Measuring the Power of Learning.™

Research Report
ETS RR–18-09

A Preliminary Investigation of the Factors Related to the Design and Scoring of Video-Based Oral Communication Performance Tasks in Higher Education

Katrina Crotts Roohr

Kri Burkander

Liyang Mao

December 2018

ETS Research Report Series

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 Preliminary Investigation of the Factors Related to the Design and Scoring of Video-Based Oral Communication Performance Tasks in Higher Education

Katrina Crotts Roohr,[1] Kri Burkander,[1] & Liyang Mao[2]

1 Educational Testing Service, Princeton, New Jersey
2 IXL Learning, San Mateo, California

Oral communication has been identified as an important skill by higher education institutions and by the workforce community. Despite its importance, minimal research has been conducted around the development of tasks to measure oral communication skills and behaviors. The purpose of this preliminary study is to evaluate the different factors related to the design and scoring of oral communication tasks. For this study, oral communication refers specifically to video-based oral presentations using a webcam. The study administered 2 informative and 2 persuasive speech tasks of varying length (2 or 4 minutes) to 8 college students. Performance was scored holistically on overall performance, and on content and delivery dimensions by 2 raters. Results show that (a) interrater reliability estimates may be impacted by task type and length, (b) there is no variation in the three scoring criteria, (c) a minimum of 2 tasks is necessary to achieve satisfactory reliability on an oral communication measure, and (d) supporting materials may impact student performance. Results of this preliminary study provide guidance on how to further this foundational research. Future research and next steps are discussed.

Keywords Oral communication; oral presentations; interrater reliability; higher education

doi:10.1002/ets2.12197

Oral communication has been identified as an important skill by higher education institutions (e.g., Association of American Colleges and Universities, 2011) and by the workforce community (e.g., Casner-Lotto & Barrington, 2006; Hart Research Associates, 2015). As a result, there is a stronger push to evaluate college students' oral communication skills to determine whether students are prepared to enter the workforce upon graduating college. Research has recommended that an assessment of oral communication be developed to measure both oral communication skills and behaviors (Morreale, Backlund, Hay, & Jennings, 2007). To capture both oral communication skills and behaviors, a more authentic form of assessment (i.e., a performance assessment) is needed where students perform a speaking task to demonstrate their skills rather than self-evaluation surveys or traditional multiple-choice questions. Thus, it is critical to investigate what factors need to be considered when designing and scoring open-ended speaking tasks. Specifically, in this study, we wanted to evaluate what factors are critical for students to accurately demonstrate their skills and behaviors and whether these responses result in valid and reliable scores.

When developing a performance assessment of oral communication, a number of factors need to be considered, such as which speaking behaviors to assess, the number of speaking tasks a student should complete, the length of the performance, and the amount of preparation time (Morreale, Backlund, et al., 2007). These different variables can greatly impact the length of the test administration and the amount of time and requirements needed to appropriately score the performance tasks. For instance, in relation to the length of the performance (i.e., task length), it is important to evaluate what the ideal task length is to respond to an open-ended speaking task and ensure that students have enough time to demonstrate their oral communication skills (e.g., that they are able to effectively organize a speech). It is also important to understand how task length can impact rater scoring. For instance, for a rater to accurately score a students' oral communication performance, do they need to watch a student speak for a minimum amount of time? In fact, in relation to task length, Morreale, Backlund, et al. (2007) suggested that future research on oral communication assessment development focus on "determining the size and diversity of speech samples required for a reliable indication of competence" (p. 13).

Few studies have evaluated factors such as performance task length for oral communication measures. This is likely because in higher education, most of the measures for evaluating oral communication are standardized rubrics and do

Corresponding author: K. C. Roohr, E-mail: kroohr@ets.org

not include predefined prompts or tasks. Instead, the tasks are determined by the users or developers of the assessment (e.g., giving an in-class presentation). As a result, much of the existing literature around performance tasks has focused on second language learners. For instance, Malabonga, Kenyon, and Carpenter (2005) found that the amount of time a test taker takes to respond to an oral task depends on a number of factors, such as a test taker's affect or emotion, word and language knowledge, and the test taker's ability to meet the demands of the task. The authors argued that "adequate response time may allow examinees to demonstrate their language abilities in a testing situation" (Malabonga et al., 2005, p. 65) but noted that it may not always be permissible given testing restraints. Although their study involved second language learners, Malabonga et al. provided insight into response time on speaking tasks. The authors found that more difficult tasks elicit longer responses from students and that more proficient students provide longer responses.

Other research has mainly evaluated the length of preparation time for speaking tasks on English proficiency measures (e.g., Abdi, Eslami, & Zahedi, 2012; Li, Chen, & Sun, 2015; Wigglesworth & Elder, 2010). Specifically, these studies have investigated the impact of preparation time on the quality, fluency, and accuracy of oral speech. Studies have shown mixed results regarding the effect of planning time, with some studies showing no significant differences in performance with different amounts of planning time (e.g., Wigglesworth & Elder, 2010). However, other studies have suggested that planning time does impact performance. For instance, Li et al. (2015) found that for a 90-s task, the quality and accuracy of speech improved with increased planning time (1 – 3 minutes); however, too much planning time (i.e., 5 minutes) showed diminishing returns.

In summary, previous research has only nominally evaluated factors such as response time and planning time for speaking tasks and has mainly focused on English proficiency or language assessments; however, speaking and oral communication are critical skills for all students, not just second language learners.

Defining Oral Communication

Oral communication has been defined in various ways throughout higher education (e.g., Adelman, Ewell, Gaston, & Schneider, 2014; Morreale, Rubin, & Jones, 1998; Rhodes, 2010), workforce (e.g., Employment Training Administration, 2014; National Institutes of Health, 2017), and K – 12 communities (e.g., National Communication Association, 1998; National Governors Association Center for Best Practices & Council of Chief State School Officers, 2010). Such definitions vary based on the type of communication or form of spoken discourse (e.g., public speaking or presentation skills, small group, interviewing). Also, various terms are used to describe oral communication in these contexts, such as *communication fluency*, *listening and speaking*, *effective communication*, and *communication and collaboration* (Roohr, Mao, Belur, & Liu, 2015). For the purpose of this study, we define oral communication in relation to oral presentations, as this has been a focus in defining oral communication in the context of higher education (see Adelman et al., 2014; Binkley et al., 2012; CAS Board of Directors, 2008; Morreale et al., 1998; Quality Assurance Agency, 2008; Rhodes, 2010).

In synthesizing the existing literature, Roohr et al. (2015) proposed an assessment framework that targets both knowledge and skills for effectively communicating formal and informal messages in a higher education context. This proposed framework has two dimensions: content and delivery. Content focuses on the purpose and reason for communicating, developing an appropriate structure or organization according to the purpose, and using appropriate language conventions and grammar. Delivery focuses on adapting the language to the targeted audience, individual, or context and demonstrating vocal (e.g., pacing, pitch, volume) and nonverbal (e.g., eye contact, facial expressions, gestures) behaviors to enhance the message. These aspects of oral communication were important to the development of the scoring rubrics for this study (see the Methods section).

Purpose and Research Questions

The purpose of this preliminary study was to evaluate the different factors related to design and scoring needed to produce valid and reliable scores of oral communication behaviors in higher education. For this study, eight college students were administered four oral communication performance tasks (informative and persuasive speeches) of varying length (2 or 4 minutes) as well as a self-report measure of oral communication performance to validate their observed score. Student performance was measured holistically across three scoring criteria: (a) overall score, (b) content dimension score, and (c) delivery dimension score. The following research questions were investigated:

1 What is the rater consistency across the three scoring criteria (overall, content, and delivery) for each of the oral communication performance tasks?
2 How much variance in student scores is due to task type, task length, rater, and scoring criteria? Are there significant differences in student performance across task type, task length, and scoring criteria?
3 What is the relationship between student performance on tasks and other related variables (e.g., perceived preparation time, perceived response time, and self-rated oral communication skill)?

