# A Proof-of-Concept Study on Scoring Oral Presentation Videos in Higher Education

ETS RR–19-22

Gary Feng
Jilliam Joe
Christopher Kitchen
Liyang Mao
Katrina Crotts Roohr
Lei Chen

*December 2019*

# 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 Proof-of-Concept Study on Scoring Oral Presentation Videos in Higher Education

Gary Feng, Jilliam Joe, Christopher Kitchen, Liyang Mao, Katrina Crotts Roohr, & Lei Chen

Educational Testing Service, Princeton, NJ

This proof-of-concept study examined the feasibility of a new scoring procedure designed to reduce the time of scoring a video-based public speaking assessment task. Instead of scoring the video in its entirety, the performance was evaluated based on content-related (e.g., speech organization, word choice) and delivery-related (e.g., vocal expression, nonverbal behaviors) dimensions. Content-related dimensions were scored based on speech transcripts, while delivery dimensions were scored using a video thin-slicing technique, where scores were assigned based on brief vignettes of a video rather than the complete performance. Initial feasibility data were collected from 4 novice raters scoring 10 video performances. Results indicated that for delivery dimensions, four 10 second thin slices yielded interrater consistency reliability somewhat similar to that of full-video scoring, and additional slices resulted in only small improvements. For transcription-based scoring, while raters were consistent, their scores had low correlations with criterion scores, which was likely due to differences in the modality (video vs. text). Video thin-slicing appears to be a promising scoring technique for relevant constructs. Further testing of a combination of audio and transcript is recommended for scoring content-related constructs.

**Keywords** Oral communication assessment; higher education; human scoring; multimodal assessment; public speaking

doi:10.1002/ets2.12256

Oral communication skills are consistently rated by university administrators (e.g., Association of American Colleges and Universities, 2011; Educational Testing Service, 2013) and employers (e.g., Casner-Lotto & Barrington, 2006; Hart Research Associates, 2015) as critical skills to be successful academically and professionally. To date, a number of existing oral communication performance-based assessments in higher education can be used to measure public speaking either in person or using a video such as the Competent Speaker Speech Evaluation Form (Morreale, Moore, Surges-Tatum, & Webster, 2007), the oral communication VALUE (Valid Assessment of Learning in Undergraduate Education) rubric from the Association of American Colleges and Universities (Rhodes, 2010), or the Public Speaking Competence Rubric (PSCR; Schreiber, Paul, & Shibley, 2012). A challenge with many of these assessments is the amount of time it can take to score presentations using these rubrics. For instance, Dunbar, Brooks, and Kubicka-Miller (2006) found that it took a considerable amount of time to establish acceptable interrater reliability prior to scoring prerecorded speeches using the Competent Speaker Speech Evaluation Form and found that because the assessment was administered to students at the end of the semester, it created a significant burden on faculty to score.

Balancing the cost and quality of performance scoring is not an easy feat. In a study designed to maximize quality, Joe, Kitchen, Chen, and Feng (2015) asked human raters to watch videos of public speaking performances and score at their own pace (i.e., they were allowed to rewind and review the video as much as they needed). With a rubric consisting of 10 content- and delivery-related dimensions (developed by Schreiber et al., 2012), the raters took three to four times real time to score. At this rate, a 5-minute performance task would take up to 30 minutes to double-score, not counting the time to train and monitor the human rater performance. In a large-scale administration context, the amount of time needed to score would be extremely burdensome to raters (especially if the raters were faculty). Additionally, the cost corresponding to that amount of time to score would be unrealistic.

On the other hand, in some large-scale assessments, operational scoring procedures only allow the rater a single, continuous viewing of the performance (i.e., no stop, repeat, or rewind) before scoring. While this controls the time and cost, the operational procedure relies heavily on the rater's memory of the performance, leaving it vulnerable to various scoring biases. Working memory overload is one concern for performances lasting minutes. Another potential bias is that when the rater has to rely on a holistic impression while also analytically scoring different constructs (e.g.,

*Corresponding author:* Katrina Crotts Roohr, E-mail: kroohr@ets.org

on content and delivery dimensions), scores from one dimension may bleed over to another, artificially heightening the correlations among dimensional scores. Such effects would be difficult to detect by looking at interrater consistency measures if raters were similarly affected. In fact, in this case, raters could appear very consistent and the assessment very reliable because different constructs could fall on a single dimension. However, if part of the validity claim is that different aspects of oral communication are assessed analytically (e.g., that subscores may be used for feedback or even training purposes), then we need to carefully examine the human scoring process to guard against such biases.

In addition to scoring time, individual differences in raters also introduce much variance in the scores. The interrater consistency of performance scoring depends heavily on the rubric, with low to medium intraclass correlations (ICCs) frequently observed (Schreiber et al., 2012). One solution is to have multiple raters per performance. For example, in the context of medical competency assessment, Williams, Klamen, and McGaghie (2003) recommended 7–11 raters. Given the time required for traditional video scoring procedures (i.e., watching the entire video at least once), a high number of raters may be unsustainable for an assessment of oral communication. As a result, we sought a method that allows scoring by multiple raters without substantially increasing scoring time.

