AC 2012-4619: WORKFORCE COMMUNICATION INSTRUCTION: PRELIMINARY INTER-RATER RELIABILITY DATA FOR AN EXECUTIVE-BASED ORAL COMMUNICATION RUBRIC

Dr. Tristan T. Utschig, Georgia Institute of Technology

Tristan Utschig is a Senior Academic Professional in the Center for the Enhancement of Teaching and Learning and Assistant Director for the Scholarship and Assessment of Teaching and Learning at the Georgia Institute of Technology. In this role, he consults with faculty about planning and assessing educational innovation in the classroom. He also serves as an evaluator on educational research grants. Formerly, he was tenured Associate Professor of engineering physics at Lewis-Clark State College. Utschig has regularly published and presented work on a variety of topics, including assessment instruments and methodologies, using technology in the classroom, faculty development in instructional design, teaching diversity, and peer coaching. Utschig completed his Ph.D. in nuclear engineering at the University of Wisconsin, Madison, where he worked on safety issues for fusion reactor designs.

Dr. Judith Shaul Norback, Georgia Institute of Technology

Judith Shaul Norback received her B.A. from Cornell magna cum laude and her master's and Ph.D. from Princeton. She has worked in the area of workplace communication skills for 25 years, starting at Educational Testing Service in 1987, then founding and directing the Center for Skills Enhancement, Inc., in 1993. Her clients included the National Skill Standards Board, the U.S. Department of Labor, and many universities. Norback joined Georgia Tech in 2000 to focus on the workplace communication skills of engineers and is general faculty and Director of Workplace and Academic Communication in the Stewart School of Industrial and Systems Engineering. In 2003, she founded the Workforce Communication Lab, which has had more than 16,000 student visits to date. The instruction she developed has been shown to make a significant difference in students' presentation skills during five semesters so far. Norback has published in refereed journals and conference proceedings, presented at national conferences, and is now Program Chair for her division in ASEE, VP of External Relations for INFORMS-ED, and Chair for Student Involvement for the 2012 Capstone Design Conference. She is working on a book called "Oral Communication Excellence for Engineers: What the Workforce Demands" for John H. Wiley & Sons (due in 2013) and several articles, while continuing to teach capstone design communication instruction and a course on journal article writing for graduate students. Her current research focus includes evaluating the reliability of the scoring rubric she and Tristan Utschig developed from executive input and identifying the cognitive schema used by students to create graphs from raw data.

Jeffrey S. Bryan, Georgia Institute of Technology

Jeffrey S. Bryan is currently in his first year of Georgia Tech's M.S. program in digital media. He attended Southern Utah University as an undergraduate, and majored in English education. He worked for several years as a trainer for AT&T, teaching adult learners, and as an Editor for an opinion research company. He currently works as a Graduate Research Assistant in Georgia Tech's Center for the Enhancement of Teaching and Learning (CETL), where he assists with assessment and data analysis for ongoing CETL projects. His master's thesis involves an investigation of choice and transgression in video game storytelling.

Workforce Communication Instruction: Preliminary Inter-rater Reliability Data for an Executive-based Oral Communication Rubric

Abstract

We have conducted a preliminary study to measure the degree of agreement among different raters (inter-rater reliability) for an executive-based scoring rubric used to rate oral engineering student presentations on 19 skills. We explore the question: do different raters give the same feedback on the same presentation? We have collected scores from raters in three different contexts: (1) the researchers and teaching assistants rating videotaped presentations from capstone design, (2) the researchers and a group of ASEE workshop attendees rating different videotaped presentations, and (3) one researcher and students rating presentations in a course. We analyzed the data collected from the raters using three different measures: first, we conducted pairwise comparisons among raters of the same presentations for frequency of exact matches; second, we conducted pairwise comparisons for frequency of 1-point score consistency (on a five-point scale); and third, we calculated criterion referenced frequencies (rater consistency compared to a "true" score given by the rubric developers) for both exact matches and 1-point score consistency. Our analysis of ratings for each of the 19 skills shows, in general, that the reliability of the overall rubric is acceptable. However, there is some variation in the reliability of each skill. In particular, we see

a. High reliability for 7 of the skills,
b. Good reliability for an additional 4 skills,
c. Reasonable but lower reliability for 8 skills.

Schools will be able to use the results of this study to identify the most reliable parts of this rubric and to enhance the reliability between different raters using the rubric. Further, these results will be used, in conjunction with both formal and informal user feedback, to improve the overall reliability of this rubric, and specifically, the reliability of those skills showing less consistency among raters.