This preliminary study has important implications for the design and development of authentic speech performance tasks for an oral communication assessment in higher education. Oral communication is not easily measureable using traditional multiple-choice assessment and instead requires face-to-face interaction (e.g., in a classroom assessment) or the use of audio- or video-recorded format. In this preliminary study, we focused on the video-recorded format and investigated various factors related to design of oral communication performance tasks, including the amount of preparation time (5 vs. 7 minutes), the use of supporting materials, the type of task (informative vs. persuasive), and the task length (2 vs. 4 minutes). In relation to scoring, we investigated three scoring criteria (overall, content, and delivery). This preliminary study can inform future research in determining the important factors needed to design and score oral communication tasks to obtain reliable and valid scores for students in U.S. higher education institutions.

Method

Participants

Participants for this study included eight summer undergraduate interns at Educational Testing Service (ETS) all attending 4-year institutions in bachelor's programs. Interns were compensated with a $10 gift certificate to be used to purchase lunch or coffee on the ETS Princeton campus. The majority of the sample was male (75%), and all students were native English speakers (see Table 1). Fifty percent were Asian/Asian American students. Three students were classified as college sophomores, three as juniors, and two as seniors based on their credit hours completed. Students' primary college majors varied across the eight students and included education, psychology, integrated marketing communication, banking/finance, English, engineering and engineering technologies, and philosophy. More than half the sample (62.5%) did not take any college-level classes related to oral communication or public speaking prior to participating in this study.

Instruments

Oral Communication Performance Tasks

Given that our targeted context for this study was higher education, we focused on the oral presentation aspect of oral communication. Morreale, Backlund, et al. (2007) noted that there are approaches that can be used to assess speaking skills, including an observational approach, such as assessing and observing students in a classroom setting, and a more structured approach, assessing students on a specific performance or structured task. For this study, we focused on the structured approach to evaluate the various factors that should be considered when developing these tasks for large-scale use. Many existing assessments are self-report (Morreale, Backlund, et al., 2007) or do not require the test taker to complete specific tasks but instead employ the use of a standardized rubric with open-ended prompts that are determined by the user of the rubric (Roohr et al., 2015). Although assessments with standardized rubrics and open-ended prompts offer flexibility to the user, there is significant variation in how the rubrics are used, which can impact interrater reliability and make comparisons across scores less meaningful.

For this study, we focused on video-based oral presentation tasks. Video-based tasks can make larger scale assessment more feasible, as multiple students can record their performance simultaneously. Additionally, video-based performance can create a more controlled scoring environment for raters (Morreale, Backlund, et al., 2007). Each oral communication performance task asked the test taker to give a speech presentation. Typically, speeches serve three general purposes, including informing, persuading, and entertaining (Wrench, Goding, Johnson, & Attias, 2012). Informative speeches involve synthesizing information and informing the audience on a topic. These speeches can be demonstrative or explanatory. Persuasive speeches involve taking a position and convincing the audience to take a different point of view. These types of speeches can be argumentative in nature to attempt to alter the viewpoint of the audience. Entertaining speeches

Table 1 Student Demographics

Demographic	*n*	%
Males	6	75
Race/ethnicity		
White	3	37.5
Asian/Asian American	4	50
Black/African American	1	12.5
Native English speakers	8	100
Full-time students	7	87.5
Credit hours		
30–60 (sophomores)	3	37.5
61–90 (juniors)	3	37.5
>90 (seniors)	2	25
College grade point average		
3.50–4.00	6	75
3.00–3.49	1	12.5
2.50–2.99	1	12.5

Table 2 Within-Subjects Design

Task type	Preparation time (minutes)	Task length (minutes)	Supporting materials
Persuasive	5	2	5 slides: pros/cons of school uniforms
Persuasive	5	4	No supporting materials
Informative	7	2	9 slides: obtaining a passport
Informative	7	4	8 slides: history of tobacco use

can be either informative or persuasive, but the main purpose is to provide the audience with enjoyment, whether funny or serious. A toast or introduction can be considered an entertaining speech.

For the purposes of this study, we focused on informative and persuasive speeches, given that these two types of speeches are more likely to occur in an academic or workplace setting as compared to an entertaining speech. For this study, four oral communication performance tasks were administered to each student using a within-subjects design (see Table 2), meaning each student participated in each condition. The four tasks (two informative and two persuasive) were newly developed by assessment developers at ETS. Tasks were tested internally with two experts in oral communication and public speaking prior to this study. These internal tests focused on evaluating whether there was sufficient preparation time to navigate the supporting materials, reviewing the tasks for typos or inconsistencies, and making sure that the tasks were understandable and clear to the speaker.

Each of the speech tasks asked students to construct either an informative or persuasive message that was either 2 or 4 minutes long. Both the informative speeches had supporting materials. The 2-minute speech had nine slides of information about obtaining a passport. The 4-minute speech had eight slides about the history of tobacco use. The 2-minute persuasive task asked the test taker to prepare a speech to a school board regarding school uniforms at public schools. This task had five slides of supporting materials. The 4-minute persuasive speech did not include preparation materials; test takers were asked to think of a place they had visited and persuade their classmates to go there.

There are numerous ways in which supporting materials can be presented to students for an oral communication task. For instance, students can be provided with the prompts ahead of time to allow the test takers to prepare their own slides, or students can be provided with the supporting materials ahead of time to give them more time to prepare. As previously discussed, we wanted to standardize the prompts as much as possible to allow for better comparisons across student performance. As a result, we opted to present the prompt and supporting materials to the student at the time of the assessment and gave students an allotted amount of preparation time before giving their presentations. Because students were given the supporting materials at the time of the assessment, raters were advised to consider this when scoring the student's oral performance, as students had less time to prepare.

In relation to preparation time, students were given 5 minutes (not including time to read the task or prompt) for the persuasive tasks and 7 minutes of preparation time for the informative tasks (see Table 2). Preparation time for the informative tasks was longer given that there were more supporting materials to navigate as compared to the persuasive tasks. Preparation time allowed students to navigate the supporting materials that were provided to complete the task and take notes on provided note cards. Supporting materials were presented in the form of PowerPoint slides and contained information from Web pages and other sources. Students were discouraged from reading their notes during the presentation and were expected to use the supporting materials in their speeches. All student responses were recorded using a webcam.

Communication Competence Self-Report Questionnaire

In addition to the four oral communication performance tasks, students also completed the Communication Competence Self-Report (CCSR), a self-report measure intended for college students to evaluate speech performance (Rubin, 1985). This measure included 19 Likert-type items asking students to rate the frequency of their speaking and interpersonal behaviors. Research involving the self-assessment of second language testing has found that self-assessment can provide fairly robust concurrent validity evidence (Ross, 1998). The CCSR was included in this study to capture other information about student oral communication skills that could be used to evaluate concurrent validity of the four tasks. The internal consistency of this measure is .87. Additional information and psychometric properties of the CCSR can be found in Rubin (1985). For this study, reliability was lower than reported in Rubin (1985), with a coefficient alpha of .62.

Background Information Questionnaire

After completing the oral communication performance tasks and CCSR, a student background information questionnaire (BIQ) was administered. This BIQ captured information on student gender, ethnicity, college major, and other demographic and college-level variables. The BIQ also asked students to rate the following statements for each task on a scale of 1 (*strongly disagree*) or 5 (*strongly agree*):

1 The amount of time allowed for preparing my answer for this task was appropriate.
2 The amount of time allowed for responding to the task was appropriate.

Scoring Rubric

Videos were human scored by two experienced raters using two holistic rubrics based on a synthesis of existing frameworks, definitions, and rubrics of oral communication (see Roohr et al., 2015). Examples of the synthesized frameworks, definitions, and rubrics included the Association of American Colleges and Universities' VALUE (Valid Assessment of Learning in Undergraduate Education) rubric (Rhodes, 2010), Lumina's Degree Qualification's Profile (Adelman et al., 2014), the Competent Speaker Speech Evaluation (CSSE) form (Morreale, Moore, Surges-Tatum, & Webster, 2007), and the Public Speaking Competence Rubric (PSCR; Schreiber, Paul, & Shibley, 2012), to name a few. Many of these existing rubrics require the rater to score on multiple dimensions that may have some overlap. Additionally, scoring on multiple dimensions may be difficult for raters and impact interrater reliability. For instance, the PSCR requires the rater to score the test taker on 11 different aspects of the presentation (Schreiber et al., 2012). Despite that the PSCR has 11 dimensions, the authors noted that two key factors emerged in their analyses—content and delivery, that is "putting together the speech content, and delivering the content well" (p. 222). The CSSE also has multiple dimensions or competencies (eight in total). Additionally, at the holistic level, raters are also asked to score on preparation and content and on presentation and delivery (Morreale, Moore, et al., 2007). As a result, we decided to develop rubrics that focused on two dimensions: content and delivery, which also aligned with our construct definition of oral communication.

