Final Results of Reliability Testing for the Norback-Utschig Presentation Scoring System and Implications for Instruction

Dr. Judith Shaul Norback, Georgia Institute of Technology
Dr. Tristan T. Utschig, Georgia Institute of Technology

Dr. Tristan T. Utschig is a Senior Academic Professional in the Center for the Enhancement of Teaching and Learning and is Assistant Director for the Scholarship and Assessment of Teaching and Learning at the Georgia Institute of Technology. Formerly, he was a tenured Associate Professor of Engineering Physics at Lewis-Clark State College. Dr. Utschig consults with faculty across the university about bringing scholarly teaching and learning innovations into their classroom and assessing their impact. He 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. Dr. Utschig completed his PhD in Nuclear Engineering at the University of Wisconsin–Madison.

Mr. Anthony Joseph Bonifonte, Georgia Institute of Technology

Anthony Joseph Bonifonte is currently in his 3rd year of Georgia Tech's PhD program in Operations Research in the Industrial and Systems Engineering Department. He attended Oberlin College as an undergraduate, majoring in math and biology. He has served as teaching assistant five times for math and industrial engineering courses. 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 thesis research involves mathematical models and decision making in cardiology.

Gloria J Ross, Georgia Institute of Technology

Gloria Ross is currently a PhD candidate in History and Sociology of Science and Technology at Georgia Tech. Her research focuses on the spatial and demographic factors that shape urban food distribution systems. She currently works as a graduate research assistant in Georgia Tech's Center for the Enhancement of Teaching and Learning (CETL) where she assists with assessment and data analysis for ongoing CETL projects.

©American Society for Engineering Education, 2014

Final Results of Reliability Testing for the Norback-Utschig Presentation Scoring System and Implications for Instruction

Abstract

In this paper we report the results of our final work completed to improve the reliability and usability of the Norback-Utschig Presentation Scoring System developed at Georgia Tech, and based on executive input. Our approach was the modified Delphi method, a multi-stage feedback process used to generate consensus among diverse stakeholders. The method was used to collect data for seven of the 13 presentation skills not yet having high reliability between raters: "initial and final impressions," "logical flow," "key points," "layout and design," "graphics," "vocal quality," and "personal presence." Data was collected from a variety of individual stakeholders, and modifications were made to the scoring system. For example, "initial connection" was renamed to "first impression." The definition of the following skills were clarified to make them easier to understand: "logical flow," "key points," "layout and design," "graphics," "vocal quality" and "personal presence." Once the skills were modified, the new scoring system was tested for reliability in three settings—industrial engineering teaching assistants and the two developers of the scoring system, biomedical engineering teaching assistants and one developer, and a class of nuclear engineering seniors. Results indicate that the reliability between raters of all skills tested improved at a significant level. The revised skills now have good to high reliability. Implications for instruction will be discussed.

Introduction

The Norback-Utschig Presentation Scoring System for Engineers is based on interviews with 72 executives, with engineering degrees, who work for a variety of companies employing engineers. Faculty input has been gathered over the years as well. The skills identified were those essential for a "stellar presentation." In this paper we report on the results of our final steps taken at Georgia Tech to revise and test the Norback-Utschig Presentation Scoring System to improve its usability and reliability. We expand upon our previously published work, which is briefly reviewed below for context. We also share implications for instruction: teaching tips for helping students perform better on each skill[1]. Examples include "key points: central message clear throughout by linking details to big picture," and "personal presence: effectively combines energy, eye contact, and movement."

One year ago we reported on our work using a modified Delphi method to revise a 19-skill presentation scoring system. The method is a multi-stage feedback process used to generate consensus among diverse stakeholders[2]. In the earlier paper we outlined lessons learned from discussing the use of the scoring system with users. We also described how we, first, summarized feedback we had collected from a small alumni-funded study, second, distributed the summary to stakeholders for their review and, third, modified the scoring system according to the newest feedback. The result was a 13-skill presentation scoring system with enhanced usability and clarity. Figure 1 summarizes changes made to the scoring system after round 1 of the Delphi update. For example, "flow" was added to "vocal quality;" both "engaging graphics" and "appropriate graphics" were combined into "graphics," "first/last impression" and "audience

connection" were combined into "initial connection." "Sequencing" was redefined as "logical flow." Additionally, Appendix A shows details for how the definitions for skills were updated as a result of these changes.

