Using Rater Agreement Analysis to Refine an Oral Presentation Rubric and Improve Inter-Rater Reliability

Gary May, Kathryn O'Neill, Neelam Sharma

Clayton State University/RockTenn /RockTenn

Assurance of learning is a hot topic in higher education. State legislatures, regional and professional accreditation agencies, and employers are asking a key question: are we graduating students who actually have the knowledge and skills that we promise (Martell & Caldron, 2005; Suskie, 2004)? Reflecting this movement, the Association to Advance Collegiate Schools of Business (AACSB) has established new accreditation standards requiring business schools to produce direct evidence of learning in their courses and programs (AACSB, 2003).

Implementing and maintaining the type of on-going comprehensive assessment plan called for in the AACSB standards presents many challenges related to time, resources, and culture, and often generates significant resistance from faculty. In many cases, faculty perceive the increased emphasis on assessment as a threat to their academic freedom, an additional demand on their time, and another tool to be used as a form of performance evaluation (Walvoord, 2004).

One especially challenging issue is requiring faculty who teach different sections of the same course to agree on measurable learning outcomes and use a common assessment method, such as embedded exam questions or a rubric, to assess the learning. While embedding common exam questions for cognitive outcomes is a fairly straightforward process, the use of rubrics is more difficult. Typically, rubrics are used to assess application and integration skills, such as writing and oral presentations, and are more subjective in nature. The use of rubrics raises a key question, especially when the data is collected across class sections from different instructors: is the data reasonably reliable? In other words, if two instructors viewed the same performance, would they agree on the rating for each performance element? Without some reliability in interpretation of the rubric, the composite data from multiple sections will not be very useful for demonstrating the mastery of skills and knowledge, and for improving teaching and learning.

This case study documents how two business school professors are working collaboratively with two external business professionals to refine an oral presentation rubric and to improve inter-rater reliability. The study demonstrates the use of rater agreement analysis to identify problems with rubric design and to focus discussion on ways to improve rater agreement. The study makes a clear case for the need for rater training if assessment data from rubrics is to be useful for improving teaching and learning. Before we share our experience, it will be helpful to set the context by providing some background on the school.

Contextual background

Thinking about contextual variables is important when attempting to generalize or compare assessment practices. Smaller schools with a primary mission of teaching have a different set of concerns and challenges at the faculty level regarding assessment than large research-oriented institutions. For example, small-school faculties tend to have heavier teaching loads and access to fewer support resources. On the other hand, gaining faculty buy-in and coordinating faculty efforts is more difficult in large schools due to the number of faculty and the expected focus on research (AACSB, 2007). The context for this study is a small business school environment.

Clayton State University (CSU), a unit of the University System of Georgia, is primarily an undergraduate institution with a mission focused on teaching excellence. CSU serves the southern metropolitan area of Atlanta with an enrollment of about 6,000 students. While student dorms have recently been added, CSU is still primarily a commuter school, with approximately 60% of the students classified as non-traditional (typically students working full time and with families). The average age is 28.

The School of Business, with 24 full-time faculty, is AACSB accredited and offers four majors (Accounting, Marketing, Management, and General Business). The School has been working on adapting to the new AACSB assurance of learning standards since 2004. The organizational structure is flat, with a Dean and Associate Dean. All faculty members report directly to the Associate Dean. An assessment committee, made up of a chair and four members representing the different disciplines, guides the assurance of learning effort with full faculty involvement.

Because of the importance of communication skills to employers, written and oral communications are at the top of the School's overall program outcomes. Therefore, the faculty has spent a great deal of time building a robust assessment process for these important skills. It has been an iterative process, with many adjustments along the way, and is still a work in progress. The School's Managerial Communication course serves as the primary assessment vehicle for written and oral communication skills. This required course is usually taken at the junior level. The two professors who teach the Managerial Communication course led the work on the assessment design and enlisted external business professionals to assist with the development and validation of the assessment process. In the case of oral communications, the focus of this paper, two corporate professionals, one with a masters in Human Resource Development, and the other with a Ph.D. in Human Resource Development, were enlisted to assist with the rubric development and validation process. Both professionals are consultants in the Talent Management group of a large, multinational manufacturer. Both consultants have extensive experience in coaching managers on oral presentation skills.