Raters scored students using holistic rubrics on a 0 to 4 scale for two different dimensions: content and delivery (see Appendix A for the informative task rubrics and Appendix B for the persuasive task rubrics). Content captured whether students organized the speech effectively, including introducing and concluding the speech, as well as whether students conveyed ideas fluently and demonstrated facility with the conventions of the English language. Delivery captured whether students displayed strong vocal variation and pacing of the speech; exhibited confidence with posture, gestures, and facial expressions; appropriately adapted the speech to the intended audience; and adapted appropriately to the environment in which the speech was being given (in this case, over a webcam). After raters completed scoring on the content and

delivery dimensions (there was no particular order in terms of scoring these two dimensions), raters were also asked to provide an overall score on a 0 to 4 scale. Raters used both the content and delivery rubrics together to provide an overall score (there was not a specific overall performance rubric that was used).

Study Design and Procedure

Using a within-subjects design, all eight students were administered each of the four tasks. A counterbalanced design was implemented to avoid order effects. Using a within-subjects design both increases our power and avoids any subject-to-subject variation (Seltman, 2014).

Students were tested in a private computer lab, and the video and screen capture was recorded using Morae software (Techsmith, 2015). Students were given instructions for the assessment and were shown instruction slides to simulate what the actual testing session would look like. This gave students the opportunity ask questions and test the recording system. During the assessment, students were shown each task with instructions. These instructions indicated the amount of preparation time and speech time for that particular task. Students were then prompted to navigate through the preparation materials and could take notes at this time. At the end of the preparation time, students were prompted to begin recording their speeches. A clock with a countdown of the time was presented on the screen. When time was up, students were presented with the next task. Collectively, the four tasks took approximately 40 minutes to complete. Students could not see themselves on screen during the presentation but could only see the tasks. After completing the four tasks, students were prompted to complete the CCSR followed by the BIQ. Total testing time took approximately 1 hour.

Rater Training

Two human raters scored each video-based response. With eight students and four tasks, a total of 32 videos were evaluated and scored. The two raters were previously trained raters who have experience scoring written communication assessments such as the *GRE*® general test. Raters were compensated for their time. The two raters were trained for approximately 2 hours about the purpose of the study and how to use the rubric. Given that this was the first time this rubric was used for scoring, we did not have specific training materials; instead, much of the training was around clarifying any questions around the scoring rubrics. We hypothesized that this could impact our interrater reliability results. As a result, during the scoring of the videos, raters were also asked to provide feedback on the rubrics to better understand any inconsistencies in the ratings.

Data Analyses

Research Question 1

To investigate the consistency across raters or interrater reliability, agreement summary scores were reported for the content and delivery dimensions and for the overall score. Specifically, we evaluated percentage exact and adjacent agreement, Spearman's rho (ρ), and the quadratic weighted kappa (κ_{QW}) coefficient. Spearman correlations were used instead of conducting parametric analyses (e.g., intraclass correlation) due to small sample sizes (Joe, Kitchen, Chen, & Feng, 2015). The kappa coefficient indicates the proportion of agreement between the two raters beyond what is expected by chance (Fleiss & Cohen, 1973). Kappa ranges from −1.00 to 1.00, where less than .00 is poor agreement, .00 – .20 is slight agreement, .21 – .40 is fair agreement, .41 – .60 is moderate agreement, .61 – .80 is good agreement, and .81 – 1.00 is very good agreement (Landis & Koch, 1977). Quadratic weighted kappa is used to evaluate the degree of disagreement between raters when the scores are ordinal by applying weights to kappa to capture larger disagreements in scores. Examining the interrater reliability allowed us to evaluate the consistency in their scores across the various dimensions and across tasks. Inconsistencies in scores meant that the tasks themselves were difficult to measure or that the rubric needed some revision.

Research Question 2

To evaluate how much variance in student scores was due to the rater (Rater 1 or 2), task type (persuasive or informative), task length (2 or 4 minutes), and scoring criteria (overall, content, and delivery), we conducted a generalizability (G) study.

The G study provided information about the sources of error influencing measurement. We used a fully crossed design with four facets (the student was the object of measurement, and rater, task type, task length, and scoring criteria were the four facets): Student × Rater × Task Type × Task Length × Scoring Criteria. The G study results allowed us to quantify the sources of error in test score related to these various facets and the interactions between these facets. The goal here was to generalize these results to a larger set of test items.

Following the G study, we conducted a decision (D) study to investigate how changing the number of tasks, task length, and scoring criteria impacted reliability. Specifically, we used the variance components estimates to optimize the measurement for a particular decision-making purpose (Shavelson & Webb, 1991). Using the results from the G study, we investigated how increasing the number of tasks, task length, and scoring criteria impacted our overall error variance and reliability coefficients (*G* and phi coefficients). The software GENOVA (Crick & Brennan, 1983) was used to conduct these analyses.

Additionally, to evaluate differences in scores across task types with varying response lengths and to evaluate differences across the task types, we conducted a repeated measures multivariate analysis of variance (MANOVA) due to the multivariate structure of the data. For our analyses, the within-subjects factor was the four conditions, and our dependent variables were the average score across the two raters for the three scoring criteria (content, delivery, and overall scores) for each condition. MANOVA analyses allowed us to determine whether there were overall differences in scores across the four task types of varying length. Effect sizes for significant main effects were evaluated using partial eta squared where .01 is considered a small effect, .06 is medium, and .14 is large (Cohen, 1988). Cohen's *d* was used to evaluate effect sizes for the planned contrasts, where .20 is small, .50 is moderate, and .70 is large (Cohen, 1988). The relationships between the three criteria with each of the performance tasks were also evaluated using Spearman correlations.

Research Question 3

To evaluate the relationship between score (overall, content, and delivery) and student perception regarding preparation time and task length, we calculated Spearman correlations between student performance and the two Likert-type item responses that were asked about each of the four tasks. These analyses allowed us to investigate whether a test taker felt impacted by preparation and testing time, which could help to explain variation in performance. Students were asked to rate their oral communication skills on a 5-point scale ranging from 1 (*excellent*) to 5 (*poor*), and we evaluated the relationship between those ratings and the overall score using a chi-square analysis. We also evaluated the relationship between overall item-level performance and scores on the CCSR using a Spearman correlation to evaluate concurrent validity evidence.

Results

Interrater Reliability

Interrater reliability was evaluated between the two raters across the three scoring criteria for the four tasks (Research Question 1). Table 3 shows the frequency score distributions, percentage agreement, Spearman correlations, and quadratic weighted kappa coefficients across all tasks and individually for each of the four tasks. Overall, results across all tasks show that the two raters had exact or adjacent agreement 94% of the time across the four tasks. Moderate correlations between .58 and .68 were found across the three scoring criteria. Quadratic weighted kappa coefficients indicated moderate to good agreement across the three scoring criteria, with the delivery dimension showing the strongest agreement ($\kappa_{QW} = .72$), followed by overall performance ($\kappa_{QW} = .66$) and content dimensions ($\kappa_{QW} = .49$).

When looking at the separate tasks, results show greater consistency in scoring criteria for the two 4-minute tasks as compared to the two 2-minute tasks with reliability estimates (κ_{QW}) between .44 (moderate) and .86 (very good) for the 4-minute tasks compared to a range of .26 (fair) to .60 (moderate) for the 2-minute tasks. For the 2-minute persuasive task, fair consistency was found across the two raters on the content ($\kappa_{QW} = .39$) and delivery ($\kappa_{QW} = .33$) dimensions. For the 2-minute informative task, fair consistency was found on the content dimension ($\kappa_{QW} = .26$). In general, the raters had the least consistency on the content dimension in terms of scoring and were typically the most consistent on the delivery dimension.