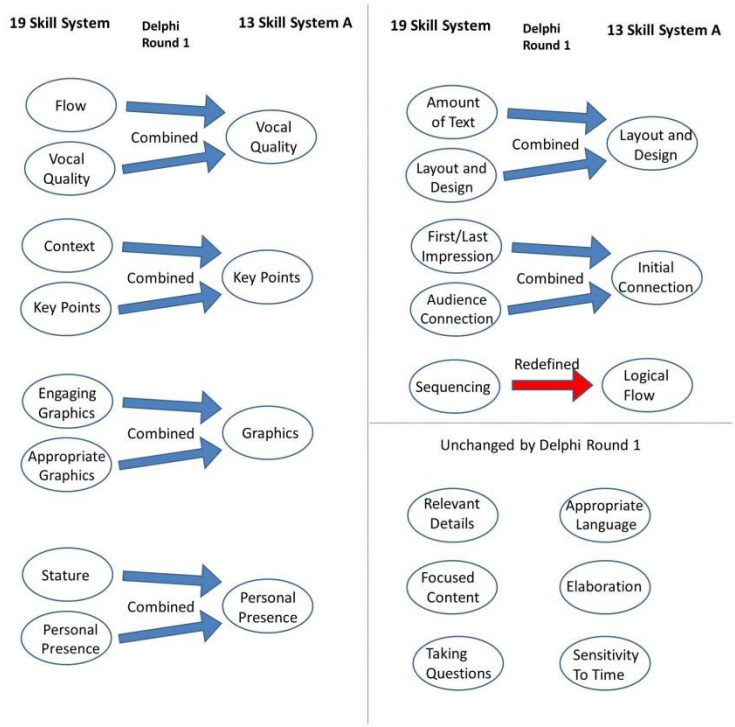

Figure 1 – Delphi Round 1 changes

This second round of feedback and modifications completed our modified Delphi method procedure. To complete this second round of feedback and modifications we first collected and summarized a second round of feedback from a variety of individual stakeholders, second, modified the scoring system once more based on the newest feedback and third, tested the scoring system for inter-rater reliability in several different settings. The changes made in Delphi round 2 include renaming "initial connection" to "first impression" and changing the definitions of about half of the skills to address user feedback on clarity and ease of understanding those skills. Additionally, two skills, "sensitivity to time" and "taking questions," were moved to the end of the survey.

Below we describe literature related to our work on the Norback-Utschig Presentation Scoring System, within engineering and then in the expanded context of STEM. Then we discuss information available on engineering written communication scoring systems. Next, we focus on the two methods used: the modified Delphi system and the testing of the scoring system for inter-rater reliability in several different engineering settings. We describe the result: a highly usable and reliable set of 11 skills divided into four categories (customizing to the audience, telling the story, displaying key information, and delivering the presentation) plus two additional skills which can be used when appropriate at the end of a presentation. The two skills are "sensitivity to time" and "taking questions."

Finally, we share a few tips that have worked well in coaching capstone design students to improve each of the 13 skills. One example consists of teaching students to be "definite!" in order to enhance the first impression they make with their audience. Another example is having students practice, throughout their presentation, using reminders of how details relate to the big picture.

Literature Review

Based on our literature review, the publication of inter-rater reliability studies or other development work for research-based rubrics or scoring systems is not common. As noted above, we focused first on scoring systems used for evaluating both oral communication within and outside of engineering settings and then we turned to scoring systems for written engineering communication. From what we could see, other work may be going on in engineering, but it is not frequently documented in publications. Here, we briefly review what we have found to date.

Earlier we found that the most comprehensive work to date in developing an oral presentation rubric for engineering settings was done at Iowa State, and is now in use at several other colleges and universities[3,4]. We also found the TIDEE Consortium had studied inter-rater reliability of student communication in team settings for engineering design coursework[5]. Many other oral presentation rubrics are used elsewhere, but the theoretical or research-based underpinnings for these rubrics are unclear. Other than finding Thambyah's discussion of how to create rubrics for engineering outcomes, we reviewed similar studies of reliability in communication rubrics—expanding from the engineering setting to STEM fields[6].

These include

- Mott's work on assessing scientific expression using media[7],
- LaBanca's eight-item rubric primarily targeting high school science teachers[8],
- Barney's study on the effects of student self-assessment of their learning outcomes when using an oral presentation rubric[9],
- Anderson and Anderson's description of elements of oral communication for students in the workplace, that should be modeled by instructors in their lectures[10],
- Battacharyya's use of a research-based approach to identify four key traits that industry looks for in technical presentations: technical competence, effective delivery skills, IT competency, and cultural awareness[11], and
- Morton and Rosse's research focusing on identifying characteristics related to the use of personal pronouns in defining persuasive oral presentations in engineering[12].

Although two of these studies relate to workplace communication specifically for engineers, it does not appear that any of this work directly links criteria for scoring systems to workplace preferences or traditional rhetorical approaches. To our knowledge, then, inter-rater reliability studies for instruments measuring the quality of engineering oral communication are very limited.