The literature on oral presentation rubric design and rater training

There appears to be only limited empirical investigation of rubric designs and reliability related to oral presentations in an academic environment; we could find only one article published in a refereed journal. Dunbar, Brooks, and Kubicka-Miller (2006) used *The Competent Speaker*, a

rubric developed by the National Communication Association, to evaluate student performance in general education public speaking courses. They found students below satisfactory standards on five of the eight competencies defined by the rubric. The study highlights the challenges in using a rubric for assessment and provides recommendations for rubric design. The authors recommend that rubrics be tailored to the specific assignment and performance definitions by element, and be written in a very specific manner to reduce the amount of inference by evaluators. In addition, the study emphasizes the importance of conducting rater training.

Compared to research on oral communication rubric designs, the literature on rater training is extensive. Two comprehensive literature reviews provide a good summary of the prescriptive techniques for training raters. Smith (1986), in his review of the rater-training literature, identified three training methodologies that lead to improved rating accuracy:

1. *Rater Error Training*: presenting raters with examples of common rating errors such as leniency, halo, central tendency, and contrast errors, and alerting raters to potential biases such as the similar-to-me effect.
2. *Performance Dimension Training*: familiarizing raters with the dimensions and rating scale by which the performance is rated.
3. *Performance Standards Training*: providing raters with a frame of reference for rating performance by providing feedback on practice ratings compared to "true" ratings assigned by trained experts.

According to Smith, the research suggests that the best way to increase rating accuracy is to combine the three approaches. He concluded that "the more actively involved raters become in the training process, the greater the outcome. Providing raters with the opportunity to participate in a group discussion along with practice and feedback exercises produces better results than presenting the training material through a lecture" (p. 37).

Woehr and Huffcutt (1994), in their review of the rater-training literature, included a fourth dimension to the training approaches – behavioral observation training. This approach focuses on strategies to improve and recall observations of specific behaviors related to the performance dimensions through the use of techniques such as note taking, diaries, and frequency counts. According to Woehr and Huffcutt, the research data suggest that behavioral observation training may be a very effective approach to increase rating accuracy. Like Smith (1986), they also argue for a combination of rater-training strategies to increase the effectiveness of rater training.

Methods

Subjects. Participants in the study included 106 undergraduates (mostly juniors) from multiple sections of the Managerial Communications course in the School of Business at Clayton State University over three semesters. All students in each course section were included in the data analysis if they completed an IRB voluntary consent form. Some sections were taught by different instructors, including one adjunct. The gender mix was 71% female and 64% were classified as minority (predominantly African-American). The average age was 26, and 42% considered themselves to be "non-traditional" students, i.e., out of school for 4 or more years prior to returning for a degree.

Rubric design. The project team, consisting of two business school professors and two external business professionals, adapted the existing oral communication Likert-scale rubric and, using published criteria for rubric design (Arter & McTighe, 2001; Huba & Freed, 2000, Wiggins & McTighe, 2001), created a three-level rubric for assurance of learning assessment purposes. The three levels of the rubric include "Unsatisfactory," "Satisfactory," and "Good." The three levels were used to be congruent with other School of Business assessment initiatives using rubrics, which report the percentage of students scoring at each level by performance element. The oral presentation rubric assesses five performance elements: opening, body, closing, visuals, and physical delivery. Each performance element, in turn, lists three or more competencies. The rubric can be scored at the element level or at the more detailed competency level. The rubric was tailored to the specific oral presentation final assignment in the Managerial Communication course, a problem-solution persuasive presentation. Table 1 provides an example portion of the initial version of the rubric.