Table 3 Interrater Reliability Across the Scoring Criteria for the Four Performance Tasks

		Frequency score distribution				% agreement				
Dimension	*N* (scores)	1	2	3	4	Exact (*E*)	Adjacent (*A*)	*E* + *A*	Correlation (ρ)	Quadratic weighted kappa (κ_{QW})
All tasks										
Content	64	5	19	28	12	41	53	94	.58	.49
Delivery	64	7	18	30	9	59	41	100	.68	.72
Overall	64	5	21	28	10	53	47	100	.68	.66
2-Minute persuasive										
Content	16	1	3	8	4	13	88	100	.35	.39
Delivery	16	1	4	10	1	38	63	100	.32	.33
Overall	16	2	2	9	3	38	63	100	.41	.60
4-Minute persuasive										
Content	16	1	2	8	5	50	38	88	.76	.44
Delivery	16	2	2	8	4	75	25	100	.75	.85
Overall	16	1	2	10	3	63	38	100	.77	.68
2-Minute informative										
Content	16	2	8	5	1	50	38	100	.29	.26
Delivery	16	2	5	8	1	38	63	100	.42	.57
Overall	16	1	9	5	1	50	50	100	.66	.50
4-Minute informative										
Content	16	1	6	7	2	50	50	100	.77	.63
Delivery	16	2	7	4	3	75	25	100	.87	.86
Overall	16	1	8	4	3	63	38	100	.84	.76

Variance in Student Scores and Reliability

Variance in student scores (Research Question 2) was calculated using G-Theory. Variance components estimates are shown in Table 4. It is important to note, however, that due to the small sample sizes, these estimates may not be stable. These results are only intended to serve as a preliminary understanding of the variance components. Results show that most of the variation in scores was due to the variation in students. Results show some variation in the two task types and little to no variation in task length and rater. That said, raters did have some variation in scores across the individuals (Person × Rater) and scoring criteria (Rater × Score Criteria), which likely explains some of the lower estimates of interrater reliability. Despite the fact that there were clear differences in interrater reliability estimates across the 2- and 4-minute tasks, the interaction between task length and rater accounted for less than 1% of the explained total variance. Additionally, results show no variation in the three scoring criteria. D study results indicate that this model had satisfactory reliability with a generalizability coefficient of .83 and phi coefficient of .77.

In addition to the G study, we also calculated D studies to see how the change in the number of task types, task length, and score criteria impacted reliability (Table 5). The number of raters remained stable at two raters because we would want at least two raters to score each video in an operational setting. With two raters, we were able to double score a response as a quality measure (i.e., we are able to evaluate the interrater agreement with two raters; McClellan, 2010). Results show that a minimum of two items were needed to obtain adequate reliability above .70. For instance, Design 5 showed adequate reliability with at least one task type (informative or persuasive) of two varying lengths (2 or 4 minutes). Similarly, Design 8 showed adequate reliability with at least two task types (both informative and persuasive) of one length (either 2 or 4 minutes).

Performance Differences

Mean scores for the four tasks across the three scoring criteria are shown in Table 6. A repeated measures MANOVA was used to evaluate differences across task type and across the average scores for the three scoring criteria. Results show that there were significant differences in the four tasks, $F(3, 19) = .80$, $p < .01$, $\eta^2_{\text{partial}} = .52$ (Pillai's trace = .52). Specifically, significant differences were found between the 2-minute informative task and the 2-minute persuasive task, with students

Table 4 Variance Component Estimates from Generalizability (G) Study

Effect	*df*	Variance component	% total variance
Person	7	0.40	43.72
Error	14	0.11	11.35
Task type	1	0.06	6.20
Person × Task Type	7	0.05	5.64
Person × Task Length × Score Criteria	14	0.05	4.83
Person × Rater	7	0.04	4.67
Person × Task Type × Score Criteria	14	0.04	4.03
Person × Task Type × Rater × Score Criteria	14	0.03	3.70
Person × Task Type × Task Length	7	0.03	3.38
Rater × Score Criteria	2	0.03	3.30
Person × Task Type × Task Length × Rater	7	0.03	3.06
Person × Task Length × Rater × Score Criteria	14	0.01	1.37
Person × Task Type × Rater	7	0.01	1.13
Task Length × Rater	1	0.01	0.97
Person × Task Length × Rater	7	0.01	0.72
Task Type × Task Length × Score Criteria	2	0.004	0.56
Task length	1	0.004	0.48
Task Type × Task Length × Rater × Score Criteria	2	0.004	0.48
Rater	1	0.004	0.40

Note. All other effects and interactions showed zero variation.

Table 5 Decision (D) Study Results

D study design	No. test takers	No. task types	No. task lengths	No. raters	No. score criteria	Gen. coeff.	Phi coeff.
1	8	1	1	2	1	0.59	0.53
2	8	1	1	2	2	0.66	0.59
3	8	1	1	2	3	0.69	0.62
4	8	1	2	2	1	0.66	0.59
5	8	1	2	2	2	0.72	0.64
6	8	1	2	2	3	0.74	0.66
7	8	2	1	2	1	0.69	0.64
8	8	2	1	2	2	0.76	0.71
9	8	2	1	2	3	0.79	0.73
10	8	2	2	2	1	0.76	0.70
11	8	2	2	2	2	0.81	0.75
12	8	2	2	2	3	0.83	0.77

Note. The first column includes the number of D study designs that were conducted. Columns 3–6 represent the changing variables that were used to calculate the reliability for each D study. Number of test takers and number of raters remained constant for each D study.

performing better on the persuasive task ($d = -0.63$; see Table 7). Significant differences were also found between the 4-minute persuasive task and both the 2-minute informative ($d = -0.80$) and 4-minute informative tasks ($d = -0.49$). No significant differences were found between the two persuasive tasks or between the two informative tasks; however, it is important to note that due to the small sample sizes, we may have been more likely to obtain nonsignificant results due to lack of power. With a larger sample size, we may have seen different results. This was the same issue for the correlation results we present subsequently.

Between-subjects effects show that there were no significant differences between content, delivery, and overall score types, $F(2, 21) = .04$, $p = .96$. Figure 1 shows the plot of mean task scores by score criteria, which could indicate that Tasks 1 and 2 (the 2- and 4-minute persuasive tasks) were easier for students on average as compared to the informative tasks and that there was little variation in the three scoring criteria. These results could also indicate that the raters were more severe in their application of the scoring rubrics for the informative tasks. Additionally, Spearman correlation results show that the three scoring criteria were highly correlated for most tasks. Interestingly, for the 4-minute persuasive task (the

Table 6 Mean Content, Delivery, and Overall Scores for the Performance Tasks

Task	Score type	*M*	*SD*
2-Minute persuasive	Content	2.94	0.73
	Delivery	2.69	0.59
	Overall	2.81	0.84
4-Minute persuasive	Content	3.06	0.73
	Delivery	2.88	0.95
	Overall	2.94	0.73
2-Minute informative	Content	2.31	0.65
	Delivery	2.47	0.74
	Overall	2.38	0.64
4-Minute informative	Content	2.63	0.74
	Delivery	2.50	0.96
	Overall	2.56	0.86

Table 7 Within-Group Pairwise Comparison Effect Sizes (Cohen's *d*)

Task	1	2	3	4
2-Minute persuasive (1)	–			
4-Minute persuasive (2)	.20	–		
2-Minute informative (3)	−.63*	−.80**	–	
4-Minute informative (4)	−.33	−.49*	.24	–

*$p < .05$. **$p < .01$.

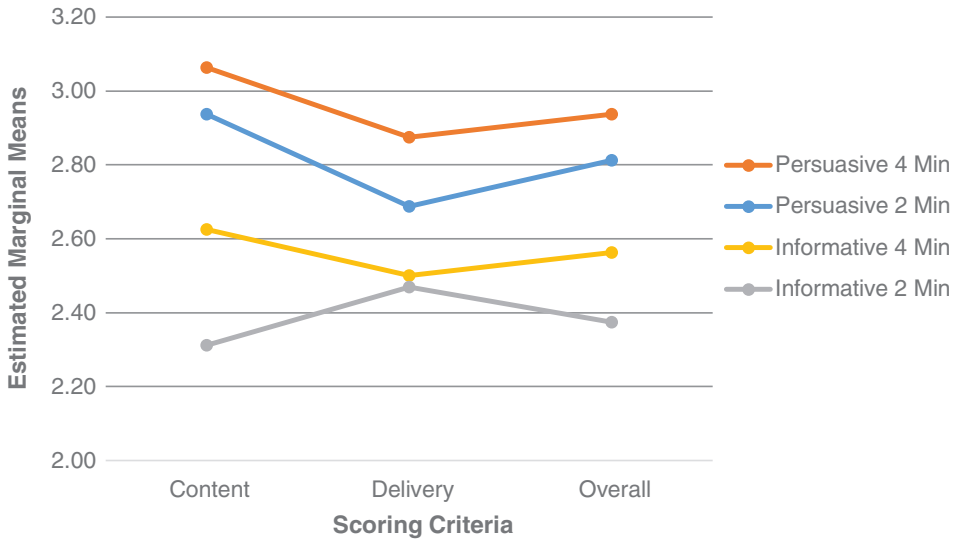

Figure 1 Average task scores across the three scoring criteria.

speech with no supporting materials), the content and delivery scores were not significantly correlated with each other. Similarly, the delivery dimension was not significantly correlated to the overall score (Table 8).