Introduction

Since the Accreditation Board of Engineering and Technology (ABET) passed criteria in 2000 and 2004[1], many engineering schools have been working to implement effective oral presentation in their instruction. But the problem of engineering students' lack of oral presentation skills persists. At Georgia Tech in the Stewart School of Industrial and Systems Engineering, we have been able to build, implement, and test a scoring system based on empirical research (in particular, executive input). The input was provided by over 66 executives, with engineering degrees, representing a wide variety of settings[2].

We have conducted a preliminary study to measure the degree of agreement among different raters for an executive-based scoring rubric used to rate oral presentations. We developed the oral presentation rubric based on several learning theories and input from executives representing a wide range of settings, all with engineering degrees. To date, the rubric has been used for five

semesters in capstone design. Previous analysis has shown that students' presentation skills have improved significantly when the rubric and its supplemental materials have been used over the course of a semester[3, 4]. In this study we explore the degree of agreement between raters or inter-rater reliability for each of the 19 skills represented in the rubric. Our goal is to answer the question: do different raters give the same feedback on the same presentation?

To study the reliability across the raters using the rubric, we have collected scores from raters in three different contexts: (1) two researchers and four teaching assistants rated 21 videotaped presentations from capstone design, (2) two researchers and a group of nine ASEE workshop attendees rated two different videotaped presentations, and (3) one researcher and his students rated 20 peer presentations in a class of 78 students.

We analyzed the data collected from the raters using the following procedure: (1) we conducted pairwise comparisons among raters of the same presentations for frequency of exact matches, (2) we conducted pairwise comparisons for frequency of 1-point score consistency (on a five point scale), and (3) we calculated criterion-referenced frequencies (rater consistency compared to a "true" score given by the rubric developers) for both exact matches and 1-point score consistency.

The results identified in this study will be used in conjunction with both formal and informal user feedback to modify the existing rubric. These modifications will have two aims – enhance ease of use and understanding of the rubric, and enhance reliability among different raters using the rubric. With these modifications completed, other schools using the modified rubric will be able to enhance reliability among scorers, and apply the rubric with even greater confidence that the feedback students receive will enhance their performance in oral presentations.

Background

The Norback-Utschig Presentation Scoring System is based on three learning theories, input from 66 executives with engineering degrees working in a variety of settings, and input from engineering faculty.

The first learning theory is *task value theory*[5, 6, 7]. The theory focuses on the question: "Why am I doing this task," and includes the importance of the task, interest in the task, and the utility and cost value of the task. Students taking our instruction are repeatedly reminded that the scoring system is based on input from executives in many firms.

The second theory is from the study of rhetoric and emphasizes *discourse communities* in which communication occurs. The second theory is *genre analysis*[8], which focuses on not only the form and technical directness of a created document, but also the purpose of the document, its social context, and the action it invokes in others. To apply this theory, a detailed analysis of the students' audience, whether it includes faculty or technical and non-technical client representatives, is included in our instruction.

The third theory is *situated learning*, in which the context of an assignment can be as important as the assignment[9, 10, 11, and 12]. For example, the capstone design teacher should be clear about

how communication assignments serve to support engineering design. And, including authentic tasks from engineering work will help students transfer their learning to the workplace. Our instructors point out that presentations are an effort to learn skills needed on the job, and faculty and client questions are referred to as information exchanges that help improve the students' projects.

Executives provided input for the Presentation Scoring System through focus groups, panel discussions, and telephone interviews. This information was condensed, staying true to the original comments, from six pages to one page of 19 skills. The skills, such as "consistently refers to know key points into the big picture," were sorted into four categories: customizing to the audience, telling the story, displaying key information, and delivering the presentation. The individual skills will be discussed in the analysis and results sections.

To date, the Presentation Scoring System has been used in Georgia Tech Capstone Design instruction for five semesters. The System and supporting instructional materials have been presented at ASEE[3], the Process Education Conference[13], the INFORMS (Institute for Operations Research & Management Science) Annual Conference[14], and the Capstone Design Conference[4]. Several papers are in preparation for publication. To date the materials have been requested by 180 academics who are considering them for use in their settings.

In order to ensure that these users will derive the maximum potential benefit from the rubric, we need to understand more about the reliability of the rubric. Can it be used consistently by different raters viewing the same presentation? Do these different raters give the same feedback for the same presentation? Does it matter whether these raters have been trained in using the rubric or not? In order to answer these questions, we have applied the rubric in a variety of settings with a variety of participants who have received varying amounts of training before applying the rubric to a specific presentation. These include: (1) two researchers and four teaching assistants rating videotaped presentations from capstone design, (2) two researchers and a group of 9 ASEE workshop attendees rating two different videotaped presentations, and (3) one researcher and his students rating 20 peer presentations in a class of 78 students. From the data collected in these settings we calculated the percent frequency of exact matches and within 1-point ratings for each skill represented on the rubric. We have performed these calculations to compare results among the two researchers, compare all raters on an equal footing (pairwise comparison), and compare other users to the researchers who have developed the rubric (reference criterion comparison). The details of this process are explained in the methods section.