Because of the dearth of literature describing the development and testing of systems to rate oral communication, we have begun to review written communication as well. We identified a number of relevant articles. For example

- Covill examines writing rubrics specifically in the context of self-assessment. She includes a summary of existing empirical research on the effect of rubric use on writing quality, but claims most rubric studies are based on middle school students[13].
- Andrade and Du found that undergraduate students perceived the writing rubric as a useful tool that clarified teacher expectations, structured their own approach to each assignment, and provided a reference to review and edit their work[14].
- In one study involving student use of written assignment rubrics, Rawson et al. found that students who used the rubric were able to acquire and use new field-specific terminology in appropriate ways[15].
- In a study comparing writing rubrics, Morozov concluded that students viewed the more detailed and extensive rubric more positively than less-extensive rubrics[16]. In this study, an effective rubric model emphasized skills, elaboration of skill, and critical thinking.
- One recent study compared the reliability of two writing rubrics across three different settings and reported moderate reliability for most skills represented in the two rubrics[17].
- Multiple studies address the effect of Calibrated Peer Review (CPR) on student writing[18, 19, 20]. CPR involves the electronic evaluation of student writing by their peers. None of these studies specifically address rubric development or inter-rater reliability. One study indicates significant differences in the effect of feedback received from TAs as opposed to feedback received from peers[18].

Methods

Two major processes are included in this study. The first is the completion of round 2 of the Delphi method to systematically modify the Norback-Utschig Presentation Scoring System to be more usable and reliable. The second is the testing of the scoring system for inter-rater reliability in several different engineering settings.

Final (Round 2) Modifications Using the Delphi Method

To make the final round of modifications to our scoring system, we first developed a process for collecting stakeholder input about the scoring system resulting from the first round of the Delphi method. Our stakeholders included faculty, graduate and undergraduate students (mostly teaching assistants or TAs), and executives. Our academic participants represented industrial engineering, biomedical engineering, nuclear engineering, and a Center for the Enhancement of Teaching and Learning. In industrial engineering, our participants were two female faculty members, three undergraduate TAs (one female and two male), and four graduate student TAs (one female and three males). From biomedical engineering, one male faculty member and six graduate student TAs (five female, one male) contributed. One male faculty from the Center and one male undergraduate from a class in nuclear engineering rounded out the academic input. Finally, three executives (one female and two male), from a variety of settings, took part, bringing the total number of stakeholders to 21.

Our process of collecting feedback from the stakeholders consisted of two parts. The first part was a quantitative questionnaire for users to rate, first, the ease of use and, second, the clarity of wording on a three-point scale for each skill. The second part was qualitative and involved interviews, focus groups, and email communication. Users explained their ratings from the

quantitative section of the survey and provided other comments about their interpretation and use of the thirteen skills when scoring presentations.

Once the stakeholder feedback was collected, it was systematically analyzed to provide a summary that could be used to make modifications to the scoring system. The methods used included summative coding and content analysis. Qualitative comments about the scoring system were first grouped by skill number and reviewed by two graduate research assistants or "coders." The assistants identified common themes and keywords in the comments and tallied each comment under a theme, with some themes having only one comment. To ensure reliability and consistency in the coding analysis, the process was independently replicated by one of the scoring system developers, and the two sets were merged by a fourth person (the other developer of the scoring system) upon discussion with both coders. The themes with the most comments represented several stakeholders who shared common feedback for how to modify the scoring system. For example, for the skill on "taking questions," five stakeholders suggested that the skill definition needed further clarification and specific indicators. Once these themes were verified and finalized, they were used to guide modifications of the scoring system.

Final Inter-Rater Reliability Testing

Data Collection

It should be noted that a number of skills were not modified at all because they were already highly reliable. For these skills, changes to the supplemental instructional materials will be made to reflect the suggestions provided by the scoring system stakeholders.

To analyze the inter-rater reliability of the final version of the scoring system, we have collected scores from raters in three different contexts. The Institutional Review Board approved this research project. Prior to each rating session, permission was obtained from each presenter and rater to use their work in this research. Each of these contexts is representative of a common setting where the rubric might be employed.

Setting 1 – **"Industrial Engineering Session"** – In this capstone design context students were preparing and presenting several presentations to clients and to academic faculty. A mixture of videotaped interim and final presentations was used for this session, where 20 presentations were rated by seven TAs who had moderate familiarity with the scoring system (two hours of training and about two weeks working with students) and one of the developers of the scoring system.

Setting 2 – **"Biomedical Engineering Session"** – In this capstone design setting, students were presenting progress reports on their work to classmates and their instructors at a point about seven weeks into a 15-week semester. Fifteen videotaped presentations were rated by four TAs who had no experience with the scoring system, and by one of the developers of the scoring system.