Relationships Between Scores and Other Measures

Results evaluating the relationship between overall scores and the two Likert-type item responses about students' perception of preparation time and task length showed no significant correlations. Additionally, there were no significant relationships between students' self-rating of oral communication skills and overall score across the four tasks, except for the 4-minute persuasive task, $\chi^2 = 16$, $p < .05$, which is notable for not having supporting materials. We also found no

Table 8 Spearman Correlations Between Scoring Criteria

Task	Content	Delivery	Overall
2-Minute persuasive			
Content	1.00	.74**	.93**
Delivery		1.00	.90**
Overall			1.00
4-Minute persuasive			
Content	1.00	.59	.91**
Delivery		1.00	.67
Overall			1.00
2-Minute informative			
Content	1.00	.83*	.94**
Delivery		1.00	.81*
Overall			1.00
4-Minute informative			
Content	1.00	.88**	.94**
Delivery		1.00	.98**
Overall			1.00

*$p < .05$. **$p < .01$.

significant Spearman correlations between total CCSR score and overall score across task types. Again, these nonsignificant results may have been a result of our small sample size for this study.

Evaluation of Preparation Time

Although the preparation time was intended to be standardized across all participants, technological difficulties and participant error contributed to some variation in the actual time participants had to prepare for the tasks, with some students having extra time to prepare. For instance, for the 4-minute persuasive task, students on average had more than 5 minutes of preparation time ($M = 5.58$ minutes, $SD = 1.34$). However, for the remaining three tasks, average preparation time closely matched the allotted time: 5 minutes for the 2-minute persuasive task and 7 minutes for the two informative tasks. As a result, we further investigated if there was any impact of this increased preparation time on student performance and students' perception of preparation time, finding that actual preparation time was not significantly correlated with participants' overall scores or with their perception of the appropriateness of the preparation time length.

We also reviewed the videos and screen captures to evaluate how participants used their preparation time. In most cases, participants moved linearly through the slides at a regular pace and took notes throughout, but a few participants moved back and forth through the slides, returning to earlier slides several times throughout. Although some students navigated back and forth, there were no clear patterns between this navigation process and student performance. Participants seemed to spend more time on slides that had more text on them, which was to be expected.

Discussion

The purpose of this study was to provide preliminary evidence on the different factors related to the design and scoring of oral communication tasks in higher education. For this study, we looked at two different oral communication task types (persuasive and informative speeches) of varying length (2 or 4 minutes) and evaluated interrater reliability and explained variance across raters, task type, task length, and scoring criteria; differences in scores; and relationships with other variables.

Interrater Reliability

In understanding how to develop tasks and rubrics to score oral communication tasks, we needed to evaluate the interrater reliability of scores. For this study, some of the lower interrater reliability estimates were not surprising, because this was the first time this rubric was used and we did not have training materials (e.g., benchmark samples, sample responses to

demonstrate various aspects of the rubric) and because the raters were new to this particular assessment context. Despite this, some of the dimensions did demonstrate higher interrater reliability estimates than expected (e.g., the delivery dimensions). Results for this study show that interrater reliability was much lower on the 2-minute tasks as compared to the 4-minute tasks, suggesting the need to improve the accuracy of scoring on shorter tasks. This could also mean that tasks need to be longer to see stable interrater reliability. It is important to note that the low interrater reliability on the 2-minute tasks may have impacted our results around performance differences and relations with other measures. As a result, we should use caution when interpreting these findings based on these rater scores. Additionally, results also show that in general, interrater reliability estimates were higher on the informative tasks as compared to the persuasive tasks. These results are consistent with results from two studies investigating the interrater reliability of the PSCR (Joe et al., 2015; Schreiber et al., 2012), a rubric used to evaluate public speaking performance in higher education. In both studies, interrater reliability estimates tended to be higher on the informative tasks as compared to the persuasive tasks, suggesting that there may be inherent differences in the tasks. It could be that it is more difficult to evaluate the persuasiveness of a test taker's performance. Additionally, these results could also indicate that there are deficiencies in the scoring rubric for persuasion and that modifications to the rubric and rater training may be necessary to increase interrater reliability.

To further investigate rater discrepancies, we evaluated the raters' rationale for their score choice. On the content dimension for the 2-minute persuasive speeches, the two raters had the most disagreement in regard to whether the speech was tailored to the intended audience. For instance, for one of the videos, Rater 1 indicated that the participant "seemed to adapt speech to school board," but Rater 2 noted that there was "little acknowledgment of audience." For the delivery dimension, there was some disagreement about eye contact and vocal variation but no consistency in this disagreement across the participant scores. Lastly, for the content dimension of the 2-minute informative speech, there was disagreement about how to penalize students for poor organization or providing misinformation from the supporting materials. In some cases, Rater 1 provided a higher score than Rater 2, despite the participant misinforming and including errors in fact. In other cases, Rater 2 provided a higher score for the same rationale, thus resulting in inconsistent scores. Disagreement on the delivery dimension was typically due to disagreement around the impact of verbal fillers and distracting mannerisms. These comments from the two raters are important for understanding where some of the language in the scoring rubric may need to be tightened to ensure higher consistency in scores. Future research should further evaluate these comments and make revisions to the scoring rubric to help increase interrater reliability. Additionally, interrater reliability should be further explored with a larger sample to see if longer tasks are more suitable when assessing oral communication. Future research would also benefit from further investigating the characteristics in the raters themselves (e.g., scoring experience, expertise in public speaking) to see if this affects scoring.

Interrater reliability results also show that consistency between raters was typically better when providing an overall score instead of a separate content or delivery score. Although the literature around oral communication assessment has suggested that tasks be scored both analytically and holistically (National Communication Association, 1998), it may be easier for raters to provide only one score instead of three for an individual task. Results from the G study show no variation in the three scoring criteria (content, delivery, and overall scores), which could also suggest that providing one overall score would be sufficient. However, because both raters scored on all three scoring criteria, this could be halo error. Additionally, MANOVA results indicate no significant differences between score types, and Spearman correlations between the two dimensions and with the overall score were high for most tasks. The only task that shows discrepancies was the 4-minute persuasive task, where we found that content and delivery scores were not significantly correlated and that the overall and delivery dimensions were also not significantly related. It is possible that because this speech did not have supporting materials, it was easier to capture the distinct differences across the content and delivery dimensions.

Although the statistical results indicated no variation or differences in the types of information that these three scores are providing within a task, institutions may still want to have information about a student's content performance as well as his or her delivery performance. Capturing the information holistically on overall performance would make it difficult to know students' strengths and weaknesses when delivering a speech. If we were to only capture information holistically on overall performance, we would want to make sure the rubric is comprehensive enough to clearly distinguish the types of skills needed to give a successful speech so that institutions could still obtain meaningful and actionable data.