Literature Review

The purpose of our literature review is to establish whether others have performed similar evaluations of rubrics and, from that, to identify the commonalities with our methodology. We also seek to establish a metric by which to indicate representative ranges for what constitutes reasonable and/or high levels of agreement between raters. In order to establish historical precedence, we refer to L.L. Thurstone's *law of comparative judgement* from *Psychological Review* (1994 originally published in 1927)[15] for the validity and purpose of pairwise comparisons, and to Jacob Cohen's *Statistical Power Analysis* (1988)[16] for the validity and

purpose of Pearson product moment correlational values, then review six articles directly addressing "rubric reliability"[17-22]. In selecting these articles, we have given primacy to engineering articles, or articles that appeared in engineering journals, as we are working in an engineering setting.

Overwhelmingly, among the six articles, the methodologies used are consistent with our own. These studies use similar measurement scales to determine reliability, and the rubrics were constructed using similar Likert scoring scales with between 3 and 5 points (or levels of performance) used to rate each question (ours is a five point scale). The studies overwhelmingly supported the general method of using experts to establish preferred scoring ranges and comparing those experts' scores with the scores of relative novices, generally students. Further, both Reddy and Andrade[17] and Stellmack et.al.[18] performed aggregate literature reviews and averaged their results across studies of 20 or more different articles. They come to similar conclusions with articles that also evaluated reliability across or among raters, or inter-rater reliability. They also concur with "the potential for rubrics to identify the need for improvements in courses and programmes" [17] and that rubrics have "utility as an instructional tool" [18]. Combining an evaluation of all of these sources, we are able to come to a general consensus as to what are acceptable ranges for percent agreement measurements and reliability measurements for rubrics.

To compare the frequency (calculated as a percent) by which raters come to agreement within a scoring range, the *law of comparable judgment* is necessary. The *law of comparative judgment* "applies fundamentally to the judgments of a *single observer* who compares a series of stimuli by the method of a paired comparison"[15]. Thurstone uses the law to compare five cases of differing assumptions in order to establish the frequency of observations. The same law is also used in order to limit the number of assumptions available to the reviewer, and in order to establish that at least one assumption of the "correlation between discriminal deviations"[15] can be established throughout a stimulus series. It is therefore used to assume that when "a *group* of observers perceives a specimen […], the distribution of excellence that they read into the specimen is normal"[15]. These percent agreements establish the normal range of a group of scorers' scoring variance. Further, as Stemler states, one of the greatest advantages to using percent-agreement measurements is their "strong intuitive appeal" (qtd. in Oakleaf[19]).

In our literature review, most of the sources used percent agreement to establish the normal range of reliability among raters or inter-rater reliability. So, for instance, Mott found the general percent for agreement between raters with exact matches averaged around 70% for most criteria, while at within ±1 they averaged 99%[20]. Newell et al. also found a high percent agreement for raters within ±1 at 97%[21]. However, though Stellmack et al. indicated a standard of difference for inter-rater agreement within a ±1 point variance at 90%[18], they also found inter-rater reliability for exact matches at only 37%[18]. Similarly, while Davis noted percent agreement within ±1 at between 85-100%[22], he too found lower agreement for exact matches, at between 20-60%[22]. The other articles we reviewed did not delineate between scoring ranges, and simply found general percent agreement ranges for scorers. Reddy, for instance, found the average agreement for many studies is around 75%[17] and Stemler suggests that for any indices with a limited point variance, "it would be surprising to find agreement lower than 90%" (qtd. in Oakleaf)[19]. That is, the smaller the number of points (or levels of performance) available to raters, the more one would

expect raters to have a high percentage of agreement. Our score range is a five point scale; for our case, based on the literature, a high percent of agreement for within ±1 is at 90% or above, with acceptable agreement at above 70%. For exact matches, acceptable levels of agreement should be at 30-50%, while agreement should be above 50% for high reliability.

