Measuring the Power of Learning.™

Research Report
ETS RR–15-36

A Prototype Public Speaking Skills Assessment: An Evaluation of Human-Scoring Quality

Jilliam Joe

Christopher Kitchen

Lei Chen

Gary Feng

December 2015

Discover this journal online at
Wiley Online Library
wileyonlinelibrary.com

ETS Research Report Series

EIGNOR EXECUTIVE EDITOR

James Carlson
Principal Psychometrician

ASSOCIATE EDITORS

Beata Beigman Klebanov
Senior Research Scientist - NLP

Heather Buzick
Research Scientist

Brent Bridgeman
Distinguished Presidential Appointee

Keelan Evanini
Senior Research Scientist - NLP

Marna Golub-Smith
Principal Psychometrician

Shelby Haberman
Distinguished Presidential Appointee

Donald Powers
Managing Principal Research Scientist

Gautam Puhan
Principal Psychometrician

John Sabatini
Managing Principal Research Scientist

Matthias von Davier
Senior Research Director

Rebecca Zwick
Distinguished Presidential Appointee

PRODUCTION EDITORS

Kim Fryer
Manager, Editing Services

Ayleen Stellhorn
Editor

Since its 1947 founding, ETS has conducted and disseminated scientific research to support its products and services, and to advance the measurement and education fields. In keeping with these goals, ETS is committed to making its research freely available to the professional community and to the general public. Published accounts of ETS research, including papers in the ETS Research Report series, undergo a formal peer-review process by ETS staff to ensure that they meet established scientific and professional standards. All such ETS-conducted peer reviews are in addition to any reviews that outside organizations may provide as part of their own publication processes. Peer review notwithstanding, the positions expressed in the ETS Research Report series and other published accounts of ETS research are those of the authors and not necessarily those of the Officers and Trustees of Educational Testing Service.

The Daniel Eignor Editorship is named in honor of Dr. Daniel R. Eignor, who from 2001 until 2011 served the Research and Development division as Editor for the ETS Research Report series. The Eignor Editorship has been created to recognize the pivotal leadership role that Dr. Eignor played in the research publication process at ETS.

ETS Research Report Series ISSN 2330-8516

RESEARCH REPORT

A Prototype Public Speaking Skills Assessment: An Evaluation of Human-Scoring Quality

Jilliam Joe, Christopher Kitchen, Lei Chen, & Gary Feng

Educational Testing Service, Princeton, NJ

The purpose of this paper is to summarize the evaluation of human-scoring quality for an assessment of public speaking skills. Videotaped performances given by 17 speakers on 4 tasks were scored by expert and nonexpert raters who had extensive experience scoring performance-based and constructed-response assessments. The Public Speaking Competence Rubric was used to score the speeches. Across all of the dimensions of presentation competence, interrater reliability between expert and nonexpert raters ranged between .23 and .71. The dimensions of public speaking competence associated with the lowest interrater reliability were effectual persuasion and word choice (.41 and .23, respectively). Even expert raters, individuals with a background in teaching and evaluating oral communication, had difficulty agreeing with one another on those dimensions. Low-inference dimensions such as visual aids and vocal expression were associated with much higher levels of interrater reliability, .65 and .75, respectively. The holistic score was associated with an interrater reliability of .63. These results point to the need for a significant investment in task, rubric, and training development for the public speaking competence assessment before it can be used for large-scale assessment purposes.

Keywords Public speaking assessment; human scoring; interrater reliability; multimodal technology

doi:10.1002/ets2.12083

Leaders in the United States K–12 and higher educational systems recognize oral communication competence as a skill students need in the 21st century workplace, a skill that is critical to this nation's social and economic growth (Association of American Colleges and Universities [AACU], 2007). Effective communication has the power to shape ideas, relationships, organizations, and national and global communities.

In the classic sender-receiver channel theory of communication (Shannon & Weaver, 1949), human communication is conceptualized as a linear process between a sender and a receiver. Also included in this model is entropy or noise. This noise represents those aspects of the communication environment and process that interfere with effective communication, such as culture, language, and social and cognitive differences between communicators. For example, consider the number of opportunities for misinterpreted facial expressions or misconstrued messages that occur in a video conferencing environment, particularly when the video and audio are not clear but are fuzzy. A furrowed brow of contemplation can easily be misinterpreted as an expression of anger. The goal of effective communication, then, is to account for and reduce noise in the communication process.

The sender-receiver channel model, though far too simplistic to capture all facets of human communication, is useful for describing public speaking, communication between a single speaker and multiple audience members (e.g., monologue or lecture). The model also holds several implications for the assessment of public speaking competence. Public speaking calls on several kinds of abilities concomitantly: psychological, cognitive, linguistic, and paralinguistic. Assessing public speaking competence is no less complex, particularly when it is performance based. In this context, the evaluator or rater is the audience member. As the receiver of the message, raters process myriad aural and visual information to arrive at a judgment about the speaker's competence and skill level as an orator. And because it is a measurement process, decision making is guided by a scoring rubric that structures the type of behaviors that should be considered and the criteria by which to evaluate those behaviors. However, the extent of how raters in the public speaking domain use the information gathered to make quality score decisions has yet to be fully examined. One of the challenges of assessing public speaking competence is reliably and accurately detecting the true quality of a performance—filtering the construct-irrelevant factors from the construct-relevant factors. In the communication model described earlier, this is characterized as noise.

Corresponding author: G. Feng, E-mail: gfeng@ets.org

Developing a strong training program and scoring design is important to reducing the effect of a noisy communication process on the quality of raters' judgments.

Research on public speaking competence assessment and how to design scoring to reduce noise and promote reliable scoring is limited. A few studies have examined reliability of publicly available rubrics used to evaluate public speaking competence (Morreale, Moore, Surges-Tatum, & Webster, 2007; Schreiber, Paul, & Shibley, 2012) as well as cognitive differences between expert and nonexpert raters (Joe, Harmes, & Hickerson, 2011). A study of multimodal technology applications in presentation skills training was conducted by University of Southern California researchers Batrinca, Stratou, Shapiro, Morency, and Scherer (2013). The focus of the study was the degree to which experts' evaluation of presentation skills and multimodal information on body movement, gestures, and facial features aligned. Two expert raters were asked to score 14 videotaped presentations on 21 behavioral dimensions using a 7-point Likert scale. They were also asked to assign an overall (holistic) score. Content-focused dimensions such as development of the speech and adaption to audience were not evaluated. Batrinca et al. reported an interrater reliability coefficient (ρ) of .65 between raters' overall (holistic) score. Agreement between raters was not examined at the trait level. More research is needed to better understand human scoring for large-scale public speaking assessments and measurement error related to raters that undermines the quality of scoring and, ultimately, the validity of conclusions made based on scores derived through human raters. Research is also needed to determine the ways in which technology can be used to support human scoring and improve the reliability of scores for large-scale public speaking assessments.

In 2013, a multimodal data collection and analysis system and prototype presentation skills assessment were developed at Educational Testing Service (ETS). The system includes a Microsoft Kinect for Windows device for tracking three-dimensional (3D) motions and a high definition (HD) camcorder to track video and voice. Positions and velocities of major body parts, including hands and torso, can be computed from the tracked 3D body motions data. In addition, acoustic features, such as pitch, energy, and pauses, can be extracted from voice recordings. These visual and aural features can be used to measure several critical attributes of presentation skills such as vocal variation and vocal intensity. The presentation skills assessment was administered to an adult population of speakers. This research represents an emerging body of work being conducted in the multimodal technologies for assessment space.

The purpose of this paper is to summarize the evaluation of human-scoring quality for the 2013 multimodal presentation assessment and to provide recommendations for future studies. The scores gathered through this study will contribute to the identification of multimodal features of interest for future automated-scoring development and human-scoring support. Findings will also inform decisions about the appropriateness of the assessment and scoring procedures for large-scale assessment use.

Method

Participants

A total of 17 participants were recruited from among ETS staff; many of whom participate in the Toastmasters club (see Chen et al., 2013). Toastmasters International is an organization that promotes communication and leadership development among professionals by encouraging its members to complete various public speaking exercises, receive feedback on evaluative criteria, and attend agenda-driven club meetings regularly (see http://www.toastmasters.org). In total, seven participants were also active members in Toastmasters Club #5087, the club chapter that is most often attended by ETS employees and convenes on the ETS Rosedale campus. No other participants were members of Toastmasters International. Seven of the participants were male and 10, female. Only two participants were nonnative speakers of English.