To summarize, the challenge is to maintain the validity and reliability of human scoring while keeping its cost at an affordable level for higher education institutions that may be paying faculty to score oral communication assessments or using outside resources for scoring. Thus this proof-of-concept study intends to investigate the use and feasibility of technology to assist human scoring on a public speaking assessment in higher education. Specifically, we divided the public speaking construct into content and delivery dimensions. We proposed scoring content dimensions on the basis of speech transcripts and delivery dimensions using a video thin-slicing technique.

## Proof-of-Concept Versus Validation

An important caveat is that this is a proof-of-concept study, not a validation study. Confusing the two can lead to serious consequences. Whereas validation studies examine the quality and fairness of developed assessment products or procedures, a proof-of-concept study aims to demonstrate the feasibility of an ideation, often with extremely limited time and resources. The solution developed in a proof-of-concept is often unpolished, and the condition under which it is tested differs significantly from the operational application. Data from the present report should not be used to evaluate the efficacy of the new methods. The question of interest here is not how well the new methods compare to standard scoring procedures (that would require a validation study); rather, the questions we try to answer are whether this new proposed method is feasible and whether it shows promise that warrants further investigation.

In designing this proof-of-concept study, we focused on the feasibility of the scoring methodology for video-based performances from a public speaking assessment to reuse materials from our previous research and to limit data collection to a minimum. Hence data from this study do not say much about how the method will perform in an operational large-scale assessment setting. Instead, this report provides a rapid initial evaluation of the potentials of the new methods, with all the caveats about their limitations.

## Technology-Assisted Scoring

To reduce human scoring time while maintaining reliability and validity, and, we hope, allowing for distributed scoring to increase the number of ratings per performance, our approach is divide and conquer. We introduced two techniques that have been used separately in assessment: (a) *transcription-based* scoring (e.g., Leung & Mohan, 2004) focusing on content-related constructs and (b) *video thin-slicing*, allowing for rapid scoring of delivery-related constructs (Ambady, Krabbenhoft, & Hogan, 2006; Ambady & Rosenthal, 1993).

### Transcription-Based Scoring

There are two motivations for scoring transcriptions of a speech, instead of videos of the speaker engaged in delivering the speech. First, we read faster than we can speak. A 5-minute speech is likely to contain about 500–750 words, which takes a skilled reader about 2–3 minutes to read if the transcription is available (Carver, 1992). Second, focusing on the

written record only can potentially circumvent some of the issues with video scoring, such as conscious or unconscious biases against racial or linguistic characteristics of the candidate or undue influences across dimensions when scoring relies on the memory of a video performance.

A potential risk, though, is that scoring speech transcripts may underestimate the communicative effectiveness because we have removed the support from other modalities, such as the vocal quality and nonverbal behaviors (e.g., gestures, facial expression, eye contact). To what extent content-related constructs can be separated from delivery-related constructs is a conceptual issue that should warrant further research. However, because this is a proof-of-concept study, we simply chose to investigate the transcript-scoring approach to see if it works at all.

We used minimal technology for transcript scoring in this study for two reasons: (a) It is very time consuming to develop sophisticated software or interface support and (b) if simple reading works, it will be good news for cost reduction. For this study, speech transcriptions were obtained for the target videos using an online human transcription service (see Joe et al., 2015), and raters received hard copies of transcriptions for scoring.

Although the approach in this study uses minimal technology for transcription, in future operational applications, several technologies may be necessary.

- *Automatic speech recognition* (*ASR*): The audio track in the video recording will go through ASR to obtain a raw transcription. This is routinely done in video services, such as YouTube, with low cost and high efficiency. The quality of the transcript depends on the speaker, the topic, and the quality of recording. ASR can be optimized for a particular prompt as data accumulate. Human proofreading is also a possibility if resources allow; this may be crowd-sourced to further lower the cost.
- *Synchronization between audio and transcription*: Speech transcripts can be automatically synchronized with the audio/video at the sentence or word level, using force-alignment algorithms. Then, using Web technologies such as Synchronized Multimedia Integration Language (SMIL; see later), we can highlight words or sentences in the transcriptions as the audio/video is played, or conversely, we can present the speech transcription such that clicking on a word will start playing the audio/video from that position. This level of flexibility affords support for the rater to locate and evaluate critical evidence for scoring.
- *Variable-speed playback* (*VSP*): In the case audio/video streams will be played along with the transcript, an important functionality for time/cost saving is the ability to play back the audio/video streams at a higher speed. VSP has been widely used in modern media players. In particular, the phase vocoder technology allows intelligible playback of speech signals (e.g., to play audio at 2 – 4 times real time) with little change in the pitch of the speech. There is evidence that at modest (1.5 times) to fast (2.0 times) speed, listeners can maintain comprehension (Ritzhaupt, 2008). VSP of video is also widespread; YouTube, for example, includes the functionality to play at up to 4 times speed for videos encoded in certain codecs. Our informal testing showed that for a typical presentation in which the speaker uses only a modest amount of body gestures, playing back at 2 times felt quite normal. We did not test the synchronized transcript – audio/video playback or the VSP technologies in this study.