Table 1
Oral Presentation Rubric (Sample Section)

Performance Element	Unsatisfactory	Satisfactory	Good
Opening			
✓ Attention	❑ No attempt to gain audience's attention	❑ Gains audience's attention with a startling statement, anecdote, question, or quotation	❑ Gains audience's attention with a startling statement, anecdote, question, or quotation and establishes common ground
✓ Purpose and Benefits	❑ No clear purpose statement or indication of benefits for the audience	❑ Provides a general statement of purpose and identifies at least one benefit for the audience	❑ Describes the problem, the questions to be answered, the benefits to the audience, and the rhetorical purpose of the presentation
✓ Overview	❑ Does not provide an overview of the presentation	❑ Provides a general overview of the topics to be covered	❑ Provides a general overview of the topics to be covered,; notes the expected length of the presentation, suggests a plan for handling questions, and asks for affirmation

Evaluation of rater agreement. Since our purpose was to determine the degree of rater agreement on the three possible ratings (unsatisfactory, satisfactory, good) for each performance competency, we used a simple agreement matrix and calculated the percentage agreement for each competency. Table 2 shows an example agreement matrix. The diagonal scores in bold represent rater agreement and the percentage agreement is calculated as a sum of the diagonal numbers divided by the total number of observations.

Table 2

Example Rater Agreement Matrix for a Given Competency

		Rater 1		
		1 (Unsat)	2 (Sat)	3 (Good)
Rater 2	1 (Unsat)	**9**	6	1
	2 (Sat)	2	**7**	4
	3 (Good)	0	3	**7**
			Percentage Agreement	**59.0%**

We also experimented with Cohen's kappa statistic as a measure of rater agreement (Cohen, 1960). Kappa is usually preferred over simple agreement percentages because it corrects for random chance agreement (Bakeman & Gottman, 1986). The kappa coefficient has a range from 0 to 1.00, with larger values indicating better reliability. As a general rule of thumb, kappa scores ranging from .40 to .60 can be characterized as fair agreement, .60 to .75 as good, and over .75 as excellent (Fleiss, 1981). However, due to our small sample size, we were not able to calculate kappa for some agreement matrices. For example, a few data sets recorded all zeros in the "Unsatisfactory" and "Good" columns of the agreement matrix. Because the kappa formula includes multiplication of column totals times row totals, the formula returns a zero or a distorted number. SPSS reports such situations as incalculable.

After some discussion with our assessment committee, we determined that simple percentage agreement was adequate for our purposes of stimulating dialogue about rubric design and identifying problems with rater agreement. Our goal is to improve student learning through improvement of our assessment processes. Martel and Calderon (2005) make the distinction between "scholarly rigor" and "academic rigor" in the development of assessment practices:

> Scholarly research has a set of standards (proven methods, replicated results, scientific sampling) that is not usually appropriate for program assessment of student learning. Not only are these standards often impossible to meet since assessment, as a field of inquiry, is still in its infancy, but demands for rigor can stall – even strangle – progress. An honest effort to investigate student learning through direct measures is what is required – not meeting standards for peer-reviewed research. (p. 24)

*Assessment process***.** The last half of the semester, students were assigned an individual project to design and deliver a problem-solution persuasive oral presentation based on a case situation. The students worked and presented in collaborative learning groups, and each student within a group had a different case topic. This assignment represents 10% of the course grade and was graded by the instructor at the time of delivery. For the purposes of this study, the presentations were captured by digital video and uploaded to a Web site, allowing the external professionals to assess each student's presentation independently using the three-level rubric.