Five raters were recruited to score the speech presentations. Two raters were ETS assessment specialists who have a background in oral communication/public speaking instruction at the higher education level. For the purposes of this report, they will be labeled *expert raters*. Both of these expert raters possess degrees in literature or poetry at the master's level and a minimum of 5 year's experience in either teaching or evaluating public speaking and communication skills. Furthermore, both of these raters also possess a secondary graduate degree in either literature or finance and marketing. The other three raters were recruited from the ETS rater pool and were selected based on the recommendations of the expert raters. These raters have extensive experience in scoring constructed responses. Among these three raters, the most recent hire for in this program was in 2005, meaning each rater possesses at least 9 year's experience with constructed-response scoring but little experience scoring public speaking performances with the instrument chosen for the study. For the purposes of this report, they will be labeled *nonexpert raters*.

Presentation Tasks

Four presentation tasks were developed to measure public speaking competencies (see Appendix A). These tasks were modeled after several speaking tasks used in the Toastmasters club. Task A was a short introductory task that served as a warm-up to the main study. It was not scored. Two of the remaining tasks were informative. Task B asked speakers to present a financial report to an imaginary group of board members of a small business, and Task C asked speakers to present a lesson on ancient Egypt to an imaginary group of middle school students. The other two tasks were persuasive and extemporaneous. Task D asked speakers to consider a movie they did not enjoy and extemporaneously convince an imaginary general audience why they should see it. Task E asked speakers to consider a place a general audience would find inconvenient to live in and extemporaneously discuss what is nice about living in such a place.

PowerPoint slides were developed by the test developers and given to participants as visual aids for Tasks B and C. No visual aids were provided for Tasks D and E. All data were collected in a laboratory environment, where participants for the 2013 multimodal presentation assessment investigation were asked to complete these presentation tasks within 2–5 minutes respective to the type of task (4–5 minutes for Tasks B and C and 2–3 minutes for Tasks D and E). Speeches were video and audio recorded. Body movements were tracked using the Kinect system described earlier in the paper. An actual audience (aside from the research assistant) was not present during the presentation.

Scoring Instrument

The Public Speaking Competence Rubric (PSCR; Schreiber et al., 2012) was selected for this study (see Appendix B). Other instruments considered were the competent speaker evaluation form (Morreale, 1990) and a modified version of The Competent Speaker, Scoring Rubric for Oral Communication Behavior Assessment (Hickerson, 2006). The PSCR is intended to be used as an assessment of public speaking competency in adult populations, particularly at the higher education level. It is also intended to be a discipline-neutral instrument, one that can be used to assess globally accepted public speaking competencies. There are nine dimensions considered to be core competencies. Two additional dimensions are included depending on the appropriateness of the task. A holistic score is also applied that characterizes the rater's general impression of the quality of the speech. Each dimension has five performance levels, where 0 reflects *deficient* competence and 4 reflects *advanced* competence. The developers of the PSCR defined the 11 competencies or dimensions as follows (the optional dimensions below are italicized; Schreiber et al., p. 214):

1. Topic selection: The speaker selects a topic that is appropriate to the audience and occasion.
2. Introduction: The speaker formulates an introduction that orients the audience to the topic.
3. Organization: The speaker uses an effective organizational pattern.
4. Supporting materials: The speaker locates, synthesizes, and employs compelling supporting materials.
5. Conclusion: The speaker develops a conclusion that reinforces the thesis and provides psychological closure.
6. Word choice: The speaker demonstrates a careful choice of words.
7. Vocal expression: The speaker effectively uses vocal expression and paralanguage to engage the audience.
8. Nonverbal behavior: The speaker demonstrates nonverbal behavior that reinforces the message.
9. Adapts to audience: The speaker successfully adapts the presentation to the audience.
10. *Visual aids:* The speaker skillfully makes use of visual aids.
11. *Effectual persuasion:* The speaker constructs an effectual persuasive message with credible evidence.

Although the amount of validity evidence for the PSCR is limited and much of the research that has been done on the instrument was conducted by its developers, its representativeness of the public speaking competence domain and psychometric properties demonstrated through prior research indicates promise for use by nonexperts.

Schreiber et al. (2012) conducted two sets of studies to explore interrater reliability and the internal structure of the PCSR. One study was conducted with five trained faculty member raters (two of whom were the instrument developers) to gather validity evidence for the PSCR. The assessment included two tasks: one informative and the other persuasive. A total of 45 speech presentations were given by college students during an introductory course in speech communication. The intraclass correlation coefficients (ICCs; a measure of variance in scores associated with raters) across the 10 dimensions and holistic score ranged from .54 to .93 for the informative speech and from .37 to .88 for the persuasive speech. The holistic score for both tasks tended to be associated with the highest agreement between raters. Overall, across the two

speech tasks, the ICCs ranged from .69 to .93. At the dimension level, however, Dimensions 2–8 were associated with ICCs greater than .70 for the informative speech, and Dimensions 4–7, 10, and 11 were associated with ICCs greater than .70 for the persuasive speech. *Topic selection* and *adapts to audience* tended to give raters the most difficulty when scoring both speech types. The ICCs were lowest for these dimensions ($\rho_{ICC} = .64$ and .54, respectively). Schreiber and colleagues (2012) conducted a second study with an expert and three nonexpert raters following the same procedures as before, but with a different set of 50 speeches, and they found improved agreement for all of the dimensions across both speech types ($\rho_{ICC} = .69$ to .95 for the informative speech and $\rho_{ICC} = .46$ to .92 for the persuasive speech). Possible reasons for this improvement were not given. The findings of the studies suggest experts and nonexperts can use the PSCR reliably.

The factor structure of the measure was also examined in the same set of studies (Schreiber et al., 2012). Specifically, three rounds of principal components analysis (PCA) were completed. The first round of PCA used scores for the first nine dimensions of the rubric. The second round PCA used scores for the first 10 dimensions of the rubric. The third round PCA used scores for Dimensions 1–9 and 11. A sample of 50 speeches was used for these analyses. A three-component solution was replicated in each round: (a) speech presentation, (b) topic adaptation, and (c) nonverbal delivery. Dimensions loaded onto each component in ways that made theoretical sense. For example, Dimensions 2 (introduction), 3 (organization), 4 (supporting materials), 5 (conclusion), 6 (word choice), and 7 (vocal expression) all loaded on speech presentation with loadings that ranged between .374 and .989. Within each component or domain, internal consistency reliability (α) across items was above .75.

The rubric structure differed slightly for the present study. Only Dimensions 2, 3, and 5–9 were used (see Appendix B). The tasks administered did not provide opportunity for topic selection; the topic was given to the speaker (Dimension 1). In addition, supporting materials were not required (Dimension 4); there was not enough advanced preparation time given to speakers to gather supporting materials. Dimension 10 was assessed only when the task required use of visual aids (Tasks B and C). Dimension 11 was considered for all tasks even though only two were persuasive in nature. A holistic evaluation component was included in the scoring design as well. An objective of this study is to determine how reliably experts and nonexperts can use the modified PSCR and how closely scoring performance aligns with prior research.

Scoring Design

Scoring of the 2013 corpus proceeded in two phases. The activities that took place in each phase are summarized in Table 1. Each presentation received at least two independent scores.

The purpose of Phase 1 of the scoring process was to provide criterion scores for a subset of the videos collected in 2013, and from those videos, exemplars were selected (presentations that represent the levels of the scoring rubric) for future training purposes. Two expert raters were involved in this stage. This phase presented the first opportunity for raters to use the scoring instrument and provide feedback on its scorability (i.e., its usability for the set of tasks and presentations).

The purpose of Phase 2 was to score the remainder of the presentations ($N = 55$; one presentation was excluded because of video quality) with expert and nonexpert raters. Presentations within each task were randomly assigned to raters. Experts were randomly paired with nonexperts. The expert rater always provided the first score for each presentation. Nonexperts were never paired with one another. If evidence shows that the experts had a reasonable level of agreement with one another, then their scores could be used as the criteria to assess the accuracy of nonexperts.