## Video Thin-Slicing

The idea of thin-slicing is to present the rater with a sample of vignettes or clips from the video, instead of the whole video. It is motivated by two observations: (a) Most speech acts are informationally redundant and (b) we can form a rapid and often accurate assessment of a speaker based on a sliver of observation (Ambady et al., 2006). The literature of social psychology, for example, has shown that approximately 30 seconds of observation can predict psychological adjustments after divorce (Mason, Sbarra, & Mehl, 2010) and racial biases (Richeson & Shelton, 2005). Thirty seconds or less (down to 6 seconds in some studies) of exposure can predict ratings of teachers by students and by principals (Ambady & Rosenthal, 1993). There is even evidence that intuitive impressions formed based on thin-slicing outperform deliberate decisions in lie detection (Albrechtsen, Meissner, & Susa, 2009). We hypothesize that video thin-slicing can reduce scoring time without significant loss in scoring quality. Several practical issues remain to be resolved, including total exposure time, slice duration, slicing units, and order of presentation.

### *Total Exposure Time*

The literature has suggested that time as short as 20 – 40 seconds is adequate for impression forming (Ambady & Skowronski, 2008). On the basis of the literature, we planned several approximately 40-second scoring blocks to see whether scores change as a function of additional exposure. Raters were asked to rate the performance after each 40-second block. Our hypothesis is that after one or two blocks, the quality of rating will reach the asymptote and additional slices will give diminishing returns.

### *Slice Duration*

Five- to 10-second slices have been reported in the literature (e.g., Ambady & Skowronski, 2008). In our experiments, 5-second slices appeared to be too short for raters to comprehend the speech. The 10-second window was generally adequate.

### *Slicing Units*

While the simplest slicing method would be strictly time based (e.g., to obtain 10-second slices from predetermined starting times), such an approach likely will result in slices containing fractions of utterances. An alternative is to ensure that each slice is semantically meaningful and linguistically intact (i.e., each slice is a sentence). Random slicing is easy to implement without much cost, whereas a sentence-based method requires technologies for detection of sentence boundaries, which can be done using speech transcripts and force-alignment algorithms.

### *Order of Presentation*

Presenting the slices in the order they occurred is a natural choice as it preserves somewhat the structure of the performance. However, we had to balance it with another goal of this proof-of-concept study, which is to determine how many slices are adequate to give quality ratings. Given the small number of raters ($n = 4$) in this study, we were limited to a within-subject design, where we asked raters to score after each block of slices to see whether the rating quality increased over a number of blocks. In this case, the slices in each block needed to be a representative sample of the performance as a whole; otherwise, the ratings after each block would be biased. This left us no choice but to randomize the order of slices. In future operational applications where the number of slices is predetermined, a sequential presentation would probably be preferred.

## Study Purpose and Research Questions

Based on a desire for cost and time savings when scoring video-based presentations in a higher education context, the purpose of this current proof-of-concept study was to test the feasibility of two innovative human scoring methods for video-based oral communication tasks: transcript scoring and video thin-slicing. Content-related constructs (e.g., speech organization, word choice) were scored based on the speech transcript. Delivery constructs (e.g., vocal expression, nonverbal behaviors) were scored using video thin-slicing. Specifically, we compared two slicing methods: random slices (of 10 seconds) versus sentence slices. The order of presentation for the video thin slices was randomized with three blocks of video thin slices presented in sequence, each containing four randomly sampled slices. Ratings of delivery constructs were obtained after each block to determine the point of diminishing return. For this study, we focused on developing the procedure and technologies by answering the following research questions:

1. What is the interrater reliability across scoring dimensions within each condition?
2. Are raters' scores using the new method (transcript and video thin-slicing) consistent with the criterion scores from Joe et al. (2015)?
3. How much time do raters need in these new scoring conditions to produce reliable scores?
4. How do the confidence scores change over time using thin-slice scoring?

Because this is a proof-of-concept study, the data will not provide definitive answers to these questions. Data here cannot determine whether these methods can be used in operational scoring. Instead, our goal is to identify evidence on whether — and which part of — the new scoring methods show enough promise to warrant further investigation.

# Method

## Participants, Setting, and Videos

The raters involved in this study included four employees (research associates) from ETS's R&D department. No person serving as a rater had prior formal experience with evaluating or scoring public speaking performances in professional or academic settings. Several raters did have prior experience with employing performance-based scoring rubrics or providing textual annotations for essays and aural speech files, however.