Importance of Multiple Tasks on an Oral Communication Assessment

Results from this preliminary study can provide some initial insight into the development of an oral communication assessment. D study results show that a minimum of two tasks is necessary to achieve satisfactory reliability estimates. However, results from this study are mixed in terms of the amount of time needed to produce valid and reliable scores. Although only two tasks may be necessary on an oral communication assessment to achieve satisfactory reliability estimates, it is critical that an oral communication assessment have adequate construct coverage. This means that more than two tasks may be necessary. For instance, a persuasive task could be argumentative in nature to attempt to change a person's attitude about something. Also, an informative task could be a demonstrative task (e.g., showing an individual how to assemble a piece of furniture) or could simply provide information about an individual or a significant event. Additional task types may also be critical moving forward to measure different aspects of oral communication. As discussed earlier, the way in which students are presented with supporting materials, or asked to prepare supporting materials themselves, could expand the oral presentation construct. Tasks could also require students to stand to give a presentation over a webcam rather than be seated, which could capture upper body posture and hand gestures differently. Lastly, if we want to expand the oral communication construct more broadly, it may be important to measure students' interpersonal skills, which would include both speaking and listening. Future research will be needed to evaluate how to administer and score these types of tasks.

A secondary finding for this study was that there were no differences in performance between the 2- and 4-minute persuasive tasks and between the 2- and 4-minute informative tasks; however, there were significant differences in performance between the persuasive and informative tasks. This was also evident in the G study results, which showed very small variation in the task length (<1%) but some variation in the task type (∼6%). Although these results are dependent on the scores provided by raters, these results suggest that there may be differences in difficulty between persuasive and informative tasks or that the two tasks are capturing different information. These results support the need for multiple task types on an oral communication performance assessment.

Impact of Supporting Materials on Scores

Although we found no significant relationships between students' perceptions of the adequacy of preparation time and task length, it is worth noting that students performed best on the 4-minute persuasive task, the task that did not include supporting materials. This task, "Why should I go here?" asked students to think of a place they had visited that they thought their classmates would like or benefit from visiting. They were asked to list some reasons on their notecards and to prepare a speech convincing their classmates that they should go there. The open nature of this speech, combined with it being anchored in students' personal experience, seemed to make this task easier for students. The other three tasks required students to read through five to nine preparation slides and summarize that material in their speeches. Students performed best when they did not have to navigate through supporting materials but could rely simply on their own experience. Relatedly, we did not find relationships between students' self-rating of oral communication and any task score, except on this same task, $\chi^2(8) = 16$, $p < .05$. Additionally, we found significant Spearman correlations between score types for all tasks, except this task. It may be that this task was fundamentally different than the other three and perhaps allowed students to demonstrate their oral communication skills in a way that the other three do not because no supporting materials were included.

It is also important to note that while we found no significant correlations between preparation time and overall score, students did report that they felt the preparation time for several tasks was not appropriate. In particular, five of the students disagreed or strongly disagreed with the statement "The amount of time allowed for preparing my answer for the 'Tobacco History' task was appropriate." This was the 4-minute informative task, which had eight slides of information to synthesize in preparation for the speech. To further investigate this, we also looked at the relationship between actual preparation time use and their perception but did not find any significant relationships.

As we look toward a larger scale study, we may want to reconsider the amount of reading material in the preparation slides. Although it is important for students to demonstrate that they can synthesize available materials in their speech planning, we also want to ensure that we are measuring oral communication skills and not reading comprehension or speed. Without preparation slides, we may also see greater consistency between raters. To provide a score for the content dimension, for example, raters had to understand the supporting materials and take them into consideration when

providing a score to students. This involves large amount of synthesis for the rater and could explain some of the poorer consistency between the raters on that particular dimension. That said, part of the decision to retain or remove supporting materials should be based on whether navigating supporting materials is considered an important aspect of the oral communication construct.

Concurrent Validity Evidence

It is important to note that no significant relationships were found between student task performance and the CCSR. This finding may be due to the small sample size or because we had such small variation in test-taker scores, which led to a lack of correlation between performance and the CCSR. It may also be that the CCSR is not the best measure of the skills these tasks require and that the tasks in this study were measuring a different construct than was measured by the CCSR. While the CCSR contains items regarding organization of presented material, speaking clearly, and being able to express one's ideas, it also asks questions about the frequency of mispronouncing words, being able to tell the difference between fact and opinion, understanding oral directions, and being able to represent another's point of view. Future research should consider other related criteria for evaluating concurrent validity evidence, or we should consider revisions to the current assessment to better reflect the construct of oral communication.

Limitations and Future Research

One limitation to the current study is the small sample size; however, this was a preliminary study, which allowed us to test out the materials and gain valuable lessons as we plan for a larger scale study. Additionally, as the convenience sample was drawn from the pool of ETS summer interns, it was not a representative sample. It was overwhelmingly male, high achieving, and White or Asian. Another limitation is the number of tasks themselves. Although we made an effort to balance task types and provide prompts of interest to college students, these results may not be generalizable, and different tasks may result in different findings. That said, we were able to capture information on both informative and persuasive tasks and on tasks with and without supporting materials. Technological errors also impacted student preparation time; however, no relationships were found between actual preparation time used and student performance. Lastly, the fact that both raters scored on all three dimensions is another limitation, as it could have created halo error on the three scoring dimensions. Future research would benefit by having different raters make judgments about different dimensions. The order of the ratings should also be counterbalanced if the raters are to score on multiple dimensions in future studies.

We plan to build on the current study, expanding the study to include a more representative sample and collect additional data. We should consider the fact that different information may be captured by speeches with and without supporting materials and across informative and persuasive tasks. We also want to make revisions to the scoring rubric based on feedback from the raters in this study and work more closely with the raters to improve rater consistency. In particular, the scoring rubric made reference to the audience of the speech, which, due to the testing conditions, was hard to interpret. Raters also asked questions about whether failing to address one of the bullets under each scoring level was cause for reducing the overall score. Some of this feedback could explain some of the differences in consistency across the scores between the two raters. As we revise the scoring rubric, we will consult with the raters from this study to make improvements for overall clarity. We will also use some of the comments from the raters to guide the development of training materials to better clarify various aspects of the rubric.

We also need to revisit the relationship between preparation time and score. Although there were no significant relationships between students' reported perception of the adequacy of preparation time and score, some students reported that the preparation time was insufficient. Because Morae software allows us to track mouse movements during the video recording, we can examine these data along with the students' notecards for patterns that might help us understand the ways that students used their preparation time. Additionally, given the results of this preliminary study, for future research, we intend to revise the tasks and utilize Web-based technology to standardize the tasks and timing and attempt to mitigate any technological errors in preparation time. We plan to investigate how participants' use of preparation time is related to their scores in a more controlled environment and if we see the same differences in time usage between persuasive and informative tasks.

Lastly, we are continuing to explore the possibility of using automated scoring to score oral communication performance. Automated scoring may be able to capture information on nonverbal behaviors such as eye contact, smiling, and

gestures. Human scorers could then be used to score the content and verbal aspects of the speech. We will continue exploring the feasibility of utilizing automated scoring for these types of performance tasks.

Conclusion

This study provides preliminary insight into some of the design and scoring considerations when developing oral communication tasks in higher education. This study yields the following conclusions:

1. Interrater reliability estimates may vary depending on the length (2 or 4 minutes) or type (persuasive or informative) of task.
2. Despite scoring on three dimensions, G study results show no variation in the three scoring criteria.
3. A minimum of two tasks is necessary to achieve satisfactory reliability on an oral communication measure.
4. Supporting materials may impact test-taker performance on an oral communication measure.
5. Additional research with larger samples is needed to further investigate our research questions.

Although the sample size for this study was small, we will use the information from this study to further investigate some of the issues surrounding the development of an oral communication assessment. It is critical that we make revisions to our scoring rubric and further investigate the impact of student preparation time and supporting materials on performance. Overall, results suggest that we need to further investigate interrater reliability before we can identify a solid recommendation regarding the specific features (e.g., task length, task types) needed for an oral communication assessment.

Author Note

Kri Burkander is currently a research associate at Research for Action in Philadelphia, Pennsylvania.

References

Abdi, M., Eslami, H., & Zahedi, Y. (2012). The impact of pre-task planning on the fluency and accuracy of Iranian EFL learners' oral performance. *Procedia - Social and Behavioral Sciences, 69*, 2281–2288.

Adelman, C., Ewell, P., Gaston, P., & Schneider, C. G. (2014). *The degree qualifications profile 2.0: Defining U.S. degrees through demonstration and documentation of college learning*. Indianapolis, IN: Lumina Foundation.

Association of American Colleges and Universities. (2011). *The LEAP vision for learning: Outcomes, practices, impact, and employers' view*. Washington, DC: Author.