In order to quantify how reliable the raters are, we also use Pearson product moment correlation values where applicable. According to Jacob Cohen, former professor of psychology at New York University and creator of Cohen's kappa, in the social sciences correlation coefficients of a product moment often seek "whether there is *any* (linear) relationship between two variables, and this translates into the null hypothesis"[16]. Because social scientists are often attempting to establish a bias in a "soft" direction, "in many, perhaps most, of the areas of behavioral science, they turn out to be so small!"[16]. Cohen goes on to define that, in general, for the social sciences, a small effect size correlation gives *r* values between 0.1 and 0.3, a medium effect size correlation yields *r* values between 0.3 and 0.5, and a large effect size correlation yields *r* values between 0.5 and 1.0[16]. These correlations help establish a standard range minimum at which the judgments of raters might be considered reliable.

In our literature review, only half of the sources used or evaluated Pearson product moment correlational values. Of those, most indicated that a general agreement of 0.5 and above was, as Mott says, an "acceptable inter-rater reliability"[20] range for rubric correlation coefficients. That said, most of the authors sought as high a correlation coefficient as possible. Johnson, for instance, in seeking to develop better rubrics, began with *r* values at or between 0.3 and 0.6 and finished with *r* values averaging 0.8[23]. Kemppainen only accepted 0.7 to 0.8[24] for high correlations, while Mott indicates 0.7 and above as the threshold for higher than "acceptable" for analytical rubrics[20]. So, within the variability of measurements that social scientists study, it seems fairly well established that acceptable *r* values begin at around 0.5 and high correlations for interrater reliability of rubrics as measurement tools yields *r* values that begin at roughly 0.7 and above.

Methods

Data Collection

To analyze the inter-rater reliability of the rubric, we have collected scores from raters in three different contexts. Approval was granted by the Institutional Review Board for this research, and the permission of each presenter and each rater to use their work was obtained prior to each rating session. Each of these contexts is representative of a common setting where the rubric might be employed.

Setting 1 – **"TA Session"** – rubric developers working with teaching assistants
In this setting, the two researchers worked with four teaching assistants who were employed in the Georgia Tech Industrial and Systems Engineering Workforce Communication Lab serving capstone design students. The Communication Lab has five fully-equipped presentation stations and about 1,000 student visits per semester[25]. We rated a total of 21videotaped presentations from capstone design, each approximately 6 minutes in length. During the early stages of the

session the group stopped to share and discuss ratings after every few presentations with the intent of convergence towards more reliable scoring.

This setting is similar to one where an instructional faculty member or team directs a large number of individual sections taught by teaching assistants, or where a large course is taught and a number of teaching assistants do much of the grading.

Setting 2 – **"ASEE Session"** – rubric developers and workshop attendees
In this setting, the two researchers were facilitating a workshop at the ASEE Annual Conference with a group of nine attendees from various institutions. At the end of the session we rated a set of two videotaped presentations from capstone design, each approximately 6 minutes in length. Results were shared and discussed after each of the two presentations.

This setting is similar to situations where diverse groups of faculty might come together to view (and rate) student research presentations for a department. It might also represent a situation where a department or college decides to utilize a common rubric across many courses and the instructors responsible for those courses are introduced to the rubric as a group before using it on their own.

Setting 3 – **"In-class Session"** – one developer and students in the course
In this setting, one of the rubric developers employed a slightly shortened version of the rubric in their course to evaluate a series of two-minute multimedia presentations developed by teams of four students each. The instructor and selected student peers rated each presentation. The peer raters consisted of approximately 14 other students in the course for each presentation, with different students selected for each presentation. There were 20 total presentations in a class of 78 total students.

This setting represents a case where a faculty member adopts a rubric for their course and asks students to do peer reviews with that rubric "on the fly" without specific training on the use of the rubric.

Analysis

For each of the 19 skills on the rubric, we analyzed the data collected from the three sessions above using two types of calculated measures for reliability: pairwise matching and criterion-referenced matching.

In pairwise matching, each combination of rater scores for a skill are compared to each other. In this case the rubric developers are treated no differently than the other raters. The frequency of matches was then calculated for two cases:

- Pairwise 1-point score consistency (in %) where raters matched within ±1 point of each other on the five point scale for the rubric.
- Pairwise exact score consistency (in %) where raters scores matched exactly.

In Criterion referenced matching, a "true" score is determined based on the scores given by the rubric developers. If the two rubric developers did not give the same score, then the "true" score

was determined from the average of the two scores. For these averages, rounding of any decimal results was intentionally biased to round towards the score provided by the rubric developer with more experience rating presentations. This rounding is necessary because only discrete values are allowed for rubric ratings and, thus, comparisons cannot be made to decimal results. Once this "true" score was found, each of the other rater's scores was compared to this "true" score. The frequency of matches was then calculated for two cases:

- Criterion referenced 1-point score consistency (in %) where rater consistency is compared to within ±1 point from the "true" score.
- Criterion referenced exact score consistency (in %) where rater consistency is compared for exact matches with the "true" score.

