National Institute for Learning Outcomes Assessment

Being Confident about Results from Rubrics

February 2012

Thomas Judd, Charles Secolsky & Clayton Allen

Using rubrics to assess student learning is more and more common, and their use is almost certainly going to increase, as the Association of American Colleges and Universities (AAC&U) essential learning outcomes become better known and the Lumina Degree Qualifications Profile gains traction. Both outcomes frameworks require something more than what available standardized instruments measure. This is one reason the AAC&U VALUE rubrics are receiving attention.

But here's the dilemma. Rubrics may be an adequate measure of individual student learning in the classroom context. But when rubrics are used to measure student learning outcomes at the course, program or institutional level, concerns may arise about the confidence institutional leaders have in using the results of rubrics to guide institutional improvement. To date, relatively little attention has been paid to the analysis of rubric scores made by raters of portfolios, essays and the like. While inter and intra-rater reliability may be familiar concepts in establishing confidence in interpreting measures, rubrics bring a somewhat different set of issues. We examine some of the issues inherent in measuring student learning outcomes via rubrics that may affect the confidence with which generalizations can be made. The rubric criteria, the raters and the statistical interaction between criteria and raters add variance to the measures that could be misattributed to student learning differences. The outcomes assessment loop can more accurately come full circle when curricular changes are made based on student performance on rubrics for portfolios and essays. But, how much of the improvement in performance can be attributed to scores on the criteria themselves without taking into account variance of scores for raters and the rater by criteria interaction?

According to Walvoord (2004), a "rubric articulates in writing the various criteria and standards that a faculty member uses to evaluate student work" (p. 19). There exist two types of rubrics –holistic and analytic. Holistic rubrics assess overall quality of a performance or product and can vary in complexity, while an analytical rubric takes the form of a matrix comprised of two or more dimensions (criteria and rating scales and descriptors). As noted earlier, the VALUE rubrics prepared by AAC&U and available on their website, are examples of analytic rubrics that present criteria, rating scales and descriptors. Comparisons of criterion scores are typically enabled by keeping descriptors constant. Although it takes time to develop clearly-defined and unambiguous descriptors, they are essential in communicating performance expectations in rubrics for facilitating the scoring process.

Linn, Baker and Dunbar (1991) state that rubric ratings for portfolios are suited for validation studies because the recently expanded concept of validity includes the concepts of consequences, fairness, transfer and generalizability, cognitive complexity, content quality, meaningfulness and cost efficiency. In line with the promise of portfolio rubrics for validation of interpretation of rubric scores, Wolf, Bixby, Glenn and Gardner (1991) argue for a reformulated psychometric model that does not rely on simple progression from novice to expert but instead considers multiple paths to excellence (Moss, 1992).

Generalizability (G) theory is a statistical approach for evaluating the reliability or dependability of behavioral measurements (see Brennan, 2001; Cronbach, Gleser, Nanda, & Rajaratnam 1972; Shavelson & Webb, 1991; Webb, Shavelson & Steedle, 2012), and allows the researcher to determine whether raters or criteria or tasks are a major source of measurement error which, if unaccounted for, could be misinterpreted as performance differences. Generalizability theory can be used for computing variance components to separate the variance attributable to these factors and student learning.

> While inter and intra-rater reliability may be familiar concepts in establishing confidence in interpreting measures, rubrics bring a somewhat different set of issues.

For a carefully designed and extensively used rubric for Graphic Design course student projects in a northeast United States college, data in the form of rubric scores were collected on 17 students using four different criteria and four raters. The criteria for judging students' products were: Technique, Design, Creativity and Concept, and Presentation. All four raters scored the work products on all four criteria. However, the work assignment or topic embodied in the task had not been taken into account as Secolsky and Wentland (2010) recommend. Although the sample size was small, variance components were computed using SPSS which provides estimates of the variances attributable to the sources of variance described above: due to criteria, raters, and their interaction. It is posited that recommendations for improvement in the curriculum being assessed using the outcomes assessment loop are more feasible to occur when the interaction term between criteria and raters is small. This source of variation is the proportion of variation attributable to both the different criteria and the different raters. Such information could be used to supplement descriptor only rubrics as do typically exist in higher education assessment.