Setting 3 – **"Nuclear Engineering Session"** – In this upper-division engineering course populated mostly by seniors, students presented results for their completed team projects at the end of the term. Ten live presentations were rated by 30 student peers who had no prior

experience with the scoring system, and by one of the developers of the scoring system who was an instructor in the course.

Analysis

For each of the 13 skills on the rubric, we analyzed the data collected from the two sessions above using two types of calculated measures for reliability: pairwise matching and criterion-referenced matching. Pairwise matching consists of comparing each combination of rater scores for a skill to each other. The frequency of matches is then calculated. In criterion-referenced matching, a "true" score is determined based on the scores given by the scoring system developers. Once the "true" score is identified, each of the other raters' scores was compared to the "true" score. More detail appears below.

In pairwise matching, each combination of rater scores for a skill is 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 percent) where raters matched within ±1 point of each other on the five-point scale for the rubric.
- Pairwise exact score consistency (in percent) where raters' scores matched exactly.

In criterion-referenced matching, a "true" score is identified based on the scores of the rubric developers. In this process, 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 percent) where rater consistency is compared to within ±1 point from the "true" score.
- Criterion referenced exact score consistency (in percent) where rater consistency is compared for exact matches with the "true" score.

It should be noted that not all raters provided ratings for all the skills on the rubric in all instances. First, raters were not required to score each skill during the relatively short presentations. Second, the skill of "sensitivity to time" could not be rated since the videos showed one speaker of a group of speakers. Third, the "taking questions" skill was not rated in videos containing no questions from the audience.

Results

Final (Round 2) Modifications Using the Delphi Method

The total number of individual comments collected was 117 (including general comments about the scoring system that did not necessarily pertain to a particular skill). The individual coding results for these comments are shown below. In general, final coding results were collapsed to

represent the smaller number of themes between the two coders. As a result of this coding, the changes made in Delphi round 2 include

- renaming "initial connection" to "first impression"
- changing the definitions of about half of the skills to address user feedback about clarity and ease of understanding of those skills
- moving two skills, "sensitivity to time" and "taking questions," to the end of the survey.

A summary of the results for each skill shown in Table 1. Additionally, Appendix A displays the original and modified wording of the definitions for each skill.

Table 1: Coding of feedback collected from stakeholders and resulting changes to scoring system

Count of themes by skill:			Round 2 Changes to Scoring System
Skill	**# of Themes Coder 1**	**# of Themes Coder 2**	
First Impression	5	9	Renamed
Relevant Details	3	5	None
Appropriate Language	2	7	None
Logical Flow	5	7	Definition clarified
Key Points	3	3	Definition clarified
Layout and Design	4	7	Definition clarified
Focused Content	3	2	None
Graphics	3	6	Definition clarified
Elaboration	1	3	None
Vocal Quality	2	2	Definition clarified
Personal Presence	4	4	Definition clarified
Taking Questions	1	4	None
Sensitivity to Time	2	5	None
General	10	11	
Total	48	75	

Final Inter-Rater Reliability Testing

In general, the skills comprising the rubric were found to be of moderate to high reliability when used by different raters across our three settings. Our analysis of ratings for each of the 11 skills (not counting "sensitivity to time" and "taking questions") 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, for the overall results we see

a. High reliability for 10 of the skills,
b. Moderate reliability for an additional 1 skill.

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

Industrial Engineering Session inter-rater reliability

As shown in Table 2, in this session, for pairwise comparisons, every skill demonstrates very reasonable inter-rater reliability (at 80% and above within 1 point) for this rater group, with the exception of "sensitivity to time." Over half of the skills show quite strong inter-rater reliability, at over 90% matching within 1 point on our 5-point scale. The anomaly for "sensitivity to time" makes sense in this setting where the videotaped presentations were one speaker's part of a group presentation. As a result, that skill was very rarely rated. For the criterion-referenced comparison, every skill except "elaboration" demonstrated very reasonable inter-rater reliability.