Raters met with project staff on two occasions: first for training in the use of the PSCR (see Schreiber et al., 2012) with an orientation to the scoring format and second for the scoring session itself. Each session lasted roughly 2 hours. Raters completed both the training and scoring in a large workspace that had several cubical-like workstations, each with its own personal computer and desk space. Scoring files and the AMBULANT media player (discussed later) used for presenting video clips were loaded onto each workstation computer for the purpose of this study.

Additionally, video clips for this study involved a subset of video data from previous research on the multimodal analysis of oral communication (Chen et al., 2013). A subset of the videos had been double-scored by experienced scorers with no time constraints (see Joe et al., 2015) according to a published rubric (see Schreiber et al., 2012). Ten video responses (2 – 5 minutes long each) with double scores were sampled to be included in this study. Videos were selected if they demonstrated high interrater reliability between the two raters (based on results from Joe et al., 2015). The main reason for selecting videos with high interrater reliability was to better control any variance due to items, especially given that there would be variation in the raters (because we used different raters to score the videos in this study). Selecting videos that have demonstrated high agreement also allowed us to evaluate whether this same level of agreement can be obtained using the proposed transcript and thin-slicing methodology. The main limitation in selecting videos with high interrater reliability was the fact that we could have regression to the mean, where the interrater reliability in subsequent ratings would likely be lower than the original reliability.

## Rater Training

Training consisted of several discrete objectives. First, a brief overview of the study purpose was shared with raters. Raters were informed that they would be providing scores for both transcripts and video clips using the PSCR scoring dimensions. They were also informed that their scores would be compared to a more robust scoring method that was completed as part of a different project (Joe et al., 2015) and that the reliability of their scores would be examined. Participants were not informed of the inferential tests being used to determine reliability.

The second objective was to thoroughly review the content of the PSCR (Schreiber et al., 2012) and discuss nuances encountered by previous raters for this same corpus of data (Joe et al., 2015), including instances where no conclusion was present in a performance due to the participant being cut off by the experimenter, absence of visual aids in some tasks, lack of persuasive elements in some tasks, and differentiation between adaptation to audience, vocal expression, and word choice. The process of reviewing the content of the PSCR was the same in this study as it was in Joe et al. (2015). The PSCR content, scoring dimensions, and evidence fragments for each dimension were discussed in a direct sequence (e.g., Dimension 1, "introduction"; Dimension 2, "organization"), though some differentiation between later scoring dimensions was required to underline where common misconceptions may arise in evidence identification. The intent for reviewing the PSCR was to clearly delineate each dimension to be as exclusive from the others as possible and provide adequate supporting information so that raters knew what to consider as evidence for each dimension. This more conversational segment of the training yielded some important clarifications and elaboration on use of the rubric.

Last, we used video exemplars identified as part of the scoring work from previous research (Joe et al., 2015) in an effort to demonstrate how scores are derived for each scoring dimension using available evidence in video clips of various performances. For each scoring dimension, we presented at least one example of a performance that constituted high or low quality, and raters discussed scoring rationales and evidentiary content and drew consensus on each example presented as a group. These discussions were facilitated by an experimenter, who was present during both the training and scoring sessions. Upon completion of review of the set of exemplars, the group completed one full example of transcript scoring and video scoring (using the random thin-slicing method discussed later). As a group, each rater's scores and scoring rationales were reviewed collectively to build consensus and help reveal any potential lingering misconceptions in implementing the scoring rubric.

**Table 1** Content Scoring Dimensions for Performance Transcripts From the Public Speaking Competence Rubric

| Scoring dimension | Corresponding PSCR performance standard[a] |
|---|---|
| Introduction | Formulates an introduction that orients audience to topic and speaker |
| Organization | Uses an effective organizational pattern |
| Conclusion | Develops a conclusion that reinforces the thesis and provides psychological closure |
| Word choice | Demonstrates a careful choice of words |
| Persuasion | Constructs an effectual persuasive message with credible evidence and sound reasoning |

*Note*. PSCR = Public Speaking Competence Rubric.
[a]See Schreiber et al. (2012, pp. 228 – 231) for the full rubric.

At the end of the training session, raters indicated that they felt the review of the tasks, PSCR scoring instructions, identifying evidence through use of exemplars, and review of the full examples using the approaches outlined here were adequate to allow correct implementation of the scoring rubric. One rater, however, requested additional reviewing of the scoring rubric in a separate session with the experimenter rather than as a group. It was noted that this same rater did not ask any questions about the tasks or PSCR during the main training with the full group, and it is believed that he or she may have felt uncomfortable asking questions or asking for clarifications while in a group setting.

## Scoring

Two modalities were pursued for scoring performances: (a) transcription to score content features of a performance and (b) the recorded video to score delivery features of a performance. This approach constrains what scoring dimensions are applicable to the information being provided to the rater. For instance, we can hardly expect a rater to give an evaluation of vocal expression when the acoustic features are absent from the performance transcript. Word choice, on the other hand, would be applicable for evaluation using performance transcripts. Rating forms for this study can be found in Appendix A.