None of the raters were aware of the type of score (first or second) they were providing. In the event that the scores between two raters were discrepant, a third rater provided a third independent score and, again, without knowledge of the type of score he or she was providing. The third rater assigned scores to all of the dimensions. However, only the scores for the dimensions with two discrepant scores between an expert and nonexpert rater pair were used. The final operational score is the average of all scores assigned to a given dimension for a given speaker.

Procedures

Phase 1

For Phase 1, scoring sessions were held at ETS's Princeton location. The scoring instrument and a nondisclosure agreement were emailed to expert raters prior to the first session. At the beginning of the session, the research facilitator reviewed

Table 1 Scoring Design for 2013 Presentations

Phase	Purpose	*N* videos	*N* raters	Training (face-to-face)	Scoring model	Presentation distribution
1	• Assign criterion scores • Select exemplars • Refine scoring rubric	12	2 (experts)	• Review of scoring materials • Scoring practice • Calibration discussion	• On-site • View videos in media player • Record evidence and scores on paper copy of score worksheet • No certification or calibration test • 100% double-scored • No validity (monitor) presentations • Third, independent score assigned to dimensions with two discrepant scores (off by two or more).	• Phase 1 sample randomly selected from pool
2	General scoring	55	5 (expert and nonexpert)	• Review of scoring materials • Review of exemplars • Scoring practice • Calibration discussion	• On-site and remote • View videos in media player on computer • Remote scoring: access video through secure file transfer protocol (sftp) site • Record evidence and scores on paper copy of score worksheet • No certification test • 100% double-scored	• Phase 2 presentations randomly assigned to raters • Expert and nonexpert raters randomly paired • % contribution of scores per rater monitored to avoid undue influence

Note. All videos were scored using nine of the 11 PSCR dimensions, and a holistic score also was assigned.

Table 2 Exemplar Matrix

Dimension	Advanced (4)	Proficient (3)	Basic (2)	Minimal (1)	Deficient (0)	N/A
Formulates an introduction that orients audience to topic and speaker	Task B: 14_b	Task C: 19_c (LR)				
Uses an effective organizational pattern	Task B: 14_b		Task E: 16_e (HR)			
Develops a conclusion that reinforces the thesis and provides psychological closure	Task D: 19_d		Task E: 16_e (LR)			Task C: 003_c
Demonstrates a careful choice of words		Task C: 003_c				
Effectively uses vocal expression and paralanguage to engage the audience	Task B: 14_b					
Demonstrates nonverbal behavior that supports the verbal message	Task B: 13_b		Task C: 003_c			
Successfully adapts the presentation to the audience	Task B: 14_b		Task C:			
Skillfully makes use of visual aids	Task B: 14_b					
Constructs an effectual persuasive message with credible evidence and sound reasoning	Task B: 14_b					

Note. LR = low range-finder; performances that are generally or predominantly representative of a particular performance level, but have some characteristics of the preceding performance level. HR = high range-finder; performances that are generally or predominantly representative of a particular performance level, but have some characteristics of the subsequent performance level. In contrast, benchmarks are clearly one performance level or another.

the procedures (see Appendix C) and the scoring materials. Questions about the scoring rubric and the tasks were clarified during this time. Expert raters were given the opportunity to practice their scoring by reviewing a common video (Task B, 5 minutes in length). Each expert had a computer from which to access and view the videos. Headphones were provided to increase the quality of the audio component. Experts used a scoring sheet to record their evidence, scores, and score rationales. Once they scored the practice video independently, they discussed their scores with one another and the facilitator. Any adjacent or discrepant scores were reconciled through discussion. The expert raters were instructed to score the remaining 11 presentations independently.

Scoring notes were developed based on the experts' first round of scoring to give further direction to future raters. Since the purpose of the Phase 1 scoring session was to select exemplars, the experts were asked to record evidence and time stamp that evidence for each of the dimensions from the videos and to also write a rationale for each score they assigned. This often consisted of an amalgamation of individual fragments of evidence that pertain to the same scoring dimension. For example, an expert rater was observed to give a final score of 3 for vocal expression, while corresponding rationales noted by the rater were "excellent vocal variety, enthusiasm" and "uses 'um' excessively." The resultant score of 3 was therefore the rater's judgment of this scoring dimension after considering the preponderance of positive and negative fragments of evidence for that scoring dimension. Exemplars were selected from the set of scored videos that had two exact scores. Exemplars were selected from the set of scored videos that had two exact scores (agreement statistics for Phase 1 scoring are reported in the results section of this report). The initial set of videos from which the exemplars were drawn was small. Therefore, there are a number of incomplete cells in the exemplar matrix, as shown in Table 2.

It was not a requirement that exemplars for all nine dimensions belong to the same speaker or even the same task. In other words, the video did not need to have a score of 4 in every dimension to be considered a candidate for the exemplar

Table 3 Rater by Task Pairings

Rater	Task B	Task C	Task D	Task E
R1 (Expert)	x	x	x	x
R2 (Expert)	x	x	x	x
R3 (Nonexpert)	x	x	x	
R4 (Nonexpert)	x			x
R5 (Nonexpert)		x	x	x

pool. It is uncommon in complex assessments (e.g., classroom observation) such as this for a test taker (or speaker in this context) to have the same performance level on all traits or dimensions of the construct. As shown in Table 2, a Level 4 exemplar for introduction was identified in the Task B set of videos, and it just so happens that the same video (14_b) was used as a Level 4 exemplar for organization; however, the Level 4 exemplar for conclusion was identified in the Task D set of videos. There were likely more Level 4 performances than Level 1 performances in this sample because the participants were selected from a fairly competent group of speakers (i.e., Toastmasters club). Moreover, participants volunteered to participate in the study. In such cases, the more proficient or competent subset of the population tends to be overrepresented.

The scoring rate was much lower than anticipated. We expected raters would score at a factor of 1.5 (time and a half). In reality, the scoring rate was about three times the length of the video. As a result, all of the scoring could not be completed in the first session. Raters were invited to return for a second session on the next day. All of the scoring could not be completed during this session, either. The remaining videos ($N = 2$) were placed in the pool of videos to be scored in Phase 2. The expert raters scored 10 videos across four tasks (three videos for Task B, three videos for Task C, three videos for Task D, and one video for Task E). In total, they assigned 190 scores (scores were not assigned to certain dimensions for certain tasks because of applicability).

In planning for Phase 2, it was expected that raters would not be able to complete all of their scoring assignments on-site. Therefore, a secure file transfer protocol (sftp) site was used to facilitate video file access. The site could not be accessed without the proper login credentials.

Phase 2

Similar procedures were followed for Phase 2. Scoring was conducted over a period of 2 weeks. New (nonexpert) raters were required to come to ETS's Princeton campus for at least the first session. When feasible, nonexperts were scheduled along with experts. The rater pairings are shown by task in Table 3.

As with the Phase 1 scoring, scoring materials (scoring rubric and PowerPoint slides) and the nondisclosure agreement were emailed to raters prior to the scoring session. Once the raters arrived, they were given a copy of the scoring instructions (see Appendix B) and their scoring assignments. Raters were oriented to the procedures. Afterward, the facilitator reviewed the rubric with raters, highlighting key attributes of each dimension and any special notes about scoring the dimension to which raters needed to attend. The expert raters participated in the discussion and were helpful in calibrating the nonexperts to the rubric.

The facilitator then reviewed each of the exemplars with the nonexpert rater. The expert rater was allowed to begin scoring. It is important to note that raters were not required to provide time-stamped evidence or score rationales during this phase because of time constraints. Following the review of the exemplar set, the rater was asked to score the scoring practice video. The facilitator reviewed the criterion scores and score rationales with the rater. After some additional clarification of the procedures, the nonexpert rater was allowed to begin scoring.