Binkley, M., Erstad, O., Herman, J., Raizen, S., Ripley, M., & Rumble, M. (2012). Defining 21st century skills. In P. Griffin, B. McGaw, & E. Care (Eds.), *Assessment and teaching of 21st century skills* (pp. 17–66). New York, NY: Springer Science and Business Media.

CAS Board of Directors. (2008). *Council for the advancement of standards: Learning and development outcomes*. Retrieved from http://standards.cas.edu/getpdf.cfm?PDF=D87A29DC-D1D6-D014–83AA8667902C480B

Casner-Lotto, J., & Barrington, L. (2006). *Are they really ready to work? Employers' perspectives on the basic knowledge and applied skills of new entrants to the 21st century U.S. workforce*. New York, NY: Conference Board/Partnership for 21st Century Skills/Corporate Voices for Working Families/Society for Human Resource Management.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum.

Crick, J. E., & Brennan, R. L. (1983). *A Generalized Analysis of Variance System* (Version 2.1) [Computer software]. Iowa City, IA: American College Testing Program.

Employment Training Administration. (2014). *Competency model clearinghouse: Communication—Listening and speaking*. Washington, DC: U.S. Department of Labor. Retrieved from http://www.careeronestop.org/COMPETENCYMODEL/blockModel.aspx?tier_id=2&block_id=11

Fleiss, J. L., & Cohen, J. (1973). The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. *Educational and Psychological Measurement, 33*, 613–619.

Hart Research Associates. (2015). *Falling short? College learning and career success*. Washington, DC: Association of American Colleges and Universities.

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. https://doi.org/10.1002/ets2.12083

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33*, 159–174.

Li, L., Chen, J., & Sun, L. (2015). The effects of different lengths of pretask planning time on L2 learners' oral test performance. *TESOL Quarterly, 49*(1), 38–66.

Malabonga, V., Kenyon, D. M., & Carpenter, H. (2005). Self-assessment, preparation, and response time on a computerized oral proficiency test. *Language Testing, 22*(1), 59–92.

McClellan, C. A. (2010). Constructed-response scoring: Doing it right. *R&D Connections, 13*, 1–7. Retrieved from https://www.ets.org/Media/Research/pdf/RD_Connections13.pdf

Morreale, S. P., Backlund, P. M., Hay, E. A., & Jennings, D. K. (2007). *Large scale assessment in oral communication P–12 and higher education* (3rd ed.). Washington, DC: National Communication Association.

Morreale, S. P., Moore, M., Surges-Tatum, D., & Webster, L. (2007). *The Competent Speaker speech evaluation form*. Washington, DC: National Communication Association.

Morreale, S. P., Rubin, R. B., & Jones, E. (1998). *Speaking and listening competencies for college students*. Washington, DC: National Communication Association.

National Communication Association. (1998). *Guidelines for communication assessment*. Washington, DC: Author.

National Governors Association Center for Best Practices & Council of Chief State School Officers. (2010). *Common Core State Standards for English language arts and literacy in history/social studies, science, and technical subjects*. Washington, DC: Author.

National Institutes of Health. (2017). *Communications*. Bethesda, MD: Office of Human Resources Retrieved from https://hr.nih.gov/competency/communications

Quality Assurance Agency. (2008). *The framework for higher education qualifications in England, Wales, and Northern Ireland: August 2008*. Mansfield, England: Author.

Rhodes, T. L. (Ed.). (2010). *Assessing outcomes and improving achievement: Tips and tools for using rubrics*. Washington, DC: Association of American Colleges and Universities.

Roohr, K. C., Mao, L., Belur, V., & Liu, O. L. (2015, April). *Oral communication in higher education: Existing research and future directions*. Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago, IL.

Ross, S. (1998). Self-assessment in second language testing: A meta-analysis and analysis of experiential factors. *Language Testing, 15*, 1–20.

Rubin, R. B. (1985). The validity of the communication competency assessment instrument. *Communication Monographs, 52*, 173–185.

Schreiber, L. M., Paul, G. D., & Shibley, L. R. (2012). The development and test of the public speaking competency rubric. *Communication Education, 61*, 205–233.

Seltman, H. J. (2014). *Experimental design and analysis*. Retrieved from http://www.stat.cmu.edu/~hseltman/309/Book/Book.pdf

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

Techsmith. (2015). *Morae recorder and manager* [Computer software]. Okemos, MI: Author.

Wigglesworth, G., & Elder, C. (2010). An investigation of the effectiveness and validity of planning time in speaking test tasks. *Language Assessment Quarterly, 7*(1), 1–24.

Wrench, J. S., Goding, A., Johnson, D. I., & Attias, B. A. (2012). *Public speaking: Practice and ethics* (Version 1.0). Retrieved from http://2012books.lardbucket.org/pdfs/public-speaking-practice-and-ethics.pdf

Appendix A Scoring Rubric for Informative Speech Task

The Informative Speech Task will be scored on a 4- two-trait scale. Please provide one score for content and one for delivery using the rubrics below.

Content

4 (Advanced)

In addressing the specific task, a speech that earns a score of 4 will present all relevant material in a clear, well-organized speech. A speech in this category will

- contain an introduction that clearly indicates the topic, establishes credibility, and presents a clear preview of the main points of the speech
- tailor the contents of the speech so as to effectively inform the intended audience
- show outstanding organization, with the main point clearly delineated and clearly related to the thesis; display effective transitions and signposting
- conclude in a memorable and appropriate way, either summarizing or ending with a strong idea

- convey ideas fluently and precisely, using effective vocabulary and sentence variety
- demonstrate superior facility with the conventions of standard written English (i.e., grammar, usage, and mechanics) but may have minor errors

3 (Proficient)

In addressing the specific task, a speech that earns a score of 3 will present mostly relevant material in a way that the listener can easily understand. A typical speech in this category will

- contain an introduction that provides some orientation to the topic and allows the listener to understand what the speech will cover
- adapt the contents of the speech as necessary to inform the intended audience
- show a clear organization, with the main points apparent; display transitional statements and signposting for the main points, but may contain some abrupt transitions
- conclude with an appropriate summary of the main points
- demonstrate sufficient control of language to convey ideas with acceptable clarity
- generally demonstrate control of the conventions of standard written English but may have some errors

2 (Basic)

In addressing the specific task, a speech that earns a score of 2 will be less than organized in its presentation of material and will leave the listener less than informed. A typical speech in this category may

- contain an introduction that provides little insight into what the speech will cover
- show limited awareness of the intended audience when determining what points to cover
- show a limited attempt at an organizational structure, with the main points presented but not in a logical fashion; display few transitional statements that may not be effective
- conclude with some summary of points but no clear connection to the thesis or a big idea/call to action
- have problems in language and sentence structure that result in a lack of clarity
- contain occasional major errors or frequent minor errors in grammar, usage, or mechanics that can interfere with meaning

1 (Minimal)

In addressing the specific task, a speech that earns a score of 1 will generally not present relevant material in a consistently organized way. A typical speech in this category may

- contain very little introduction, instead presenting an irrelevant opening that abruptly jumps into the body of the speech
- show no awareness of the reason for delivering the speech
- present some relevant information, but is not logically organized; display mostly awkward or incorrect transitions
- trail off or give an inappropriate and unclear conclusion that may not relate to the rest of the speech
- have problems in language and sentence structure that frequently interfere with meaning
- contain serious errors in grammar, usage, or mechanics that frequently obscure meaning

0 (Deficient)

In addressing the specific task, a speech that earns a score of 0 will generally not be organized or understandable. A typical speech in this category may

- contain no introduction or preview of the main points
- present very little relevant information in a random fashion; display no transitional statements
- give no conclusion, ending abruptly
- have severe problems in language and sentence structure that persistently interfere with meaning

- contain pervasive errors in grammar, usage, or mechanics that result in incoherence

Delivery

4 (Advanced)

In delivering the material, a speech that earns a score of 4 will exhibit outstanding verbal and nonverbal behaviors that reinforce the message. A typical speech in this category will

- display superior vocal variation in both the intensity/volume and the pacing of the speech; project enthusiasm and avoid verbal fillers
- exhibit confidence and poise with posture, gestures, and facial expressions that are appropriate to the topic; maintain eye contact with the audience
- pay attention to the specific demographic features, beliefs, attitudes, and values of the audience and adapt the speech accordingly
- adapt appropriately to the specific environment in which the speech is being given (in this case, over a webcam)

