Portfolio a Portfolio?" *Educational Leadership*, **48**(5), p. 60 (1991)
5. Rogers, G.M., and J.M. Williams, "Asynchronous Assessment: Using Electronic Portfolios to Assess Student Outcomes," *Proc. of the 1999 ASEE Nat. Mtng.*, Session 2330, Charlotte (1999)
6. Morris, L.L., C.T. Fitz-Gibbon, and E. Lindheim, *How to Measure Performance and Use Tests*, Sage Publishers, Newberry Park, CA (1987)
7. Roid, G.H., and T.M. Haladyna, *A Technology for Test-Item Writing*, Academic Press, San Diego (1982)
8. Robertson, G.J., "Classic Measurement Work Revised: An Interview with Editor Robert L. Linn," *The Score*, p.1 (1989)
9. Fitzpatrick, R., and E.J. Morrison, "Performance and Product Evaluation," in *Educational Measurement*, R. Thorndike ed., American Council of Education, Washington DC (1989)
10. Erwin, T. Dary, *Assessing Student Learning and Development*, Jossey-Bass, San Francisco (1991)
11. Banta, T.W., J.P. Lund, K.E. Black, and F.W. Oblander, *Assessment in Practice*, Jossey-Bass Inc., San Francisco (1996)
12. Olds, B.M., and R.L. Miller, "Using Portfolios to Assess a ChE Program," *Chem. Eng. Ed.*, **33**(2), 110 (1999)
13. Young, V.L., D. Ridgway, M.E. Prudich, D.J. Goetz, B.J. Stuart, "Criterion-based Grading for Learning and Assessment in the Unit Operations Laboratory," *Proc. of the 2001 ASEE Nat. Mtng.*, Albuquerque (2001)
14. Walvoord, B.E., and V.J. Anderson, *Effective Grading: A Tool for Learning and Assessment*, Jossey-Bass Inc., San Francisco (1998)
15. Terzini, P.T., and E.T. Pascarella, *How College Affects Students: Findings and Insights from Twenty Years of Research*, Jossey-Bass Inc., San Francisco (1991)
16. Angelo, T.A., and K.P. Cross, *Classroom Assessment Techniques: A Handbook for College Teachers*, 2nd ed., Jossey Bass Inc., San Francisco (1993)
17. Pace, C.R., "Perspectives and Problems in Student Outcomes Research," in *Assessing Educational Outcomes*, Peter Ewell ed., Jossey-Bass Inc., San Francisco (1985) ❐