Data Analysis

Along with score distribution, agreement summary statistics are reported for each dimension and the holistic score: (a) percentage of exact, adjacent, and discrepant scores; (b) kappa (unweighted and quadratic weighted); (c) Spearman's rho (denoted as ρ); and (d) intraclass correlation coefficient for absolute agreement (denoted as ρ_{ICC}). Spearman rank-order correlations between dimensions were examined, which take into account chance agreement and similarity in

Table 4 Agreement Statistics for Phase 1 (Expert Raters)

		Frequency score distribution					% agreement				
Dimension	*N* (scores)	0	1	2	3	4	Exact	Adjacent	Exact + adjacent	Discrepant	Correlation(ρ)
1. Introduction	20	5%	20%	55%	20%	5%	60%	20%	80%	20%	0.29
2. Organization	20	0%	50%	40%	10%	0%	60%	40%	100%	0%	0.04
3. Conclusion	18	35%	15%	30%	10%	35%	67%	22%	89%	11%	0.75
4. Word choice	20	0%	15%	65%	20%	0%	50%	50%	100%	0%	0.30
5. Vocal expression	20	10%	20%	35%	35%	10%	30%	70%	100%	0%	0.54
6. Nonverbal behavior	20	5%	10%	40%	45%	5%	40%	50%	90%	10%	0.50
7. Adapts to audience	20	5%	15%	65%	15%	5%	30%	60%	90%	10%	−0.03
8. Visual aids	12	0%	5%	50%	5%	0%	67%	33%	100%	0%	—
9. Effectual persuasion	20	5%	35%	45%	15%	5%	40%	40%	80%	20%	−0.10
10. Holistic score	20	0%	25%	65%	10%	0%	70%	30%	100%	0%	0.56

Note. There were 12 presentations in the Phase 1 set, which would have led to 24 scores. However, the experts were able to complete scoring for only 10 of the presentations. The number of scores varied for the conclusion and visual aids dimensions because those dimensions did not apply in all cases.

rank-ordering of test takers (or speakers, in this case) between raters. Spearman correlations were reported because the within-task sample size was too small for a parametric analysis of bivariate relationships. The ICC describes the variance in scores that is attributed to differences between raters.

The expectation for percentage of agreement between raters depends, in large part, on the length of the score scale (i.e., the effective length, the number of scale points raters actually use) and the complexity of the rubric (e.g., the number of dimensions and level of elements within each dimension raters are required to adopt and apply). There are several rules of thumb for examining kappa. Kappa values between .81 and 1 indicate perfect agreement, values between .61 and .80 indicate substantial agreement, values between .41 and .60 indicate moderate agreement, values between .21 and .40 indicate fair agreement, and values between 0 and .20 indicate slight agreement (Landis & Koch, 1977).

Results

Phase 1 Results

Owing to small sample size, kappa and ICC values are not reported. The percentage of exact agreement between experts ranges from 30% to 70%. The holistic score was associated with the highest percentage of agreement (70%). As shown in Table 4, experts had matching or adjacent (1 point difference) scores 80% of the time or more. The lowest percentage of agreement was associated with vocal expression and adapts to audience (30%). The highest percentage of agreement among the presentation competence traits was associated with conclusion and visual aids (67%). The interrater reliability coefficients (ρ) indicate that experts' scores were moderately to highly consistent in four dimensions: (a) conclusion, (b) vocal expression, (c) nonverbal behavior, and (d) the holistic score. Scores were slightly consistent in two dimensions: introduction and word choice. Scores for organization and adapts to audience were uncorrelated. Interestingly, there was a slight inverse correlation between experts' effectual persuasion scores. Performances that one expert ranked as advanced or proficient, the other expert ranked as deficient or minimal.

Phase 2 Results

Expert and nonexpert raters double-scored 55 videos in total across the four tasks. Recall that 12 from the pool of 68 videos were scored in Phase 1, one video was excluded due to video quality, and the remaining 55 were scored in Phase 2. In total, raters assigned 1,032 scores (total scores assigned is less than 1,100 because a visual aids score was not given to Task D and Task E). The following results address the question of whether nonexpert raters apply the same scale (i.e., scoring criteria) as the expert raters. Reported in Tables 5 and 6 are the agreement summary statistics and score distributions for each dimension and the holistic score.

Table 5 Scoring Performance Statistics for Phase 2 (All Raters) Across Tasks

		Frequency score distribution					% agreement				Kappa		Correlation	
Dimension	*N*	0	1	2	3	4	Exact	Adjacent	Exact + adjacent	Discrepant	unwtd	q.wtd	ρ	ICC
1. Introduction	110	0%	2%	22%	45%	31%	58%	31%	89%	11%	0.37	0.30	0.36	0.47
2. Organization	109	0%	2%	25%	50%	22%	49%	42%	91%	9%	0.21	0.33	0.37	0.50
3. Conclusion	103	3%	9%	23%	42%	17%	40%	36%	76%	24%	0.19	0.33	0.40	0.50
4. Word choice	109	0%	0%	12%	60%	27%	55%	40%	95%	5%	0.18	0.25	0.24	0.41
5. Vocal expression	110	0%	1%	28%	38%	33%	47%	53%	100%	0%	0.21	0.59	0.60	0.75
6. Nonverbal behavior	110	0%	8%	22%	49%	21%	49%	42%	91%	9%	0.23	0.46	0.47	0.64
7. Adapts to audience	110	0%	5%	14%	47%	35%	58%	29%	87%	13%	0.35	0.34	0.36	0.51
8. Visual aids	53	0%	3%	7%	25%	13%	44%	44%	88%	11%	0.18	0.47	0.54	0.65
9. Effectual persuasion	108	0%	5%	31%	38%	24%	38%	44%	82%	18%	0.12	0.13	0.15	0.23
10. Holistic score	110	0%	0%	25%	51%	24%	62%	33%	95%	5%	0.39	0.45	0.47	0.63

Note. Corr = correlation; q.wtd = quadratic weighted; unwtd = unweighted; ICC = intraclass correlation coefficient.

Table 6 Mean Scoring Differences Between Expert and Nonexpert Rater

	Mean		
Dimension	Expert	Nonexpert	Mean difference
1. Introduction	3.07	3.04	0.04
2. Organization	2.87	2.98	−0.11
3. Conclusion	2.61	2.71	−0.10
4. Word choice	3.09	3.09	0.00
5. Vocal expression	3.00	3.00	0.00
6. Nonverbal behavior	2.84	2.82	0.02
7. Adapts to audience	2.96	3.27	*−0.31*[a]
8. Visual aids	2.73	3.26	*−0.53*[a]
9. Effectual persuasion	2.78	2.85	−0.07
10. Holistic score	2.89	3.07	−0.18

[a]These difference values, which are italicized, are significant at $p < .05$ level.

Interrater Agreement

As shown in Table 5, raters tended to have matching or adjacent scores 76% of the time or more. Kappa values indicate fair agreement between experts and nonexperts on five of the 10 dimensions and good to moderate for vocal expression, nonverbal behavior, visual aids, and the holistic score. Scores between experts and nonexperts had poor agreement for persuasive. The interrater reliability coefficient (ρ) for these scores ranges from low to moderate (.15 to .60). The dimensions on which raters reached an acceptable level of agreement, where variance between experts and nonexperts was fairly low, were vocal expression ($\rho_{ICC} = .75$), nonverbal behavior ($\rho_{ICC} = .64$), and visual aids ($\rho_{ICC} = .65$). Agreement for all other dimensions, including the holistic score, ranged between .23 and .63. The largest sources of discrepant scores (more than 1 point difference between raters) were conclusion (24%) and effectual persuasion (18%).

Score Distributions

The effective score scale (score range) varied across dimensions either because of raters' scoring criteria or because of the true distribution of scores in the sample (it is difficult to say which because we do not have accurate or criterion scores for each speech presentation). For example, raters assigned scores that ranged from 1 to 4 for introduction. However, for word choice, they assigned scores that ranged from 2 to 4, using an effective 3-point scale for this dimension. The most frequently assigned score in this dataset was 3. In general, any information about oral communication competence that is to be gained from these data is located in the middle to upper range of the score scale. Less information was captured about performances at the lower end of the score scale.

Table 7 Contingency Matrix for Conclusion

		Nonexpert					
	Performance level	0	1	2	3	4	Total
Expert	0	0	0	1[a]	2[a]	0[a]	3
	1	0	1	4	1[a]	0[a]	6
	2	0[a]	1	4	3	1[a]	9
	3	0[a]	2[a]	4	13	4	23
	4	0[a]	0[a]	2[a]	4	4	10
	Total	0	4	15	23	9	51

Note. N is less than 55 because some videos were given an N/A for *conclusion*, applicable in instances where the speaker was clearly cut off from speaking or otherwise did not get to all materials being presented (PowerPoint-aided speeches).
[a]These figures, which are shaded, indicate discrepant score counts.