3 (Proficient)

In delivering the material, a speech that earns a score of 3 will exhibit verbal and nonverbal behaviors that reinforce the message. A typical speech in this category will

- display good vocal variation in both the intensity/volume and the pacing of the speech; project interest and generally avoid verbal fillers
- generally exhibit confidence with posture, gestures, and facial expressions that are appropriate to the topic; spend the majority of the time looking into the camera instead of looking down
- pay some attention to the specific demographic features of the audience and attempt to adapt the speech accordingly
- show awareness of the specific environment in which the speech is being given (in this case, over a webcam) and try to adapt to it

2 (Basic)

In delivering the material, a speech that earns a score of 2 will exhibit adequate verbal and nonverbal behaviors. A typical speech in this category may

- demonstrate some attempt to vary speech and speak audibly and clearly; use fillers that are frequent but do not overly detract from the message
- overtly refer to notes in a way that detracts from the message but generally avoid distracting mannerisms
- show minimal attention to the audience and why the topic is important to it
- not adapt to the specific environment in which the speech is being given (in this case, over a webcam)

1 (Minimal)

In delivering the material, a speech that earns a score of 1 will exhibit verbal and nonverbal behaviors that need improvement. A typical speech in this category may

- speak too softly or indistinctly for the listener to hear comfortably; use distracting fillers and pauses
- look down at notes and/or speak in a stiff and unnatural style
- make no attempt to adapt the speech to the audience
- make no attempt to adapt to the specific environment in which the speech is being given (in this case, over a webcam)

0 (Deficient)

In delivering the material, a speech that earns a score of 0 will exhibit distracting or inappropriate verbal and nonverbal behaviors. A typical speech in this category may

- speak inaudibly, enunciate poorly, or speak in a monotone; use distracting fillers and pauses
- exhibit nervousness or other distracting nonverbal behaviors or exhibit behaviors that contradict the message
- be delivered with no attempt to establish common ground with the audience or be contrary to the audience beliefs, values, and attitudes
- ignore the specific environment in which the speech is being given (e.g., for a webcam, be out of the frame of the picture)

Appendix B Scoring Rubric for Persuasive Speech Task

The Persuasive Speech Task will be scored on a 4-point, two-trait scale. Please provide one score for content and one for delivery using the rubrics below.

Content

4 (Advanced)

In addressing the specific task, a speech that earns a score of 4 will present all relevant material in a clear, well-organized speech. A speech in this category will

- contain an introduction that clearly indicates the topic, establishes credibility, and presents a clear thesis
- tailor the contents of the speech so as to effectively persuade the intended audience
- show outstanding organization, with the main points clearly delineated and clearly supporting the thesis; display effective transitions and signposting
- conclude in a memorable and appropriate way, either summarizing or ending with a strong idea or call to action
- convey ideas fluently and precisely, using effective vocabulary and sentence variety
- demonstrate superior facility with the conventions of standard written English (i.e., grammar, usage, and mechanics) but may have minor errors

3 (Proficient)

In addressing the specific task, a speech that earns a score of 3 will present mostly relevant material in a way that the listener can easily understand. A typical speech in this category will

- contain an introduction that provides some orientation to the topic and allows the listener to discern a thesis
- adapt the contents of the speech to persuade the intended audience
- show a clear organization, with the main points apparent; display transitional statements and signposting for the main points but may contain some abrupt transitions
- conclude with some reference back to the thesis or a strong idea/call to action
- demonstrate sufficient control of language to convey ideas with acceptable clarity
- generally demonstrate control of the conventions of standard written English but may have some errors

2 (Basic)

In addressing the specific task, a speech that earns a score of 2 will be less than organized in its presentation of material and will leave the listener less than informed. A typical speech in this category may

- contain an introduction that does not fully explain the purpose of the speech
- show little acknowledgment of the intended audience in deciding what points to make
- show a limited attempt at an organizational structure, with the main points presented but not in a logical fashion; display few transitional statements that may not be effective
- conclude with some summary of points but no clear connection to the thesis or a big idea/call to action
- have problems in language and sentence structure that result in a lack of clarity

- contain occasional major errors or frequent minor errors in grammar, usage, or mechanics that can interfere with meaning

1 (Minimal)

In addressing the specific task, a speech that earns a score of 1 will generally not present relevant material in a consistently organized way. A typical speech in this category may

- contain very little introduction, instead presenting an irrelevant opening that abruptly jumps into the body of the speech
- show no discernible intended audience
- present some relevant information but is not logically organized; display mostly awkward or incorrect transitions
- trail off or give an inappropriate and unclear conclusion that may not relate to the rest of the speech
- have problems in language and sentence structure that frequently interfere with meaning
- contain serious errors in grammar, usage, or mechanics that frequently obscure meaning

0 (Deficient)

In addressing the specific task, a speech that earns a score of 0 will generally not be organized or understandable. A typical speech in this category may

- contain no introduction or preview of the main points
- present very little relevant information in a random fashion; display no transitional statements
- give no conclusion, ending abruptly
- have severe problems in language and sentence structure that persistently interfere with meaning
- contain pervasive errors in grammar, usage, or mechanics that result in incoherence

Delivery

4 (Advanced)

In delivering the material, a speech that earns a score of 4 will exhibit outstanding verbal and nonverbal behaviors that reinforce the message. A typical speech in this category will

- display superior vocal variation in both the intensity/volume and the pacing of the speech; project enthusiasm and avoid verbal fillers
- exhibit confidence and poise with posture, gestures, and facial expressions that are appropriate to the topic; maintain eye contact with the audience
- pay attention to the specific demographic features, beliefs, attitudes, and values of the audience and adapt the speech accordingly
- adapt appropriately to the specific environment in which the speech is being given (in this case, over a webcam)

3 (Proficient)

In delivering the material, a speech that earns a score of 3 will exhibit verbal and nonverbal behaviors that reinforce the message. A typical speech in this category will

- display good vocal variation in both the intensity/volume and the pacing of the speech; project interest and generally avoid verbal fillers
- generally exhibit confidence with posture, gestures, and facial expressions that are appropriate to the topic; spend the majority of the time looking into the camera instead of looking down
- pay some attention to the specific demographic features of the audience and attempt to adapt the speech accordingly
- show awareness of the specific environment in which the speech is being given (in this case, over a webcam) and try to adapt to it

2 (Basic)

In delivering the material, a speech that earns a score of 2 will exhibit adequate verbal and nonverbal behaviors. A typical speech in this category may

- demonstrate some attempt to vary speech and speak audibly and clearly; use fillers that are frequent but do not overly detract from the message
- overtly refer to notes in a way that detracts from the message but generally avoid distracting mannerisms
- show minimal attention to the audience and why the topic is important to it
- not adapt to the specific environment in which the speech is being given (in this case, over a webcam)

1 (Minimal)

In delivering the material, a speech that earns a score of 1 will exhibit verbal and nonverbal behaviors that need improvement. A typical speech in this category may

- speak too softly or indistinctly for the listener to hear comfortably; use distracting fillers and pauses
- look down at notes and/or speak in a stiff and unnatural style
- make no attempt to adapt the speech to the audience
- make no attempt to adapt to the specific environment in which the speech is being given (in this case, over a webcam)

0 (Deficient)

In delivering the material, a speech that earns a score of 0 swill exhibit distracting or inappropriate verbal and nonverbal behaviors. A typical speech in this category may

- speak inaudibly, enunciate poorly, or speak in a monotone; use distracting fillers and pauses
- exhibit nervousness or other distracting nonverbal behaviors or exhibit behaviors that contradict the message
- be delivered with no attempt to establish common ground with the audience or be contrary to the audience beliefs, values, and attitudes
- ignore the specific environment in which the speech is being given (e.g., for a webcam, be out of the frame of the picture)

Suggested citation:

Roohr, K. C., Burkander, K., & Mao, L. (2018). *A preliminary investigation of the factors related to the design and scoring of video-based oral communication performance tasks in higher education* (Research Report No. RR-18-09). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/ets2.12197

Action Editor: Keelan Evanini

Reviewers: Lawrence Davis and Michelle Martin-Raugh