Table 2 – Industrial Engineering session inter-rater reliability

Industrial Engineering session	**Pairwise**				**Criterion Referenced**		
	N	EXACT	WITHIN_1		N	EXACT	WITHIN_1
First Impression	490	43%	92%		127	45%	89%
Relevant Details	518	42%	91%		131	40%	89%
Appropriate Language	532	50%	98%		133	41%	98%
Logical Flow	525	37%	90%		132	37%	86%
Key Points	518	37%	89%		131	33%	84%
Layout and Design	532	33%	80%		133	32%	82%
Focused Content	532	41%	86%		133	44%	89%
Graphics	518	35%	84%		131	36%	86%
Elaboration	525	32%	84%		132	20%	78%
Vocal Quality	525	43%	92%		132	39%	91%
Personal Presence	525	44%	92%		132	45%	92%
Taking Questions	137	44%	93%		37	38%	95%
Sensitivity to Time	17	29%	76%				

Biomedical Engineering Session inter-rater reliability

The BME session results, shown in Table 3, demonstrate slightly lower inter-rater reliabilities than the industrial engineering session. This is true for both the pairwise comparisons and the criterion-referenced matching. For pairwise comparisons, eight of the 11 skills demonstrate inter-rater reliability above 80% within 1 point. For the criterion-referenced comparison, the "key points" skill stands out as having rather low reliability, despite being reasonably reliable for the pairwise comparisons.

Table 3 – BME session inter-rater reliability

BME session	Pairwise				Criterion Referenced		
	N	EXACT	WITHIN_1		N	EXACT	WITHIN_1
First Impression	170	45%	89%		68	41%	85%
Relevant Details	162	35%	76%		63	21%	70%
Appropriate Language	162	35%	86%		63	35%	79%
Logical Flow	160	39%	92%		64	31%	94%
Key Points	162	31%	80%		63	14%	59%
Layout and Design	166	35%	88%		67	27%	84%
Focused Content	162	39%	90%		63	37%	90%
Graphics	119	32%	76%		50	30%	80%
Elaboration	162	40%	77%		63	32%	75%
Vocal Quality	166	43%	84%		64	36%	69%
Personal Presence	166	38%	87%		67	42%	85%

Nuclear Engineering Session inter-rate reliability

Table 4 shows the very reasonable inter-rater reliability for the NRE session. For this session, we are using a cleaned data set that is modified from the original ratings. Several students recorded all '5' ratings for every skill. With one exception, these "all 5" ratings were not realistic representations for presentations observed, so all such results were removed from the data set. Note this is a conservative adjustment, and does not improve our ratings. All pairwise exact scores are at least 40%, and all but 2 skills have over a 90% matching within 1. Likewise the criterion referenced comparison demonstrates very high inter-rater reliability, with all exact scores at least 40%, and all but 1 skill over 90% matching within 1.

Table 4 – NRE session inter-rater reliability

NRE session	Pairwise				Criterion Referenced		
	N	EXACT	WITHIN_1		N	EXACT	WITHIN_1
First Impression	3267	48%	93%		249	56%	95%
Relevant Details	3294	44%	94%		250	49%	94%
Appropriate Language	3220	48%	97%		247	48%	95%
Logical Flow	3292	47%	94%		250	51%	96%
Key Points	3223	48%	95%		247	51%	94%
Layout and Design	3267	43%	91%		249	45%	93%
Focused Content	3241	43%	92%		248	50%	92%
Graphics	3268	46%	90%		249	43%	84%
Elaboration	3292	40%	89%		250	45%	92%
Vocal Quality	3272	45%	92%		249	45%	95%
Personal Presence	3292	40%	88%		250	46%	94%

Overall inter-rater reliability

Using the results presented above, an overall reliability for each skill was determined by combining the results from two 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 twelve 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 eight individual reliability measures for each skill. This index is used to judge whether a skill is highly reliable, moderately reliable, or marginally reliable. The indices assigned for each individual reliability measure are 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 resulted 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 Tables 6 and 7. Table 6 lists all the skills that were modified based on feedback from the previous study, and Table 7 lists those skills unchanged from the previous study. The reliability index for the previous study may be fractional since the reliability index for some changed/redefined skills are the average of two skills (see figure 1).

As shown in the tables, moderate to high reliability was achieved for all of the skills (0 or higher). In Table 6, we observe that reliability of every skill that was modified for this study was better in the current study than in the previous. An unpaired two-sample t-test concludes there is significant improvement in the overall current reliability compared to the previous reliability for those skills that were changed ($p<.001$). In Table 7, we see no significant difference between the reliability of those skills that were unchanged between studies ($p=0.76$).

Table 6 – Overall inter-rater reliability for changed / redefined skills

SKILL	Current Overall Reliability Index	Previous Study Reliability Index
Logical Flow	5	-1
First Impression	4	-1
Personal Presence	3	-1.5
Vocal Quality	3	1
Graphics	1	-1.5
Layout and Design	1	-1.5
Key Points	1	-4.5

Table 7 – Overall inter-rater reliability for unchanged skills

SKILL	Current Overall Reliability Index	Previous Study Reliability Index
Appropriate Language	5	6
Focused Content	4	1
Relevant Details	1	3
Elaboration	0	2

Now that the reliability among raters for these skills has been established, implications and suggestions for instruction need to be addressed. Table 8 includes some tips, for teaching each skill, that have worked well in coaching industrial engineering capstone students to improve their presentation skills[2].