### *Transcript Scoring Using the Public Speaking Competence Rubric*

The PSCR dimensions relevant to transcript scoring are introduction, organization, conclusion, persuasion, and word choice. A brief description for each scoring dimension is provided in Table 1, and it was hypothesized that each of these dimensions can be adequately evaluated using only the speech content provided as text. Raters were instructed to pay attention to various aspects of the speech, such as staying on topic, presenting a clear thesis, employing vivid or imaginative use of language, and showing little to no bias or grammatical error in content when reviewing the transcripts.

Transcripts were organized and presented to raters in a randomized sequence such that each rater had a different randomized sequence of transcripts within the same subset of 10 performances. Raters were asked to score all of these 10 transcripts prior to conducting the video-based scoring, as it was believed that the amount of time required to provide transcript scores would be more variable than the on-the-rails presentation sequence for video files. That is, each rater experienced roughly the same reviewing duration for the video scoring, but each rater may read transcripts at a different pace, introducing a secondary source of variance for the amount of time required to produce a score.

In addition to providing the five content dimension scores from the PSCR, raters were also asked to provide confidence ratings (i.e., the proportion of certainty they had provided an accurate score), and start and stop time stamps were assigned when they began reviewing the transcript and provided their final PSCR score and confidence rating (see Figure A1).

### *Video Scoring Using the Public Speaking Competence Rubric*

PSCR scoring dimensions of interest in the video scoring portion of the study included vocal expression, nonverbal behavior, adaptation to audience, visual aids, and holistic score, which consisted of a more general impression of the subject's overall performance. These dimensions were used to evaluate the delivery of the speech. Further elaboration on the defining features of each scoring dimension recorded as part of video scoring can be found in Table 2.

**Table 2** Delivery Scoring Dimensions for Performance Videos From the Public Speaking Competence Rubric

| Scoring dimension | Corresponding PSCR performance standard[a] |
|---|---|
| Vocal expression | Effectively uses vocal expression and paralanguage to engage the audience |
| Nonverbal behavior | Demonstrates nonverbal behavior that supports the verbal message |
| Adaptation to audience | Successfully adapts the presentation to the audience |
| Visual aids | Skillfully makes use of visual aids |
| Holistic score | Overall performance quality |

*Note*. PSCR = Public Speaking Competence Rubric.
[a]See Schreiber et al. (2012, pp. 228 – 231) for the full rubric.

### *Confidence Scoring*

We hypothesized that with increasing information (e.g., the total amount of video viewed by a rater), the rater's reported confidence in scores and actual interrater reliability of scores should increase. One of the more prominent interests in conducting this study was measuring the degree to which raters could provide a reliable score comparable to a score given by an expert rater who has the luxury of unlimited time and the ability to thoroughly review each videotaped performance. If encouraging results are obtained in this regard, we would still want to know how much of the video was required to give an accurate score. To answer this, raters reviewed each file in three groupings of four clips (each 10 seconds in length) determined by thin-slicing method. Between each grouping of four clips (i.e., 40 seconds of video clips), raters provided a score for each of the PSCR dimensions covered in video scoring. Scores across groupings were then compared to detect possible patterns of stability in scoring over successive blocks of clips. Participants were instructed during the training session to provide scores in an additive fashion reflecting the entirety of the performance rather than only the last group of four clips. In this way, scores over successive groupings should be considered as an evaluation of the whole performance in light of additional information provided to the reviewer.

### *Thin-Slicing Methods*

Two variants of thin-slicing methods were explored. We initially conceived of a number of different ways the videos could be cut, processed, and presented to reviewers, though we had no a priori reason for believing one method was superior to any other. The first method consisted of direct, random sampling of the entire video file (hereinafter *random slicing*). Using the 10 video files, we determined the entire file duration in seconds, start time stamps for 12 evenly distributed intervals corresponding to the three groupings of four clips to be presented (e.g., a file 300 seconds in length is divided into 12 equal parts with corresponding start markers 25 seconds apart), and randomized the order of start markers obtained in this way. The rationale for finding and randomly sorting start markers using this method was to ensure reviewers had an equal chance of seeing a selection from the beginning or end of the presentation for any single clip. Clip duration was held constant at 10 seconds each, without respect to long pauses or other features of the performance.