Table 8 Contingency Matrix for Effectual Persuasion

		Nonexpert					
	Performance level	0	1	2	3	4	Total
Expert	0	0	0	0[a]	0[a]	0[a]	0
	1	0	1	2	0[a]	0[a]	3
	2	0[a]	0	8	8	3[a]	19
	3	0[a]	0[a]	1	10	8	19
	4	0[a]	2[a]	4[a]	5	2	13
	Total	0	3	15	23	13	54

Note. N is less than 55 because one rater had difficulty assigning a score for this dimension for one of the videos. Final was coded as missing.
[a]These figures, which are shaded, indicate discrepant score counts.

As shown in Table 6, experts tended to award higher scores to introduction and vocal expression. Nonexperts, on the other hand, tended to award higher scores to adapt to audience and visual aids. Further, there were statistically significant mean score differences between expert and nonexpert raters. Nonexpert raters tended to score one third of a point higher than experts when scoring adapts to audience and one half of a point higher when scoring *visual aids*. Incidentally, one of the experts noted the difficulty of scoring adapts to audience for Task D and Task E because of the artificial nature of the scoring setting (i.e., rater did not score in real time as an actual audience member).

Discrepant ratings were examined more closely for these two dimensions. The contingency matrix between expert and nonexpert conclusion scores is shown in Table 7. Discrepant score counts are in the shaded cells on the off-diagonal. There was a higher frequency of discrepant scores that involved a 3 rating (10%) and a 2 rating (8%) than any other score level. However, relative to exact and adjacent scores, the amount of discrepant score was low.

The second largest source of discrepant scores was effectual persuasion. Recall that Tasks D and E were the persuasion tasks, though Tasks B and C speeches may have included some persuasive elements, depending on the speaker. The contingency matrix between expert and nonexpert effectual persuasion scores is shown in Table 8. All of the discrepant ratings involved a score of 4. Raters had the most difficulty distinguishing between Level 2 and Level 4 performances (13%).

Summary of Phase 2 Scoring Performance at Task Level

Scoring performance was also examined at the task level. Table 9 contains summary statistics for Task B. Recall that this task asks speakers to present a financial update to board members of a small business. Due to small sample size, kappa and ICC values are not reported. Effectual persuasion and conclusion were associated with the largest percentage of discrepant scores (22% and 21%, respectively). The interrater reliability coefficients indicate that expert and nonexpert scores were the least consistent for effectual persuasion and organization. For the other dimensions, interrater reliability was moderate to high ($\rho = .33$ to .81). Looking at expert and nonexpert mean scores and their differences, a pattern of leniency among nonexperts emerges. Generally, nonexperts gave slightly higher scores than experts for most dimensions. The exceptions

Table 9 Task B Scoring Performance

		% agreement				Correlation	Mean		
Dimension	*N*	Exact	Adjacent	Exact + adjacent	Discrepant	ρ	Expert	Nonexpert	Mean difference
1. Introduction	28	57%	43%	100%	0%	.49	2.93	2.79	0.14
2. Organization	28	50%	43%	93%	7%	.16	2.86	2.86	0.00
3. Conclusion	28	64%	14%	78%	21%	.40	2.62	2.77	−0.15
4. Word choice	28	64%	36%	100%	0%	.33	2.93	2.93	0.00
5. Vocal expression	28	64%	36%	100%	0%	.81	2.93	2.93	0.00
6. Nonverbal behavior	28	50%	43%	93%	7%	.65	2.50	2.64	−0.14
7. Adapts to audience	28	67%	33%	100%	0%	.68	2.86	3.21	−0.36
8. Visual aids	28	56%	33%	89%	11%	.68	2.50	3.07	−0.57
9. Effectual persuasion	28	44%	33%	77%	22%	.21	2.50	2.79	−0.29
10. Holistic score	28	67%	33%	100%	0%	.69	2.79	2.86	−0.07

Table 10 Task C Scoring Performance

		% agreement				Correlation	Mean		
Dimension	*N*	Exact	Adjacent	Exact + adjacent	Discrepant	ρ	Expert	Nonexpert	Mean difference
1. Introduction	26	77%	23%	100%	0%	.78	3.15	3.23	−0.08
2. Organization	26	54%	38%	92%	8%	.69	2.92	3.38	−0.47
3. Conclusion	26	31%	46%	77%	23%	.10	1.91	2.82	−0.91
4. Word choice	26	46%	38%	84%	15%	.32	3.08	3.08	0.00
5. Vocal expression	26	38%	62%	100%	0%	.56	3.15	3.15	0.00
6. Nonverbal behavior	26	54%	38%	92%	8%	.63	3.08	3.00	0.08
7. Adapts to audience	26	54%	23%	77%	23%	.54	2.85	3.54	−0.69
8. Visual aids	26	38%	46%	84%	15%	.24	3.00	3.46	−0.46
9. Effectual persuasion	26	23%	69%	92%	8%	.37	2.75	3.33	−0.58
10. Holistic score	26	46%	46%	92%	8%	.44	3.00	3.46	−0.46

were organization, word choice, and vocal expression, where mean scores were equal. Sample sizes were too small to test for significance of these differences.

Table 10 contains summary statistics for Task C. Recall that Task C asks speakers to present a lesson on ancient Egypt to prepare a hypothetical group of middle school students for a future exam. Interrater reliability was the lowest for conclusion and visual aids. For the other dimensions, scores were moderately to highly consistent ($\rho = .32$ to .78). Looking at expert and nonexpert mean scores and their differences, the pattern of leniency among nonexperts is consistent with Task B scoring behavior. Generally, nonexperts gave slightly higher scores than experts for most dimensions. The exceptions were word choice and vocal expression, for which mean scores were equal, and nonverbal behavior, for which experts gave slightly higher scores, on average, than nonexperts. Sample sizes were too small to test for significance of these differences.

Table 11 contains summary statistics for Task D. Recall that Task D asks speakers to extemporaneously persuade an audience to see a really bad movie. The interrater reliability coefficients indicate that expert and nonexpert scores for introduction and effectual persuasion was practically uncorrelated. The consistency of scores between the two groups was low to moderate for the other dimensions ($\rho = .14$ to .50). There is a marked difference in scoring quality between this task and the previous two tasks. Generally, nonexperts gave slightly higher scores than experts for most dimensions. The exceptions were word choice and vocal expression, for which mean scores were equal, and conclusion and effectual persuasion, for which experts gave slightly higher scores, on average, than nonexperts. Sample sizes were too small to test for significance of these differences.

Table 12 contains summary statistics for Task E. Recall that Task E asks speakers to extemporaneously discuss what is nice about living in a place that you would find very inconvenient or annoying. As with the previous tasks, effectual

Table 11 Task D Scoring Performance

		% agreement				Correlation	Mean		
Dimension	*N*	Exact	Adjacent	Exact + adjacent	Discrepant	ρ	Expert	Nonexpert	Mean difference
1. Introduction	28	50%	21%	71%	29%	-.04	2.79	3.07	−0.29
2. Organization	28	43%	43%	86%	14%	.27	2.57	2.71	−0.14
3. Conclusion	28	29%	43%	72%	29%	.43	2.92	2.57	0.35
4. Word choice	28	50%	43%	93%	7%	.24	3.21	3.21	0.00
5. Vocal expression	28	43%	57%	100%	0%	.50	3.00	3.00	0.00
6. Nonverbal behavior	28	64%	21%	85%	14%	.14	2.86	2.93	−0.07
7. Adapts to audience	28	64%	14%	78%	21%	.22	2.93	3.14	−0.21
8. Visual aids	—	—	—		—	—	—	—	—
9. Effectual persuasion	28	36%	36%	72%	29%	.06	2.93	2.64	0.29
10. Holistic score	28	57%	29%	86%	14%	.39	2.79	3.07	−0.29

Note. Visual aids did not apply to this task.