Table 8: Suggestions for instruction

Customizing to the Audience
1. **First impression (Speaker grabs audience attention while defining purpose)**
Ask the presenter to start their presentation with lots of energy and inflection to engage the audience on a personal level. Say "I can be definite!" as an example, and then ask the speaker to try it. Ask the team members for their feedback. If their feedback is "so-so", ask the speaker to try it once more. Give positive feedback when the speaker hits a confident note and then emphasize keeping that up.
2. **Relevant details (Uses concrete examples and details familiar to the audience)**
Emphasize using real-life examples to make the talk engaging and to explain ideas and processes, etc., to the audience. Discuss what could be a concrete example in one or two cases, and then ask the students to come up with their own.
3. **Appropriate language (Describes concepts at just the right level for particular audience)**
Discuss the audience's technical background and whether acronyms, for example, can be used in the talk. If the audience's technical background matches the presenters' then acronyms will be recognized. However, if part of the audience has a technical background and part doesn't, emphasize replacing acronyms with the complete phrase or using the acronym on the slide but describe it in words fully each time it comes up.

Telling the Story
4. **Logical flow (Links and transitions smoothly among different parts of the presentation)**
Ask the group to use storyboarding to create and then check their logical flow. To do this, use a guide sheet with boxes on the page, and enter the title of each slide in each box. If several slides have the same title, ask the students to help the audience know what's coming next by adding a descriptive subtitle to the title for each slide. Second, ask the students to check the logical flow by making sure there are no surprises in the order of the information, and that enough information has been provided ahead of each slide so the audience can understand the slide.
5. Key points (Central message clear throughout by linking details to big picture)
Remind students that if an audience is unfamiliar with their presentation, the audience may be presented a number of details and, in the midst of the details, forget exactly what their purpose is. Therefore, when the students are discussing several details, the presenter will need to regularly remind the audience of how the details connect to the overall purpose or the big picture. Ask students to practice including this reminder when they present details.

Displaying Key Information
6. Layout and design (Easy-to-follow organization, easily digestible amount of text, compatible color)
Ask the students several questions about their slide layout: is the information on the slide balanced and is it easy for the audience to follow? If not, how can it be improved? Is the information set out in the same order that people use to read-- from left to right and top to bottom? Have they avoided having an overwhelming amount of text on the slide? Have they followed the general guideline of eight words per bullet and eight bullets per page? Do the colors used clash with each other when shown on the exact projector and computer to be used during the presentation?
7. Focused content (For each slide, information supports only one or two key points)
Discuss with the students the number of main points appearing on each slide (for example, the executive summary includes one main point: a preview of the project). Then encourage interaction about whether each slide includes enough information to support the one or two key points, keeping in mind the audience's background.
8. Graphics (Maps/charts/graphs/pictures easy to understand and clearly illustrate points)
In the case of graphs, ask students whether they consider that the graphs on their slides include easy-to-understand main points, titles, labels for each axis, and units. If they include multiple graphs on one slide, ask why and whether the graphs can be split up on multiple slides. If there is good reason (such as a comparison) for showing all the graphs on one slide, discuss animation that would help the audience digest the information easily—for example, bringing in one graph at a time.

Delivering the Presentation
9. Elaboration (Avoids reads slides and instead expands upon slide content)
Ask students to check each other's eye contact with the audience while they describe the content of the slides. Eye contact will keep the audience engaged and enable the presenter to determine what the audience's reaction is. Emphasize using the slides as cues to more discussion instead of a document to read. Include a discussion of using notes and how, in many cases, the audience will see this as a case of the presenter not bothering to adequately prepare.
10. Vocal quality (Uses volume, pace, and inflection to emphasize key points)
If the presenter is soft-spoken, ask them to try saying, "I can be definite!" as in skill 1. Remind them that people in the back of the room will need to hear them. The presenter my say they feel like they are shouting. If so, assure them that this is natural when they are not used to projecting. After the group listens to a presenter give their portion of the presentation, ask them whether the presenter emphasized important points by avoiding a monotone and instead using inflection to keep audience interest.
11. Personal presence (Effectively combines energy, eye contact, and movement)
Ask the group to check the presenter's energy and enthusiasm while presenting. Then discuss various ways to show enthusiasm—lots of eye contact, more definite body movement (such as opening your hands away from you) during main points, using engaging expressions and varying facial expression. Advise the groups not to practice the presentation too much or they'll get tired of it or memorize it. They should change their wording each time they practice and then act as if this is the most interesting project they can imagine. Avoiding memorization is important for several reasons: first, the presenter may tend to sound robotic. Second, if the presenter is interrupted, with a question, for example, they may have trouble returning to the exact place in their script. As a result they may review again what they have already stated—possibly to the point of restating several full sentences. Memorization can also cause the speaker to start a perfectly fine phrase and then interrupt or "correct" it with a different phrase—because they remember what is on their script and feel they need to stick to it.