Second, we tested an early and more deliberate method for thin-slicing based on the content of individual utterances in each performance (hereinafter *sentence slicing*). Relying on each video file's corresponding transcript (using human transcription) and extracted .wav file, we found start and stop time stamps for individual utterances throughout each file based on punctuation in the corresponding transcript. Specifically, we went through the transcript and found the sentence boundaries and the needed timing for each of the sentences. We used a forced alignment algorithm to create this look-up table of sentence start – stop markers. The output of the forced alignment process was viewed in Praat, a free-to-download phonetics and audio editing application for quality checking.[1]

To create sentence-based thin-slicing video vignettes, we randomly sampled 12 sentences from each video, with the caveat that each segment (i.e., utterance) must be longer than 2 seconds in duration for inclusion in the final data set. Although variability in utterance duration was observed ($SD = 5.17$ seconds), the average duration for these clips was comparable to that of the clips from the random-slicing method ($M = 8.7$, compared to the constant 10 seconds). We examined whether time stamps differed between the random- and sentence-slicing methods as a check to ensure that the breadth of each performance was sampled evenly as intended. Independent samples $t$-tests revealed no significant difference in the distribution of start time stamps ($t = .59$, $p = .56$) and stop time stamps ($t = .47$, $p = .64$) between final data sets for either thin-slicing method.

Video thin-slicing was implemented using open standards and open source software. The slicing and presentation of videos was done using SMIL 3.0.[2] See Appendix B for a sample SMIL file used in this study. SMIL is widely supported on modern browsers, making it easy to implement as a Web-based distributed scoring system in the future. Another benefit of SMIL is that we did not need to physically segment the video files into smaller files, which would be difficult to manage. Instead, the video files could be kept intact. Slices can be defined in a SMIL file (in XML) that defines the start and end time stamps of the segments. The SMIL files can be automatically generated for different scoring task requirements (e.g., how long each segment is and whether to randomize the order of presentation). We defined the beginning and end time stamps of video slices using SMIL's < par> element and strung them together using the <seq> elements. Every four slices, the playback was paused to allow for scoring. This was implemented using <smilText . . . end = "activateEvent" > (see Appendix B). SMIL files were automatically generated using a custom AWK script, which takes the word time stamp data (an output of the Praat program) as input.

AMBULANT Player for Windows was used to present performances to raters using the aforementioned thin-slicing methods. AMBULANT is an open source video player that allows for writing SMIL scripts for video editing and playback.[3] These files also permit users to write in prompts and reference/playback various files within the same script, making it a versatile tool for customizing content to be displayed. Start time stamps and clip durations were written into multiple SMIL files. Random-slicing and sentence-slicing variants of these SMIL files were created for each performance (20 total SMIL files for playback). Raters eventually would only view half of either set of thin-sliced files (e.g., Rater 1 would view five random-sliced files and five sentence-sliced files). File IDs were assigned such that each file would have exactly two reviewers for each thin-slicing method.

Raters were instructed to open the SMIL file corresponding to each assigned performance in AMBULANT and follow the programmed on-screen prompts. The raters were transitioned from one grouping to the next after watching four clips. Specifically, a prompt appeared on-screen instructing raters to provide PSCR scores on their score sheets, then to click the screen to continue with the next group of four clips. Raters proceeded in this way until all 12 clips had been reviewed, and then a prompt appeared instructing the rater to close AMBULANT and load the next file.

Raters were provided with a list of file IDs to load into AMBULANT in a preassigned sequence, determined semirandomly, so that the thin-slicing method assigned for each file alternated between random slicing and sentence slicing. The order and method of thin-slicing for files presented was inverted for each rater. For instance, Rater 2 received the same assignment for thin-slicing method as Rater 1, but it was presented to Rater 2 in reverse order, and Rater 3 received the opposite thin-slicing assignment as Rater 1, but in the same order. Consequently, no rater played the same list of files with the same thin-slicing method assignment *and* in the same order, but all participants had viewed files with this alternating thin-slicing method arrangement. Counterbalancing the order of performances in this way was done in an effort to detect any role that familiarity with the scoring rubric may have on confidence ratings or scores.

## Results

Results were based on content (i.e., introduction, organization, conclusion, word choice, and persuasion) and delivery (i.e., verbal expression, nonverbal behavior, adaptation to audience, visual aids, and holistic) scores across the following conditions.

- *Original scores*: Ratings taken from Joe et al. (2015), where trained raters scored the same videos with no time limit. These scores served as the *criterion scores*. The raters from Joe et al. (2015) were not part of the current study.
- *Transcript scores*: The condition in which raters scored content-related constructs using speech transcripts only. Each transcript was scored by all four raters in the present study.
- *Random-slicing and sentence-slicing scores*: Conditions used different methods of thin-slicing for video-based delivery-related construct scoring. As mentioned earlier, the slices were either 10-second-long clips with random starting points or randomly selected sentence-long utterances. On average, the duration of the sentence-based slices was 8.7 s ($SD = 5.7$ seconds), which was largely comparable to the length in the random condition. Note that for a particular video performance, a rater scored in either the random-slicing or sentence-slicing condition, but not both. Two raters were assigned to rate in each condition.
- *Time 1, Time 2, and Time 3 scores*: In the random- and sentence-slicing conditions, raters were asked to score the overall performance after viewing each of the three video blocks. Hence raters provided scores three separate times for one video.