Table 12 Task E Scoring Performance

		% agreement				Correlation	Mean		
Dimension	*N*	Exact	Adjacent	Exact + adjacent	Discrepant	ρ	Expert	Nonexpert	Mean difference
1. Introduction	28	50%	36%	86%	14%	.25	3.43	3.07	0.36
2. Organization	28	50%	43%	93%	7%	.29	3.14	3.00	0.14
3. Conclusion	28	57%	43%	100%	0%	.70	2.86	2.71	0.14
4. Word choice	28	57%	43%	100%	0%	.26	3.14	3.14	0.00
5. Vocal expression	28	43%	57%	100%	0%	.46	2.93	2.93	0.00
6. Nonverbal behavior	28	29%	64%	93%	7%	.40	2.93	2.71	0.21
7. Adapts to audience	28	50%	43%	93%	7%	.28	3.21	3.21	0.00
8. Visual aids	—	—	—		—	—	—	—	—
9. Effectual persuasion	28	50%	36%	86%	14%	.19	2.93	2.71	0.21
10. Holistic score	28	64%	36%	100%	0%	.62	3.00	2.93	0.07

Note. Visual aids did not apply to this task.

persuasion was associated with the lowest interrater reliability ($\rho = .19$). Interrater reliability was low to moderate for the other dimensions ($\rho = .25$ to .70). Conclusion and the holistic score were associated with the highest consistency between experts and nonexperts. Generally, nonexperts gave slightly lower scores than experts for most dimensions. The exceptions were word choice, vocal expression, and adapts to audience, for which mean scores were equal. Sample sizes were too small to test for significance of these differences.

Discussion

Presentation performances were scored by nonexpert human raters as well as expert human raters who had extensive experience scoring performance-based and constructed-response assessments, but were minimally trained on a version of the PSCR (Schreiber et al., 2012), modified for the study. Speech performances given by 17 speakers on four tasks were video- and audiotaped. In addition, body movements were tracked using the Microsoft Kinect for Windows system for future data analysis and possible use as a scoring support for human raters. Each performance was evaluated on dimensions related to content, adaptation, and nonverbal delivery along a 5-point rating scale. The rubric was reviewed with raters, and each rater completed a scoring practice video to calibrate to the score scale.

Scoring was conducted in two phases. Phase 1 involved expert raters. They were tasked with scoring a subset of the videotaped performances and assigning a criterion score to each for the purpose of training nonexpert raters. Phase 2 scoring involved expert and nonexpert raters. The remainder of the videos was scored during this phase. The Phase 1

analysis of scoring performance addressed the question of whether expert raters' scores were trustworthy criteria by which nonexpert scores could be compared. Results showed that expert agreement met the 60% target for only three of the 10 scores assigned: organization, conclusion, and vocal expression. It is important to note that the final criterion score was derived through consensus discussion. The Phase 2 analysis addressed scale usage and the question of whether nonexpert raters apply the same scale (i.e., scoring criteria) as the expert raters. In general, scores from middle and upper range of the scale were awarded more frequently than scores at the lower end of the scale. This finding was expected given that a majority of the study participants were members of the ETS Toastmasters club and were fairly competent speakers.

The findings also show that nonexperts did not use the same scale as experts when scoring visual aids and adapts to audience. Nonexperts, on average, scored up to a half of a point higher than experts on these dimensions. At the task level, nonexperts generally tended to score higher than experts on Tasks B through D and lower than experts on Task E. It is not immediately clear why this was the case.

The largest sources of discrepant scores (more than 1 point difference between raters) across all four tasks were conclusion (24%) and effectual persuasion (18%). Conclusion may have presented some challenges because judging the psychological closure of the speech is very much idiosyncratic to each rater. Granted, there were several instances where one rater did not detect a conclusion and the other rater not only detected a conclusion, but also thought that it warranted a score of 3. These types of discrepancies must be addressed through training. For the tasks in which persuasion was required, Tasks D and E, rater agreement was low. This could be due to the misalignment between the scoring criteria and the tasks. For example, the scoring criteria outline three critical attributes that raters considered when evaluating effectual persuasion: (a) credibility of evidence, (b) call to action, and (c) reasoning. Recall that *supporting materials* was excluded from the rubric because speakers did not have time to gather supporting materials for their presentations. It stands to reason that a lack of supporting materials and evidence would negate the need to evaluate the credibility of evidence. It is possible that because that criterion was included in the rubric, raters felt obligated to use it. Any identified evidence may or may not have been consistent across raters. Further, because speakers were not made aware of the scoring criteria, they may not have known to make the call to action explicit in their presentations. It is also possible that some raters could have missed the call to action if one was present. If retained in the 2014 assessment, these tasks must be revised to elicit the kinds of behaviors that are consistent with the evaluation standards (as they are expressed in the rubric), or the rubrics must be revised to align to the tasks.

In addition, the results also showed that raters had difficulty coming to agreement on scores for Task D more than other tasks. Speakers who had difficulty thinking of an example of a bad movie were allowed to modify the task (e.g., bad book) and raters were told to expect some variation. However, it is unclear if this played a role in low interrater reliability. It is more likely that the lack of benchmarks for Task D contributed significantly to the quality of scoring. There was only one example of a high-scoring speech for conclusion for this task. Incidentally, Task D and Task E had fewer benchmarks and poorer scoring quality than Task B and Task C, which underscores the importance of providing sufficient scoring supports during training.

As shown in Table 13, the ICCs in this study were well below the ICCs reported in Schreiber et al. (2012). However, while not explicitly stated, it is highly probable that the five raters in Study 1 and four raters in Study 2 of Schreiber et al. each scored all of the speech presentations. In other words, the scoring was fully crossed. It is not clear from the description provided in the paper if this is true. However, if this is the case, then that would explain why the ICCs were so high compared to the ICCs from our study. The more raters, the higher interrater reliability generally will be when scores are averaged across raters. For instance, for the sample of videos in our study that received a third rater, interrater reliability (ρ_{ICC}) increased to .78 from .40 when the third rating was applied. In an operational setting, three to five raters per presentation may not be a realistic or sustainable scoring model given the expense.

Differences between the quality of scoring for the current study and the Batrinca et al. (2013) study were also examined using the relationship between experts' Phase 1 holistic (overall) scores. The interrater reliability coefficient for raters' scores for the present study was .63. Batrinca et al. (2013) reported an interrater reliability of .65. Despite the differences in the scoring procedures, the similarity in reliability of scoring between the two studies suggests that a holistic impression of public speaking competence can be measured somewhat reliably by a group of experts and nonexperts. The findings of the current study point more specifically to certain dimensions of public speaking competence where reliability

Table 13 Comparison of Results to Schreiber et al. (2012) Study

	% agreement			Kappa			Schreiber et al. (2012)	
Dimension	Task E	Task A	Task D	unwtd.	q.wtd.	ICC	ICC (Study 1)	ICC (Study 2)
1. Introduction	0.58	0.31	0.11	0.37	0.30	0.47	.64	.76
2. Organization	0.49	0.42	0.09	0.21	0.33	0.50	.79	.81
3. Conclusion	0.40	0.36	*0.24*	0.19	0.33	0.50	.84	.75
4. Word choice	0.55	0.40	0.05	0.18	0.25	0.41	.85	.73
5. Vocal expression	0.47	0.53	0.00	0.21	0.59	0.75	.84	.83
6. Nonverbal behavior	0.49	0.42	0.09	0.23	0.46	0.64	.77	.70
7. Adapts to audience	0.58	0.29	0.13	0.35	0.34	0.51	.54	.86
8. Visual aids	0.44	0.44	0.11	0.18	0.47	0.65	.70	.83
9. Effectual persuasion	0.38	0.44	*0.18*	0.12	0.13	0.23	.81	.84
10. Holistic score	0.62	0.33	0.05	0.39	0.45	0.63	—	—

Note. q.wtd = quadratic weighted; unwtd = unweighted; ICC = intraclass correlation coefficient.

must be significantly improved, such as effectual persuasive and word choice, in addition to conclusion as mentioned earlier.

A major limitation of the study was the amount of raw video available to develop training. As mentioned, the range of exemplars at each of the score levels for each dimension was limited. Exemplars are foundational to training. The extent to which the exemplar set is limited will affect the quality of raters' scoring. In some cases, there were no exemplars for a particular score level of a dimension. As an example, there was only one exemplar for effectual persuasion (Score Level 1). That dimension was associated with the lowest level of interrater reliability (ICC = .23). Another example of this was word choice, where only one exemplar could be identified. Interrater reliability was also low (ICC = .41).