Two skills that can be applied to either (1) an individual speaker or (2) the team as a whole
12. Taking questions (Adeptly answers questions to satisfy the audience)
Remind students that they need to be respectful of their questioner by letting them finish their question. If the presenter doesn't understand it, they should ask for clarification. The next steps involve answering the question concisely and truthfully (for example, if you don't know, don't talk until you think you've said enough so no one will notice you didn't answer the question. Instead, admit you don't know, thank the questioner for making the point, and state that you will look into the answer and get back to the questioner.) Once you've answered the question, look for a nod from the questioner, or check with them by asking, "Is that clear?" or "Did I answer your question?" or something similar.

13. Sensitivity to time (Stays within allotted time even with questions throughout presentation)
Ask the presenters and their team to review their slides and brainstorm possible questions before they present. Then discuss the total time and the estimated time they might allow for questions (for example, depending on the situation, perhaps 5 minutes out of 20). During practice ask the non-speaking group members to help you interrupt the speakers with questions so they can better estimate the time required for the presentation including the questions. Prepare them for the possibility of needing to skip a slide or two if, despite all their preparation, a question takes longer to answer than expected. (If the question starts to delay the talk, the presenter, depending on the situation, may be able to ask the audience member for a separate discussion after the talk ends.) Ask speakers if they know whether, as an individual, they tend to speak more quickly or more slowly in front of an audience (as compared to practice). Then, time each speaker to be sure they stay within their expected time limit.

Conclusions

We have now concluded final revisions to the Norback-Utschig Scoring System for oral presentations in engineering settings. The system has been reduced to 11 skills in four categories plus two additional skills to be rated at the end of a presentation. Each of these skills has been tested for inter-rater reliability, with good to highly reliable results. Finally, we have shared instructional tips for helping students perform well on each skill.

Current work is now targeting ease-of-use of our instructional materials, which will be shared at the presentation. Central to this work is, first, revision of the Teacher's Guide for the Norback-Utschig Presentation Scoring System and second, expanded descriptions for each skill regarding what an excellent performance looks like. Future work includes the development of a proposal for dissemination to a national audience and a study on the cognitive models used by students to create graphs and tables.

References

1. Norback, J., Utschig, T., & Bryan, J. "Workforce Communication Instruction: Preliminary Inter-Rater Reliability Data for an Executive-Based Oral Communication Rubric." *ASEE,* 2012.
2. Norback, J., Utschig, T., & Bryan, J. "Insights into the Process of Building a Presentation Scoring System for Engineers." *ASEE,* 2013.
3. Payne, D., & Blakely, B. (Eds.). (2004-2008). Multimodal Communication: Rethinking the Curriculum. Iowa City, IA: ISUComm at Iowa State University.
4. Payne, D., & Blakely, B. (Eds.). (2007). ISUComm Foundation Courses: Student Guide for English 150 and 250. Iowa City, IA: ISUComm at Iowa State University.

5. Davis, D. “Establishing Inter-rater Agreement for TIDEE’s Teamwork and Professional Development Assessments”. *ASEE Conference Paper*, 2011.

6. Thambyah, A. “On the design of learning outcomes for the undergraduate engineer’s final year project.” *European Journal of Engineering Education* 36.1 (2011): 35-46.

7. Mott, M.S., Chessin, D., Sumrall, W., Rutherford, A., & Moore, V. “Assessing Student Scientific Expression Using Media: The Media-Enhanced Science Presentation Rubric (MESPR).” *Journal of STEM Education: Innovations and Research* 12.1 (2011): 33-41.

8. LaBanca, Frank. “The 21st-Century Oral Presentation Tool Bag.” *Science Teacher* 78.7 (2011): 51-55.

9. Barney, Sebastian, et al. “Improving Students with Rubric-Based Self-Assessment and Oral Feedback.” *IEEE Transactions on Education* 55.3 (2012): 19-325.

10. Anderson, Randy J., and Lydia E. Anderson. "Professorial Presentations: The Link between the Lecture and Student Success in the Workplace." *Academy of Educational Leadership Journal* 14.1 (2010): 55-62. *Business Source Complete*.

11. Battacharyya, Ena, Arun Patil, & Rajeswary Appecutty Sargunan. “Methodology in Seeking Stakeholder Perceptions of Effective Technical Oral Presentations: An Exploratory Pilot Study.” *Qualitative Report* 15.6 (2010): 1549-1568.