The study proceeded in three phases by semester. The initial phase was conducted as a pilot, where the student videos (N=31) were assessed by the two raters, who rated separately and then agreed upon combined ratings, without discussion of the rubric dimensions or rater training. The purpose of the pilot iteration was to refine the rubric. After the rater agreement statistics were calculated, the raters met to discuss areas of disagreement (low percentage of agreement) and made clarifying edits to the rubric. For example, a classic error in rubric design is including more than one behavior in the performance description for a given competency. This error can be seen in Table 1, where purpose and benefits are listed together. If the observer hears a purpose statement but no benefits statement, should that be counted as unsatisfactory? A very low agreement percentage indicated confusion with the problem / benefit item. The solution, in this case, was to separate the two into separate line items as discrete behavior statements.

For the next phase (called Round 1), the raters used the revised rubric independently to assess the next semester's presentations. Again, after calculation of agreement measures, the raters met with the professors to discuss results. No additional changes were made to the rubric, but all agreed to rater training to improve agreement.

Prior to rating the third semester's presentations (Round 2), the raters participated in two hours of rater training. The rater training followed Smith's (1986) recommended training methodologies on rater error training, performance dimensions, and performance standards. After discussion of typical rating errors and a discussion of the performance descriptions by competency, the two raters rated a selection of presentations independently and then compared their ratings with each other and to a "true" rating provided by the lead instructor for the course. The raters and the lead instructor then discussed their ratings until consensus was reached. The raters then used the resulting agreements on rubric interpretation to rate independently the next group of presentations.

Results

Table 3 provides a summary of the percentage agreements by performance element and competency for Round 1 (refined rubric, no rater training) compared to Round 2 (refined rubric, rater training). The results show that the mean percentage agreement score for Round 2 increased to 70.3 compared to 59.4 for Round 1. An independent t-test indicated that increase was statistically significant, $t_{(05, 42)} = 2.27$, $p = .029$.

Discussion

The results indicate some significant improvement in rater agreement with the addition of rater training. However, the results also highlight the great challenge of achieving satisfactory rater agreement across all competencies for a complex behavioral activity like oral presentations. While it is commonly accepted that rater agreement percentages are "good" if they are in the 90s (Bakeman & Gottman, 1986), we feel even 80% agreement would make the assessment data usable. However, examination of Table 3 indicates that 13 of the 22 competencies have agreement ratings of less than 80%, with 7 in the 50% range. Clearly, we have more work to do to improve both the rubric and the rater training process. Our plan is to continue the process of

rubric review, rater training, and rater consensus discussions until we approach the 80% target for all 22 competencies.

Given these results, imagine the challenge of collecting data across multiple sections from instructors who have not participated in rater training. Most likely, the use of the data for assurance of learning purposes would be almost meaningless.

Table 3

Rater Agreement by Competency: Oral Presentation Assessments

		% Rater Agreement	
		Round One	Round Two
Performance		Rubric Edit	Rater Training
Element	Competency	N = 39	N = 36
Opening	Attention	64.1	69.4
	Purpose	43.6	52.8
	Benefits	48.7	75.0
	Overview	84.6	100.0
Body	Organization	51.3	83.3
	Audience Centered	59.0	91.7
	Persuasion: Credibility	56.4	58.3
	Persuasion: Logic	46.2	47.2
	Persuasion: Emotional Appeal	74.4	94.4
Closing	Summary	59.0	66.7
	Call to Action	61.5	91.7
	Memorable Ending	59.0	50.0
PowerPoint	Concentration	61.5	86.1
	Clarity	43.6	66.7
	Consistency	48.7	52.8
	Correctness	64.1	50.0
	Control	89.7	94.4
Delivery	Appearance	71.8	94.4
	Eye Contact	48.7	44.4
	Poise and Confidence	61.5	69.4
	Voice	46.2	50.0
	Enthusiasm	63.6	58.3
	Composite Mean	59.4	70.3

One outcome of this study is the decision to use outside evaluators for complex rubric assessments such as oral and written communication. We believe a trained external evaluator, scoring multiple class sections semester after semester, will give us more consistent and usable data for assurance of learning purposes.