Another possible limitation of the study is the background qualifications of the raters. None of the nonexpert raters had a background in evaluating public speaking competency or, to our knowledge, teaching public speaking. Future studies may want to use raters with scoring experiences that align more with the construct being assessed, particularly for new assessments where exemplars and other training material are limited.

Finally, the scoring instrument does not provide direction for how to address English-as-a-second-language factors in the scoring criteria. Several of the raters noted the difficulty in scoring and making allowances for second language issues for the two speakers whose native language was not English. We will consult with assessment developers to determine the ways in which we can refine the scoring rubric and develop a more comprehensive set of benchmarks to promote higher quality scoring.

Conclusion

The results of the current study point to the need for a significant investment in task, rubric, and training development for public speaking competence assessment. Practically across all of the dimensions, agreement among raters did not meet an acceptable level of agreement. Even expert raters, individuals with a background in teaching and evaluating oral communication, had difficulty agreeing with one another on certain dimensions (effectual persuasion and word choice).

In addition to providing raters with extensive training and opportunities to practice scoring, aggregating the dimensions into the three components identified by Schreiber et al. (2012) might improve the quality of scoring. There would be fewer pieces of information raters would need to manage cognitively. Alternatively, if measuring oral communication competence at the expanded trait level is important (e.g., for formative assessment purposes), there may be opportunity to employ computer-assisted scoring once the multimodal technology has been developed and validated. Raters would be able to draw upon more objective measures of body movement and vocal qualities upon which to base their judgments. Studies can explore how automated scoring-assisted scoring impacts the cognitive quality of human judgments at the trait level as well as the statistical outcomes of scoring. Again, there are some very basic steps that can and should be taken to improve agreement among raters such as providing a complete set of benchmarks and range finders for each of the dimensions, more robust score rationales, and more scoring practice videos.

References

Association of American Colleges & Universities. (2007). *College learning for the new global century*. Washington, DC: Author.

Batrinca, L., Stratou, G., Shapiro, A., Morency, L. P., & Scherer, S. (2013, August). *Cicero — Towards a multimodal virtual audience platform for public speaking training*. Retrieved from http://ict.usc.edu/pubs/Cicero%20-%20Towards%20a%20Multimodal%20Virtual%20Audience%20Platform%20for%20Public%20Speaking%20Training.pdf

Chen, L., Feng, G., Kitchen, C., Leong, C. W., Lee, C., & Hakkinen, M. (2013). *A preliminary study of using multimodal evidence (speech/video/3D-tracking) to analyze presentations.* Manuscript in preparation.

Hickerson, C. (2006). *Scoring rubric for oral communication behavior assessment.* Unpublished manuscript, James Madison University, Harrisburg, VA.

Joe, J. N., Harmes, J. C., & Hickerson, C. A. (2011). Using verbal reports to explore rater perceptual processes in scoring: A mixed methods application to oral communication assessment. *Assessment in Education: Principles, Policy, & Practice, 18*(3), 239 – 258.

Landis, J. R., & Koch, G. G. (1977). A one way components of variance model for categorical data. *Biometrics, 33,* 671 – 679.

Morreale, S. (1990, November). *The competent speaker: Development of a communication-competency based speech evaluation form and manual.* Retrieved from ERIC database. (ED325901)

Morreale, S., Moore, M., Surges-Tatum, D., & Webster, L. (2007). *The competent speaker speech evaluation form* (2nd ed.). Washington, DC: National Communication Association.

Schreiber, L. M., Paul, G. D., & Shibley, L. R. (2012). The development and test of the Public Speaking Competence Rubric. *Communication Education*, *61*(3), 205 – 233.

Shannon, C. E., & Weaver, W. (1949). *The mathematical theory of communication*. Urbana: University of Illinois Press.

Appendix A

Presentation Skills Assessment Task Description and Task Notes

First Prepared Presentation: *Technical Presentation*

Task	Present financial update to board members of a small business.
Preparation time	10 minutes
Presentation time	4 – 5 minutes
PowerPoint slides	Canned

The participant is asked to take a few minutes to review slides for the first PowerPoint presentation and informed that any questions regarding the material provided can be answered by the researcher prior to, but not during the task.

For this task, we will give you a few slides to review for a PowerPoint presentation you will be asked to present. The task will resemble a simple financial update being given to board members of a small business. You will be given note pages for each of the slides to review first and certain points will be highlighted in red for you to emphasize during your talk. These are points that you want to make sure the board members understand. It is not necessary to recite the material you are provided with, and instead you are welcome to improvise or develop your own presentation, as long as you cover the highlighted points.

You will not have your notes available to you during your talk, but you are welcome to study them and write your own notes for up to 10 minutes. We expect to record for roughly 4 – 5 minutes. You should consider having enough to talk about on each slide to fill the 5 minute duration. This will effectively mean spending about 1 minute per slide. If you have any questions about the material you are provided with, please let us know and one of the researchers will answer any questions you have. During your talk, we will not be able to answer your questions, however. As before, we will let you know if we have enough recorded for this task.

Second Prepared Presentation: *Lecture Presentation*

Task	Present a lesson (lecture style) to prepare a group of middle school students for a future exam.
Preparation Time	10 minutes
Presentation Time	4 – 5 minutes
PowerPoint slides	Canned

The participant is asked to take a few minutes to review slides for the second PowerPoint presentation. As in the first task, the participant may ask questions only pertaining to the material provided prior to the task itself.

The next task is much like the previous one. We will give you some slides to review for another presentation. This task will resemble a simple lecture that may be given to a group of middle school students. Like before, you will have a moment to review notes on each of the slides and highlighted in red are several points that you should emphasize to your students as these items will be on a future exam. Again, you will have up to 10 minutes to prepare and we expect to record for roughly 4 – 5 minutes, or roughly 1 minute per slide. You will not have access to your notes during your talk. If you have questions about the material you are provided with, please let us know and one of the researchers will answer any questions you have. During your talk, we will not be able to answer your questions, however.

Second Table Topics Task: *Bad Movie Recommendation*

Task	Extemporaneously persuade an audience to see a really bad movie
Preparation Time	0 minutes
Presentation Time	2 – 3 minutes
Power Point slides	None

The participant is given a topic that he or she will need to talk about for approximately 2 – 3 minutes. The participant will also be informed that there will be no preparation time, and he or she should try to talk throughout the allotted time.

In the next task, you will be asked to speak for only 2 – 3 minutes about a topic we give you. There is no preparation time for this task and we will begin recording once the topic has been read to you. Like before, we will let you know if we have recorded enough information and we can move to the next task.

For this speech, we want you to think of a really bad movie you have seen. Your task is to convince us, your audience, that this movie is really worth seeing and tell us why. You may begin.

Third Table Topics Task: *Obscure Living*

Task	Discuss what is nice about living in a place that you would find very inconvenient or annoying
Preparation Time	0 minutes
Presentation Time	2 – 3 minutes
PowerPoint slides	None

The participant again is given a topic to talk about for 2 – 3 minutes with no preparation time. The participant may discontinue the talk if he or she feel there is nothing else to say on the topic.

For the last task, you will be given another topic to discuss for 2 – 3 minutes. Like the last task, you will not have preparation time and we will begin recording once the topic is given to you. We will let you know if we have recorded enough information for this task.

First, think of a place that would be very inconvenient or annoying for you to live in (allow the participant some time to consider this). Please tell us what is nice about living in this place.