In addition, a limited amount of data was suitable for the calculation of Pearson correlation coefficients (*r*). Values of *r* were calculated to compare ratings between the two rubric developers for all presentations where both provided ratings (23 presentations), and to compare ratings between the rubric developer "true" score and the teaching assistant ratings (21 presentations).

It should be noted that not all raters provided ratings for all skills on the rubric in all instances. This is because some skills, such as sensitivity to time, were not able to be evaluated for certain presentations, and there was no explicit expectation that raters assign values for each of the 19 skills on the rubric during these relatively short presentations.

Results and Discussion

In general, the skills comprising the rubric were found to be of moderate to high reliability when used by different raters. Our analysis of ratings for each of the 19 skills shows, in general, that the reliability of the overall rubric is acceptable. However, there is some variation in the reliability of each skill. In particular, we see

a. High reliability for 7 of the skills,
b. Good reliability for an additional 4 skills,
c. Reasonable but lower reliabilty for 8 skills.

Specific results for each skill and in each setting are displayed below.

Rubric Developer Inter-rater Reliability

Table 1 displays the results of the pairwise matching and Pearson correlation results for ratings given by the two rubric developers. In the table, "N" represents the number of times the skill was compared. Pairwise matching results are displayed as percents for frequency of agreement by "exact" match and "within 1 point" matches. Pearson correlation coefficients are also presented along with the calculated statistical significance of the correlation values.

As can be seen from the results, exact matches average around 40%, while matches within 1 point on the 5-point scale average around 90%. Only one skill (key points) shows within 1-point agreement below 80%, but this skill was rated only 12 times. Several other skills (audience connection, taking questions, sensitivity to time, and focused content) were rated only a few

times, and so the results in those cases should be considered only as indications of potential reliability.

Regarding the results for Pearson correlations, more data is needed to obtain strong indications of statistical significance for four skills, as they have four or fewer data points. Eleven out of the 15 skills with between 12 and 23 data points show statisically significant correlations at the .05 level or better. Further, when rounding to one significant digit, nine of those 15 skills show good reliability with *r* values at or above 0.5, and an additional four skills demonstrate an acceptable but marginal reliability at 0.3 or above. The remaining two skills (layout and design, and key points) need additional data as they have only 12 comparisons each.

Table 1 – rubric developer inter-rater reliability

Pairwise Matching Frequency	TA session % agreement				Pearson		
	N	EXACT	WITHIN_1		N	r	signif
Audience Connection	3	33%	100%		3	N/A	
Appropriate language	15	53%	93%		15	0.428	0.056
Relevant details	18	44%	100%		18	0.697	0.001
Taking questions	4	25%	100%		4	N/A	
Sequencing	20	45%	95%		20	0.463	0.02
Key points	12	31%	81%		12	0.16	0.309
Context	16	31%	81%		16	0.518	0.02
Sensitivity to time	1	100%	100%		1	N/A	
Layout and design	12	33%	83%		12	-0.038	0.453
Focused content	3	33%	100%		3	-0.5	0.333
Amount of text	23	57%	96%		23	0.738	0
Appropriate graphics	22	27%	86%		22	0.293	0.093
Engaging Graphics	22	32%	95%		22	0.363	0.048
First/last impression	22	41%	95%		22	0.539	0.005
Flow	19	26%	89%		19	0.588	0.004
Elaboration	20	45%	90%		20	0.532	0.008
Stature	23	17%	78%		23	0.359	0.046
Vocal Quality	23	26%	83%		23	0.491	0.009
Personal presence	22	36%	95%		22	0.566	0.003

TA Session inter-rater reliability

As shown in Table 2, the session with the teaching assistants demonstrates inter-rater reliabilities very close to the levels attained by the rubric developers shown in Table 1. This is true for both the pairwise comparisons and the criterion-referenced matching. With the exception of the "sensitivity to time" skill, every skill demonstrates very reasonable inter-rater reliability (at 80% and above within 1 point) for this rater group, and approximately half of the skills show quite strong inter-rater reliability, at over 90% matching within 1 point on our 5-point scale. The anomoly for "sensitivity to time" makes sense in this setting where the presentations were

videotapted and little, if any, context was provided to indicate how this skill might be rated. As a result, that skill was very rarely rated.