As the demand for assurance of learning increases, the need to be sure about the validity and accuracy of the results will become more and more important. As demonstrated by these

findings, it is easy to underestimate the complexity of observation-type assessments and ratings. Scholars who routinely apply inter-rater reliability tests for their academic research might not think to do so for assurance of learning ratings, especially when observable performance seems to be self-evident.

At the same time, this study demonstrates the possibility of arriving at standards and procedures that lead to successful outcomes, making it easier for institutions and faculties to arrive at quality assurance of learning results more quickly. The hidden complexity of assurance of learning as demonstrated by this project should lead to further investigation of how complexity affects assessment efforts in a variety of disciplines with an eye towards identifying successful strategies.

Improvement in assurance of learning processes will lead to more confidence in the work of our universities and faculties by all constituencies in both the public and private sectors, and demonstrate more clearly the contributions our institutions make to the society.

References

AACSB (2003). *Eligibility standards for business accreditation.* St. Louis, MO: Association to Advance Collegiate Schools of Business.

AACSB (2007). Applied Assessment Seminar. Tampa, FL: Association to Advance Collegiate Schools of Business.

Arterm H., & McTighe, J. (2001). *Scoring rubrics in the class room: Using performance criteria for assessing and improving student performance*. Thousand Oaks, CA: Corwin Press.

Bakeman, R., & Gottman, J. (1986). *Observing interaction: An introduction to sequential analysis*. Cambridge: Cambridge University Press.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 20*, 37-46.

Dunbar, N., Brooks, C., & Kubicka-Miller, T. (2006). Oral communication skills in higher education: Using a performance-based evaluation rubric to assess communication skills. *Innovative Higher Education, 31*, 115-128.

Fleiss, J. (1981). *Statistical methods for rates and proportions*. New York: Wiley.

Huba, M., & Freed, J. (2000). *Learner-centered assessment on college campuses: Shifting the focus from teaching to learning*. Boston: Allyn and Bacon.

Martell, K. & Calreron, T. (2005). Assessment in business schools: What it is, where we are, and where we need to go now. In K. Martell & T. Calderon (Eds.), *Assessment of student learning in business schools*: Vol. 1, No. 1 (pp.1-26). Tallahassee, FL: Association for Institutional Research.

Smith, D. E. (1986). Training programs for performance appraisal: A review. *Academy of Management Review, 11*, 22-40.

Suskie, L. (2004). *Assessing student learning: A common sense guide*. Bolton, MA: Anker Publishing.

Wiggings, G. & McTighe, J. (2001). *Understanding by design.* Upper Saddle River, NJ: Prentice-Hall.

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

Woehr, D. J., & Huffcutt, A. L. (1994). Rater training for performance appraisal: A quantitative review. *Journal of Occupational & Organizational Psychology, 4*, 189-216.

Biographies

GARY MAY is an Associate Professor of Management in the School of Business at Clayton State University, a unit of the University System of Georgia. He holds a Business Administration degree from Duke University and a Masters and Ph.D. in Human Resource Development from Georgia State University. Gary has the distinction of being the first Chief Learning Officer in corporate America and brings 32 years of business experience to the classroom.

KATHRYN O'NEILL is Senior Consultant, Talent Management, at RockTenn, one of North America's leading manufacturers of paperboard, containerboard, packaging and merchandising displays with $3 billion in sales. She holds a Bachelor of Arts from Texas Tech University, and a Master of Arts in Communication and a Ph.D. in Human Resource Development from Georgia State University. Kathryn has been involved with workplace learning since 1978, and has taught written and spoken communication skills on the college level.

NEELAM SHARMA is Consultant, Talent Management at RockTenn. She holds a Bachelor of Science in Packaging Design from Rochester Institute of Technology, and a Master of Arts in Human Resource Development from Georgia State University. Neelam has taught Presentation Skills to subject matter experts in the workplace and other settings for over 12 years.