Appendix B

Modified Public Speaking Competence Scoring Rubric

Performance standard	Assessment criteria						
The student …	Advanced (4)	Proficient (3)	Basic (2)	Minimal (1)	Deficient (0)	N/A	Notes on implementation
REMOVED 1. Selects a topic appropriate to the audience and occasion							
2. Formulates an introduction that orients audience to topic and speaker	Excellent attention getter; firmly establishes credibility; sound orientation to topic; clear thesis; preview of main points cogent and memorable	Good attention getter; generally establishes credibility; provides some orientation to topic; discernible thesis; previews main points	Attention getter is mundane; somewhat develops credibility; awkwardly composed thesis; provides little direction for audience	Irrelevant opening; little attempt to build credibility; abrupt jump into body of speech; thesis and main points can be deduced but are not explicitly stated	No opening technique; no credibility statement; no background on topic; no thesis; no preview of points		
3. Uses an effective organizational pattern	Very well organized; main points clear, mutually exclusive and directly related to thesis; effective transitions and signposts	Organizational pattern is evident, main points are apparent; transitions present between main points; some use of signposts	Organizational pattern somewhat evident; main points are present but not mutually exclusive; transitions are present but are minimally effective	Speech did not flow well; speech was not logically organized; transitions present but not well-formed	No organizational pattern; no transitions; sounded as if information was randomly presented		
REMOVED 4. Locates, synthesizes, and employs compelling supporting materials							

Appendix B: Continued

Performance standard	Assessment criteria						
The student …	Advanced (4)	Proficient (3)	Basic (2)	Minimal (1)	Deficient (0)	N/A	Notes on implementation
5. Develops a conclusion that reinforces the thesis and provides psychological closure	Provides a clear and memorable summary of points; refers back to thesis/big picture; ends with strong clincher or call to action	Appropriate summary of points; some reference back to thesis; clear clincher or call to action	Provides some summary of points; no clear reference back to thesis; closing technique can be strengthened	Conclusion lacks clarity; trails off; ends in a tone at odds with the rest of the speech	No conclusion; speech ends abruptly and without closure	Given in instances where the speaker is clearly cut off from speaking or otherwise does not get to all materials being presented on (PowerPoint-aided speeches)	
6. Demonstrates a careful choice of words	Language is exceptionally clear, imaginative and vivid; completely free from bias, grammar errors, and inappropriate usage	Language appropriate to the goals of the presentation; no conspicuous errors in grammar; no evidence of bias	Language selection adequate; some errors in grammar; language at times misused (e.g., jargon, slang, awkward structure)	Grammar and syntax need to be improved as can level of language sophistication; occasionally biased	Many errors in grammar and syntax; extensive use of jargon, slang, sexist/racist terms, or mispro-nunciations		
7. Effectively uses vocal expression and paralanguage to engage the audience	Excellent use of vocal variation, intensity, and pacing; vocal expression natural and enthusiastic; avoids fillers	Good vocal variation and pace; vocal expression suited to assignment; generally avoids fillers (e.g., um, uh, like)	Demonstrates some vocal variation; enunciates clearly and speaks audibly; fillers (e.g., um, uh, like) present but do not detract from the message	Sometimes uses a voice too soft or articulation too indistinct for listeners to comfortably hear; often uses fillers	Speaks inaudibly; enunciates poorly; speaks in monotone; poor pacing; distracts listeners with fillers		Consider the extent that fillers are a distraction in differentiating Scores 2 and 1 and frequency of filler use for differentiating Scores 3 and 2

Appendix B: Continued

Performance standard	Assessment criteria						
The student …	Advanced (4)	Proficient (3)	Basic (2)	Minimal (1)	Deficient (0)	N/A	Notes on implementation
8. Demonstrates nonverbal behavior that supports the verbal message	Posture, gestures, facial expression, and eye contact well developed, natural, and display high levels of poise and confidence	Posture, gestures, and facial expressions are suitable for speech, speaker appears confident	Some reliance on notes but has adequate eye contact, generally avoids distracting mannerisms	Speaker relies heavily on notes; nonverbal expression stiff and unnatural	Usually looks down and avoids eye contact; nervous gestures and nonverbal behaviors distract from or contradict the message		No notes are available during any task
9. Successfully adapts the presentation to the audience	Speaker shows how information is personally important to audience; speech is skillfully tailored to audience beliefs, values, and attitudes; speaker makes allusions to culturally shared experiences	Speaker implies the importance of the topic to the audience; presentation is adapted to audience beliefs, attitudes and values; an attempt is made to establish common ground	Speaker assumes but does not articulate the importance of topic; presentation was minimally adapted to audience beliefs, attitudes, and values; some ideas in speech are removed from audience's frame of reference or experiences	The importance of topic is not established; very little evidence of audience adaptation; speaker needs to more clearly establish a connection with the audience	Speech is contrary to audience beliefs, values, and attitudes; message is generic or canned; no attempt is made to establish common ground		Task b: board members of a business firm; Task c: class of 8th graders; Tasks d and e: general audience

Appendix B: Continued

Performance standard	Assessment criteria						
The student …	Advanced (4)	Proficient (3)	Basic (2)	Minimal (1)	Deficient (0)	N/A	Notes on implementation
10. Skillfully makes use of visual aids	Exceptional explanation and presentation of visual aids; visuals provide powerful insight into speech topic; visual aids of high professional quality	Visual aids well presented; use of visual aids enhances understanding; visual aids good quality	Visual aids were generally well displayed and explained; minor errors present in visuals	Speaker did not seem well practiced with visuals; visuals not fully explained; quality of visuals needs improvement	Use of visual aids distracted from the speech; visual aids not relevant; visual aids poor professional quality	Given in instances where no visual aids are present, Tasks D and E	Do not evaluate based on the quality of visual aids, but rather solely on the presenter's use of visual aids in the speech
11. Constructs an effectual persuasive message with credible evidence and sound reasoning	Articulates problem and solution in a clear, compelling manner; supports claims with powerful/credible evidence; completely avoids reasoning fallacies; memorable call to action	Problem and solution are clearly presented; claims supported with evidence and examples; sound reasoning evident; clear call to action	Problem and solution are evident; most claims are supported with evidence; generally sound reasoning; recognizable call to action	Problem and/or solution are somewhat unclear; claims not fully supported with evidence; some reasoning fallacies present; call to action vague	Problem and/or solution are not defined; claims not supported with evidence; poor reasoning; no call to action		Discount the definition of problem/solution in tasks where a well defined problem does not appear to exist, Tasks C and potentially B

Note. A holistic score is also assigned that captures a rater's overall impression of the quality of the performance, taking into account all of the dimensions of public speaking competence.

Appendix C

Scoring Instructions

Video:

1. ***No pausing or rewinding***. Please watch each video in its entirety without pausing and rewinding.
2. ***Record evidence***. As you are watching the video, take notes or jot down evidence of behaviors that are relevant to the dimensions on the scoring rubric. Once the video has stopped, review your notes, and assign a score to nine (9) of the eleven (11) dimensions of the *Public Speaking Competence Rubric*.

Tips for taking down evidence:

1. **Evidence, Bias, Interpretation.** Take down what you see and what you hear. Avoid using descriptors such as "nice" and "good" if they are not a part of the scoring criteria. These are very subjective. Rely on the language of the rubric. Also avoid making inferences beyond that which you can observe in the video.
2. **Avoid transcribing the video**. Only attend to the behaviors that correspond to the dimensions of the presentation that are being evaluated.
3. **Take notes first and sort later**. Initially, note-taking and scoring will be made easier if you record your evidence first and assign that evidence to the appropriate dimension after the video has stopped. As you become more experienced in using the rubric, you will recognize what pieces of evidence belong to a given dimension, and you will be able to sort as you go.

Tips for Minimizing Bias:

1. **Note your biases and bias triggers**. Underlying biases (e.g., racial, gender, region, religion) and/or personal preferences (e.g., speaking style, attire, hairstyle) can make a major difference in the quality of the scores raters provide. The key to minimizing the influence of your biases and personal preferences on your score decisions is to first acknowledge what those biases are and what triggers those biases to manifest in your decision making.
2. As you are scoring, if you encounter a trigger that prohibits you from making an objective evaluation of the speaker's performance, please indicate "Hold" on the top of the form and we will have another rater score that performance.

Task Notes:

1. Since this is a pilot of this assessment for research purposes, it would be helpful if you recorded any thoughts or considerations concerning the tasks as you are scoring. Specifically, we would like to know anything about the task that interferes with your ability to score it using the PSCR.

Suggested Citation:

Joe, J., Kitchen, C., Chen, L., & Feng, G. (2015). *A prototype public speaking skills assessment: An evaluation of human-scoring quality* (Research Report No. RR-15-36). Princeton, NJ: Educational Testing Service. http://dx.doi.org/10.1002/ets2.12083

Action Editor: Keelan Evanini

Reviewers: Chaitanya Ramineni and Katrina Roohr

ETS and the ETS logo are registered trademarks of Educational Testing Service (ETS). MEASURING THE POWER OF LEARNING. is a trademark of ETS. All other trademarks are property of their respective owners.

Find other ETS-published reports by searching the ETS ReSEARCHER database at http://search.ets.org/researcher/