Table 2 – TA session inter-rater reliability

TA session	Pairwise				Criterion Referenced		
	N	EXACT	WITHIN_1		N	EXACT	WITHIN_1
Audience Connection	6	33%	100%		3	33%	100%
Appropriate language	67	46%	91%		27	48%	93%
Relevant details	38	66%	97%		10	70%	100%
Taking questions	22	41%	95%		7	14%	86%
Sequencing	139	42%	90%		46	41%	87%
Key points	97	28%	84%		39	26%	85%
Context	76	39%	87%		28	25%	86%
Sensitivity to time	4	25%	50%		1	0%	0%
Layout and design	110	33%	87%		41	34%	88%
Focused content	70	61%	96%		36	58%	94%
Amount of text	168	43%	93%		50	46%	96%
Appropriate graphics	126	46%	94%		42	48%	95%
Engaging Graphics	140	35%	87%		45	38%	82%
First/last impression	124	54%	96%		41	68%	95%
Flow	171	46%	91%		55	35%	89%
Elaboration	183	46%	89%		57	39%	95%
Stature	157	32%	82%		49	37%	88%
Vocal Quality	161	42%	90%		48	56%	92%
Personal presence	172	44%	91%		54	48%	89%

ASEE Session inter-rater reliability

As shown in Table 3, the session with the ASEE workshop participants demonstrates substantially lower inter-rater reliabilities than the TA session or that attained by the rubric developers. This is true for both the pairwise comparisons and the reference criterion matching. In this case, the "appropriate graphics" skill stands out as having rather low reliability among this group, and a number of other skills match within 1point less than 70% of the time. However, it should be noted that only two presentations were rated by this group, and most of the participants attempted to rate all 19 skills within a space of six to ten minutes after having been introduced to the rubric only an hour before they began to use it. Again, due to the videotaped presentation mode, "sensitivity to time" was very rarely rated.

Table 3 – ASEE session inter-rater reliability

ASEE session	Pairwise				Criterion Referenced		
	N	EXACT	WITHIN_1		N	EXACT	WITHIN_1
Audience Connection	73	32%	81%		16	38%	75%
Appropriate language	100	43%	94%		18	56%	94%
Relevant details	83	27%	72%		16	6%	44%
Taking questions	56	29%	79%		9	33%	100%
Sequencing	81	22%	63%		16	19%	69%
Key points	90	21%	61%		18	17%	78%
Context	73	22%	64%		16	13%	63%
Sensitivity to time	12	42%	100%		0	N/A	N/A
Layout and design	100	34%	73%		18	56%	83%
Focused content	81	22%	64%		17	18%	65%
Amount of text	110	24%	67%		18	17%	61%
Appropriate graphics	100	20%	56%		17	24%	41%
Engaging Graphics	100	28%	75%		17	6%	53%
First/last impression	100	25%	68%		17	12%	29%
Flow	100	27%	77%		9	22%	67%
Elaboration	100	30%	76%		9	56%	89%
Stature	110	19%	65%		18	28%	78%
Vocal Quality	110	34%	80%		18	56%	89%
Personal presence	110	31%	80%		18	17%	83%

In-class session inter-rater reliability

As shown in Table 4, the session in one of the rubric researcher's classes displays relatively high reliability. Three of the 19 skills on the full version of the rubric were not used for in-class session because they were minimally relevant to the presentation situation. Of the remaining 16 skills, agreement among the raters was generally high.

For pairwise matching, seven of the 16 skills showed agreement at 90% or higher, and 15 at 80% or higher, with only one skill (personal presence) below the 80% mark.

For the reference-criterion matching, one skill indicated a very low level of agreement (key points). This skill was rated particularly low by the instructor on a consistent basis, indicating that expectations for performance of this skill by the instructor may not have been clearly communicated to the students. Three other skills (audience connection, amount of text, and first/last impression) displayed significantly lower reference criterion matching frequencies than the other skills (below 70%) despite relative high pairwise matching agreement. Again, in these cases the instructor ratings were lower than typical student ratings, indicating either (1) a certain leniency on the part of the students or, (2) again, expectations for performance of this skill may not have been clearly communicated to the students. The consistent difference between the instructor and the students is important to note for this session because, of the three settings studied here, only the in-class setting carried the weight of actual grades attached to the ratings.

For the in-class session, peer ratings were used as part of the grade received by the presenters. The other session ratings had no consequence for any student grades; they were conducted for the purpose of learning how to use the rubric rather than assigning grades.