12. Morton, J. & Rosse, M. “Persuasive presentations in engineering spoken discourse.” *Australasian Journal of Engineering Education,* 17 (2011): 55-65.

13. Covill, Amy E. “College Students' Use Of A Writing Rubric: Effect On Quality Of Writing, Self-Efficacy, And Writing Practices.” *Journal Of Writing Assessment* 5.1 (2012): 1-19. *MLA International Bibliography*. Web. 20 Dec. 2013.

14. Andrade, H. G. and Du, Y. “Student perspectives on rubric-referenced assessment.” *Practical Assessment, Research, and Evaluation* 10 (2005): 1-11.

15. Rawson, R. E., Quinlan, K. M., Cooper, B. J., Fewtrell, C., & Matlow, J. R. “Writing-skills development in the health professions.” *Teaching and Learning in Medicine,* 17 (2005): 233-239.

16. Morozov, A. “Student attitudes toward the assessment criteria in writing-intensive college courses.” *Assessment Writing*, 16 (2011): 6-31.

17. Utschig, T., Newton, S., Bryan, J. “Measuring Skills across the Profile of a Quality Learner and of a Quality Engineer.” *International Journal of Process Education* 4.1(2012): 3 – 12.

18. Hartberg, Yasha, et al. (2008). Development of Student Writing in Biochemistry Using Calibrated Peer Review. *Journal of the Scholarship of Teaching and Learning.* 2, 29-44.

19. Gunersel, Adalet Baris, and Nancy J. Simpson. Instructors' uses, experiences, thoughts and suggestions regarding Calibrated Peer Review." *Assessment & Evaluation in Higher Education* 35.7 (2010): 771-781.

20. Berry, Frederick C. and Carlson, Patricia A. “Assessing Engineering Design Experiences using Calibrated Peer Review.” *International Journal of Engineering Education* 26.6 (2010): 1503-1507.

Appendix A – Skill Definitions

The definitions for each skill before and after each round of edits to the scoring system are displayed here. This paper primarily describes the second round of edits applied following the collection of stakeholder feedback in round 2 of our modified Delphi method process. Prior edits from round 1 of our modified Delphi process are described elsewhere[2].

Skill	Original Definition(s)	Definition after Delphi round 1 edits	Definition after Delphi round 2 edits
*First Impression	Audience Connection – Refers directly to audience needs to help define purpose/goals of presentation First/Last impression - Grabs audience attention at beginning and inspires them with the closing	Grabs audience attention while defining purpose/goals of presentation	Speaker grabs audience attention while defining purpose
Relevant Details	Uses concrete examples and details familiar to audience	No change	No change
Appropriate Language	Describes concepts at just the right level for particular audience	No change	No change
*Logical Flow	Formerly called sequencing - Links different parts of the presentation and uses appropriate transitions	Links different parts of the presentation and uses appropriate transitions	Links and transitions smoothly among different parts of the presentation
*Key Points	Context - Clearly illustrates major points by linking to additional relevant information Key Points - Consistently refers to how key points fit into the big picture	Material presented is clearly linked to major themes/big picture	Central message clear throughout by linking details to big picture
*Layout and Design	Amount of text - Uses an appropriate amount of text to describe essence of key point(s) Layout and design - Information is easily understood due to layout and color is used appropriately	Information is easily understood due to organization, color, and amount of text	Easy to follow organization, easily digestible amount of text, compatible color
Focused Content	For each slide, information supports only one or two key points	No change	No change

*Graphics	Appropriate graphics - Maps/charts/graphs/pictures/illustrations used clearly support key points Engaging graphics - Graphics are visually appealing, easy to understand, include helpful labeling	Maps/charts/graphs/ pictures easy to understand and clearly illustrate points	No change
Elaboration	Avoids reading slides and instead expands upon slide content	No change	No change
*Vocal Quality	Flow - Knows material well without memorization or repeated hesitations, ums, etc. Vocal Quality - Adapts tone, volume, and pace to emphasize key points	Uses tone, volume, pace, & inflection to emphasize key points	Uses volume, pace, & inflection to emphasize key points
*Personal Presence	Stature - Uses good posture and bearing Personal presence - Effectively combines energy, inflection, eye contact, and movement	Effectively combines energy, eye contact, and movement	No change
†Sensitivity to Time	Begins/ends on time even with questions throughout presentation	No change	Stays within allotted time even with questions throughout presentation
†Taking Questions	Adeptly accepts and satisfactorily answers audience questions	No change	Adeptly answers questions to satisfy audience

* This skill was counted as a changed skill when comparing results between original and final reliability indices.

† This skill was not tested in final reliability calculations due to video format of presentations making these skills difficult or impossible to evaluate.