**Table 3** Summary of Rater Assignment

| | Original | | Transcript | | | | Random | | Sentence | |
|---|---|---|---|---|---|---|---|---|---|---|
| Video | R1 | R2 | R1 | R2 | R3 | R4 | R1 | R2 | R1 | R2 |
| 1 | AA | BB | A | B | C | D | C | B | A | D |
| 2 | AA | CC | A | B | C | D | C | B | A | D |
| 3 | DD | EE | A | B | C | D | C | B | A | D |
| 4 | DD | CC | A | B | C | D | C | B | A | D |
| 5 | DD | EE | A | B | C | D | A | D | C | B |
| 6 | DD | EE | A | B | C | D | A | D | C | B |
| 7 | AA | EE | A | B | C | D | A | D | C | B |
| 8 | AA | BB | A | B | C | D | A | D | C | B |
| 9 | AA | BB | A | B | C | D | A | D | C | B |
| 10 | AA | BB | A | B | C | D | C | B | A | D |

*Note*. R = rater.

**Table 4** Means and standard deviations of content-related scores across conditions

| | Original, *M* (*SD*) | | Transcript, *M* (*SD*) | | | |
|---|---|---|---|---|---|---|
| Score | R1 | R2 | R1 | R2 | R3 | R4 |
| Introduction | 3.40 (0.70) | 3.10 (0.88) | 2.70 (0.68) | 2.90 (0.99) | 2.80 (0.92) | 2.50 (0.71) |
| Organization | 3.30 (0.48) | 3.40 (0.70) | 2.90 (0.74) | 2.60 (0.84) | 1.80 (0.42) | 2.50 (0.53) |
| Conclusion | 2.90 (1.20) | 2.90 (0.88) | 2.50 (0.85) | 1.80 (1.23) | 1.60 (1.51) | 2.30 (0.95) |
| Word choice | 3.30 (0.67) | 3.40 (0.52) | 2.70 (0.48) | 3.00 (0.47) | 2.80 (0.63) | 2.80 (0.63) |
| Persuasion | 3.10 (0.74) | 3.20 (0.79) | 2.50 (0.71) | 2.50 (0.53) | 2.50 (0.53) | 2.90 (0.74) |

*Note*. Averages were calculated based on 10 videos. R = rater.

## Descriptive Statistics

Two raters scored the 10 videos across the content and delivery dimensions in the original, random-slicing, and sentence-slicing conditions; four raters scored the videos in the transcript condition. A summary of rater assignment is presented in Table 3. Note that there were five total raters in the original condition and four raters in the transcript, random-, and sentence-slicing conditions.

Tables 4 and 5 present the mean and standard deviations of raters' scores for the content- and delivery-related scores, respectively, across the various conditions. Scores in each dimension were averaged across the 10 videos. For content-related scores, the scores in the original condition were generally higher on average than the scores using the transcription method. For the delivery-related scores, there was less of a trend across conditions; instead, differences were more prominent across raters.

## Consistency Between Raters Within Conditions

The ICC and percentage of exact agreement between raters were computed for each condition across the content and delivery dimensions (see Tables 6 and 7, respectively). In the transcript condition, because the four raters rated all the 10 videos, a two-way random model was used to calculate the ICC. The *two-way random model* (a) models both an effect of rater and of the examinee (i.e., two effects) and (b) assumes both are drawn randomly from larger populations (i.e., a random effects model). However, in the original, random-slicing, and sentence-slicing conditions, because different raters scored the 10 videos, a one-way random model was used to calculate the ICC. The *one-way random model* makes no effort to disentangle the effects of the rater and examinee, so there is only one effect.

Results found that the interrater reliability between raters for the original condition was high across content- and delivery-related scores. Specifically, ICCs were all over .70, and the percentage of exact agreement was fairly high (>60%) across all scoring dimensions, except conclusion and visual aids. This was not surprising because the 10 videos were selected partly because of the high agreement. For the transcript condition (content-related scores), interrater reliability based on the ICC was high for the introduction score but not very good for the other scores; however, both the organization