Table 4 – In-class session inter-rater reliability

In-class session	Pairwise				Criterion Referenced		
	N	EXACT	WITHIN_1		N	EXACT	WITHIN_1
Audience Connection	1780	37%	84%		250	25%	68%
Appropriate language	1794	44%	93%		264	44%	88%
Relevant details	1767	45%	96%		249	43%	98%
Taking questions	0	N/A	N/A		0	N/A	N/A
Sequencing	1742	44%	94%		225	47%	93%
Key points	1754	39%	87%		237	10%	36%
Context	1767	43%	96%		250	41%	94%
Sensitivity to time	1649	47%	93%		215	37%	95%
Layout and design	0	N/A	N/A		0	N/A	N/A
Focused content	1718	41%	94%		214	38%	89%
Amount of text	1752	35%	83%		261	27%	65%
Appropriate graphics	1753	42%	91%		261	38%	92%
Engaging Graphics	1752	38%	87%		261	32%	79%
First/last impression	1712	33%	81%		208	23%	63%
Flow	1727	39%	88%		224	37%	88%
Elaboration	1689	41%	89%		210	48%	93%
Stature	0	N/A	N/A		0	N/A	N/A
Vocal Quality	1643	35%	80%		196	36%	82%
Personal presence	1096	31%	77%		25	36%	88%

Overall inter-rater reliability

Using the results presented above, an overall reliability for each skill was determined by combining the results from the three settings. The data above includes three (3) settings for two (2) types of criteria (reference criterion and pairwise matching) and for two (2) levels of agreement (exact and within 1 point). This gives 12 individual reliability measures (3 x 2 x 2) for each skill. Among these individual reliability measures, an overall inter-rater reliability index was assigned for each of the 12 individual reliability measures for each skill. This index is used to judge whether a skill is highly reliable, moderately reliable, or marginally reliable. Indices were assigned for each individual reliability measure as shown in Table 5.

Table 5 – inter-rater reliability index assignments for each reliability data point

Index	Criteria
-1	marginally reliable (below 30% exact or below 70% within 1)
0	moderately reliable (30-50% exact or 70-90% within 1)
1	Highly reliable (above 50% exact or above 90% within 1)

Once indices were assigned, the number of times a skill was measured “high”, “moderate”, or “marginal” was added up. This results in a scale of +12 to -12, where +12 indicates the skill always displayed a high level of reliability, and -12 indicates the skill always displayed a low level of reliability. The results of this indexing process are shown in Table 6.

As shown in the table, none of the skills were highly reliable in every setting and by every measure. However, on average, moderate to high reliability was achieved for most of the skills (0 or higher). A few skills (-3 and below) show marginal to moderate reliablity and should be used with caution. For example, using the skill “key points” may prove unreliable with a group of raters that have not had time to come to consensus about how to rate this skill.

Table 6 – Overall inter-rater reliability

SKILL	Overall Reliability Index
Appropriate language	6
Vocal Quality	4
Relevant details	3
Elaboration	2
Sensitivity to time	1
Layout and design	1
Focused content	1
Audience Connection	0
Taking questions	0
Appropriate graphics	0
Personal presence	0
Sequencing	-1
First/last impression	-2
Flow	-2
Context	-3
Engaging Graphics	-3
Stature	-3
Amount of text	-4
Key points	-6

Conclusions and Future Work

Our results demonstrate the rubric generates acceptable reliability as a rule, with approximately one third of the items generating high reliability on a rather consistent basis. However, a few items need to be treated with caution when using multiple raters. This reliability is displayed for both pairwise matching and reference criterion matching methods.

Based on these results, schools who wish to adopt the rubric will be able to prioritize which pieces of the rubric to use in their courses. Teachers and communication professionals can make decisions based on the most important skills for their context and the most reliable skills of the

rubric. Schools will also be able to use the results of this study to enhance the reliability between different raters and by identifying the most reliable parts of the rubric.

During the next stages of development for our rubric, we intend to focus on *why* we are getting a few lower reliability results. A focus group with users will be employed to collect specific data, and other user feedback will be collected to triangulate our data. Thus, the inter-rater reliability results identified in this study will be used in conjunction with both formal and informal user feedback to modify the existing rubric. These modifications will have two aims – enhance ease of use and understanding of the rubric, and enhance reliability among different raters using the rubric. In addition, now that we have established that our scoring presentation system is both valid and reasonably reliable, we can also begin to look at *how* different variables (such as prior experience with presentations, taking capstone design, GPA, or amount of training received) affect the level of performance in giving presentations as rated using our rubric.

With these modifications completed, other schools using the modified rubric will be able to enhance the reliability among their scorers, and apply the rubric with even greater confidence that the feedback students receive will enhance their performance in oral presentations. Finally, we will make available the modified rubric and an outline of any suggested training to improve reliable scoring among raters.