A rubric with the four criterion scores was used to evaluate performance of students' work products. The greatest extent of quantification occurred with means of criteria by individual raters. As is the case with many rubrics that are developed for scoring purposes, no attempt was made to supplement rubric scoring with more sophisticated statistical measures.

The values of the two variance components were presented elsewhere: one for criteria, one for raters and their interaction term. We contend that if the variance attributable to criteria exceeds the variance attributable to raters and the interaction between criteria and raters is small then the outcomes assessment loop can more comfortably be closed. Otherwise, what could be viewed as a need for improvement may actually be masking a difference due to rater scoring and inter-rater-reliability. If it is the case that inter-rater reliability is already computed by the department that scores the rubric, then the interaction term can make or break the sound use of rubric data as the evidentiary basis for recommending curricular improvements.

While no rule of thumb exists for determining high and low variance for the interaction term, the percentage was not that great in the graphic design course example. A smaller percentage for the interaction term would have been preferable for improving mean measures for criteria using outcomes assessment.

The results of this analysis show that in this case, the greatest proportion of variance is related to the criteria themselves, with relatively small to moderate interaction between criteria and raters. Therefore, one can be confident that the assessment result from using the rubric is not being overly masked primarily from the interaction between criteria and raters. It is therefore more likely that the curriculum can be improved with a focus on raising mean scores on criteria without much regard to a confounding that may be due to a rater by criterion interaction.

Based on this experience, we suggest that users of rubrics pay attention to the meaning of each score for each criterion. Doing so affords the opportunity of facilitating clear definition of rubric categories with little need to consider the impact of inter-rater reliability. If some doubt still exists about inter-rater reliability, especially with a newly designed rubric, then the procedure discussed here may be valuable for improving the rubric in trial runs or early uses of the rubric.

Although it takes time to develop clearly-defined and unambiguous descriptors, they are essential in communicating performance expectations in rubrics for facilitating the scoring process.

References

Brennan, R.L. (2001). *Generalizability theory*. New York, NY: Springer-Verlag.

Cronbach, L.J., Gleser,G.C., Nanda,H. & Rajaratnam, N. (1972). *The dependability of behavioral measurements: Theory of generalizability for scores and profiles*. New York, NY: John Wiley.

Linn, R.L., Baker, E.L., & Dunbar, S. B. (1991). Complex, performance-based assessment: expectations and validation criteria. *Educational Researcher, 20*(8), 15-21.

Moss, P. (1992). Shifting conceptions of validity in educational measurement: Implications for performance assessment. *Review of Educational Research, 62*(3), 229-258.

Secolsky, C. & Wentland, E. (2010). Differential effect of topic: Implications for portfolio assessment. *Assessment Update: Progress, Trends, and Practices in Higher Education, 22*(1), 1-2, 15.

Shavelson, R.J., & Webb, N.M. (1991). *Generalizability theory: A primer*. Newbury Park, CA: Sage.

Walvoord, B. E. (2004). *Assessment clear and simple: A practical guide for institutions, departments, and general education*. San Francisco, CA: Jossey-Bass.

Webb, N.M., Shavelson, R.J., & Steedle, J.T. (2012). Generalizability theory in assessmen contexts. In C. Secolsky & D. B. Denison (Eds.). *Handbook on Measurement, Assessment, and Evaluation in Higher Education* (pp. 132-149). New York, NY: Routledge.

Wolf, D., Bixby, J., Glenn, J, & Gardner, A. (1991). To use minds well: Investigating new forms of student assessment. *Review of Research in Education 17*, 31-74.

National Institute for Learning Outcomes Assessment

Please Cite As:

Judd, T., Secolsky, C. & Allen, C. (2012, February). *Being confident in results from rubrics*. Urbana, IL: University of Illinois and Indiana University, National Institute for Learning Outcomes Assessment (NILOA).