Bibliography

1. ABET. *Criteria for Accrediting Engineering Programs*. 2000, 2004.
2. J.S. Norback and T.T. Utschig, *A preliminary Scoring System for Engineering Presentations, built on Executive Feedback*, INFORMS 2009 Annual Meeting, San Diego, CA, October 11-14, 2009
3. S. Howe, K. Caves, C. Kleiner, G. Livesay, J. Norback, R. Rogge, C. Turner, and T. Utschig, *Nifty Ideas and Surprising Flops in Capstone Design Education*, International Journal of Engineering Education, Vol. 27, No. 6, pp. 1-12, 2011.
4. T.T. Utschig and J. Norback, *Refinement and Initial Testing of an Engineering Student Presentation Scoring System*, American Society for Engineering Education Conference, Louisville, KY, June 20-23, 2010.
5. Eccles, J.S. 2005. Subjective task value and the Eccles et al. model of achievement-related choices. In *Handbook of competence and motivation*, eds. A. J. Elliot and C. S. Dweck, 105-21. New York: Guilford Press.
6. Velez, J. 2008. Instructor communication behaviors and classroom climate: Exploring relationships with student self-efficacy and task value motivation. http://etd.ohiolink.edu/send-pdf.cgi/Velez%20Jonathan%20J.pdf?acc_num=osu1211151901 (last accessed, 19 May 2010).
7. Pintrich, P. R. 1994. Continuities and discontinuities: Future directions for research in educational psychology. *Educational Psychologist* 29: 37-148.
8. Zachry, M. 2000. Communicative practices in organizations: Genre-based research in professional communication. *Business Communication Quarterly* 63 (4): 95-101.
9. Paretti, M. 2008a. Teaching communication in capstone design: The role of the instructor in situated learning. *Journal of Engineering Education*: 491-503.
10. Dias, P., A. Freedman, P. Medway, and A. Pare. 1999. *Worlds apart: Acting and writing in academic and workplace contexts*. Mahwah, NJ: Lawrence Erlbaum.
11. Katz, S.M. 1998a. Part I—Learning writing in organizations: What newcomers learn about writing on the job. *IEEE Transactions on Professional Communication* 41 (2): 107–14. New York, NY: IEEE Professional Society.
12. Katz, S.M. 1998b. Part II—How newcomers learn to write: Resources for guiding newcomers. *IEEE Transactions on Professional Communication* 41 (3): 165–73. New York, NY: IEEE Professional Society.

13. J. Norback and T.T. Utschig, *Building a Stakeholder-based Rubric to Enhance Student Communication Skills*, International Journal of Process Education, Vol. 2, no. 1, June 2010.
14. J. Norback and T.T. Utschig, *Communication Instructional Tools from Georgia Tech and Texas Tech*, INFORMS 2010 Annual Meeting, Austin, TX, November 7-10, 2010.
15. Thurstone, L. (1994). A Law of Comparative Judgement. *Psychological Review , 101* (2), 266-270.
16. Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences* (2nd ed.). New Jersey: Lawrence Erlbaum Associates.
17. Reddy, Y. M., & Andrade, H. (2010). A Review of Rubric Use in Higher Education. *Assessment & Evaluation in Higher Education , 35* (4), 435-448.
18. Stellmack, M. A., Konheim-Kalkstein, Y. L., Manor, J. E., Massey, A. R., & Schmitz, J. A. (2009). An Assessment of Reliability and Validity of a Rubric for Grading APA-Style Introductions. *Teaching of Psychology , 36*, 102-107.
19. Oakleaf, M. (2009). Using Rubrics to Assess Information Literacy: An Examination of Methodology and Interrater Reliability. *Journal of the American Society for Information, Science, and Technology , 60* (5), 969-983.
20. Mott, M. S., Etsler, C., & Drumgold, D. (Spring 2003). Applying and Analytic Writing Rubric to Children's Hypermedia "Narratives". *Early Childhood Research & Practice: An Internet Jounral on the Development, Care, and Education of Young Children* , 1-18.
21. Newell, J. A., Dahm, K. D., & Newell, H. L. (2002). Rubric Development and Inter-rater Reliability Issues in Assessing Learning Outcomes. *ASEE*.
22. Davis, D. (2011). Establishing Inter-rater Agreement for TIDEE's Teamwork and Professional Development Assessments. *ASEE*.
23. Johnson, C. S. (2006). The Analytic Assessment of Online Portfolios in Undergraduate Technical Communication: A Model. *Journal of Engineering Education , 95* (4), 279-287.
24. Kemppainen, A., Amato-Henderson, S., & Hein, G. (2010). Work in Progres: Refining a Technical Communication Rubric for First-Year Engineering Instructors. *Frontiers in Education Conference (FIE)* , T2G1-T2G3.
25. Norback, J.S., and J.R. Hardin. "Integrating Workplace Communication into Senior Design," IEEE Transactions on Professional Communication 48 (2005).