**Table 5** Means and Standard Deviations of Delivery-Related Scores Across Conditions

| | Original, *M* (*SD*) | | Random slicing, *M* (*SD*) | | | | | | Sentence slicing, *M* (*SD*) | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Score | R1 | R2 | R1/ T1 | R1/ T2 | R1/ T3 | R2/ T1 | R2/ T2 | R2/ T3 | R1/ T1 | R1/ T2 | R1/ T3 | R2/ T1 | R2/ T2 | R2/ T3 |
| Verbal expression | 3.10 (0.99) | 3.20 (0.92) | 3.20 (0.79) | 3.40 (0.52) | 3.20 (0.79) | 3.10 (0.74) | 2.70 (1.06) | 2.80 (0.92) | 2.90 (0.74) | 3.00 (0.67) | 3.10 (0.57) | 2.80 (0.92) | 2.90 (0.99) | 2.70 (1.06) |
| Nonverbal behavior | 2.90 (1.20) | 2.90 (0.74) | 3.00 (0.67) | 3.10 (0.74) | 3.10 (0.74) | 2.60 (1.07) | 2.50 (1.08) | 2.40 (1.07) | 2.40 (0.97) | 2.70 (0.95) | 2.90 (0.99) | 2.60 (1.07) | 2.50 (1.27) | 2.50 (1.18) |
| Adaptation to audience | 3.40 (0.70) | 3.40 (0.70) | 2.70 (0.82) | 2.80 (0.79) | 2.80 (0.79) | 3.20 (0.63) | 3.30 (0.48) | 3.20 (0.63) | 2.70 (0.82) | 2.70 (0.82) | 3.00 (0.50) | 3.10 (0.88) | 3.10 (0.88) | 3.10 (0.88) |
| Visual aids[a] | 2.80 (1.10) | 3.00 (0.71) | 3.00 (0.71) | 3.00 (0.71) | 3.00 (0.71) | 1.80 (0.84) | 1.80 (1.10) | 1.80 (1.10) | 2.20 (1.48) | 3.20 (0.84) | 3.20 (0.84) | 2.00 (0.71) | 2.20 (0.84) | 2.00 (0.71) |
| Holistic | 3.20 (0.92) | 3.30 (0.82) | | | 3.10 (0.57) | | | 2.70 (0.82) | | | 3.00 (0.67) | | | 2.80 (1.03) |

*Note*. R = rater; T = time.
[a]Average calculated based on 5 videos instead of 10.

**Table 6** Consistency Between Raters on Content-Related Scores Within Conditions

| | Original | | |
|---|---|---|---|
| Score | ICC | % Agree | Transcript (ICC) |
| Introduction | .87 | 70 | .85 |
| Organization | .73 | 70 | .66 |
| Conclusion | .77 | 50 | .64 |
| Word choice | .93 | 90 | .27 |
| Persuasion | .85 | 70 | .39 |

*Note*. % Agree was not calculated for the transcript condition because there were 4 raters. ICC = intraclass correlation between raters; % Agree = percentage of exact agreement between raters.

**Table 7** Consistency Between Raters on Delivery-Related Scores Within Conditions

| | Original | | Random slicing | | | | | | Sentence slicing | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Score | ICC | % Agree | T1, ICC | T1, % agree | T2, ICC | T2, % agree | T3, ICC | T3, % agree | T1, ICC | T1, % agree | T2, ICC | T2, % agree | T3, ICC | T3, % agree |
| Verbal expression | .91 | 70 | .33 | 40 | .09 | 30 | .50 | 30 | .78 | 50 | .52 | 40 | .31 | 40 |
| Nonverbal behavior | .89 | 60 | .41 | 40 | .63 | 60 | .48 | 30 | .59 | 40 | .68 | 40 | .39 | 10 |
| Adaptation to audience | 1.00 | 100 | .73 | 50 | .65 | 50 | .78 | 60 | .63 | 20 | .75 | 40 | .44 | 30 |
| Visual aids | .78 | 40 | .27 | 0 | .26 | 20 | .26 | 20 | .44 | 40 | .39 | 20 | .27 | 0 |
| Holistic | .87 | 90 | | | | | .60 | 40 | | | | | .49 | 30 |

*Note*. ICC = intraclass correlation between raters; T = time; % Agree = percentage of exact agreement between raters.

and conclusion scores had ICCs above .60 (Table 6). For the random- and sentence-slicing conditions (delivery-related scores), the interrater reliability was high for the adaption to audience score (except for Time 3 in the sentence-slicing condition), but there was noticeable disagreement between raters for other delivery-related scores and the holistic score (Table 7). Additionally, the interrater reliability for the random- and sentence-slicing conditions did not seem to increase from Time 1 to Time 3. In other words, we did not see clear evidence that more slices or viewing time resulted in more consistent ratings.

Overall, as compared to the original condition, the interrater reliability for the transcript, random-slicing, and sentence-slicing conditions was lower. There may be multiple reasons, though the current study cannot tease them apart. One reason

**Table 8** Interrater Reliability Between Criterion and Content-Related (Transcript) Scores

| | Rater A | | | Rater B | | | Rater C | | | Rater D | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Score | ICC | ρ | % Agree | ICC | ρ | % Agree | ICC | ρ | % Agree | ICC | ρ | % Agree |
| Introduction | <0[a] | −.37 | 50 | <0 | .05 | 40 | <0 | −.08 | 30 | <0 | −.12 | 50 |
| Organization | .23 | .24 | 60 | .44 | .79** | 30 | <0 | .09 | 10 | <0 | .11 | 30 |
| Conclusion | <0 | −.03 | 40 | .38 | .55 | 40 | .50 | .84** | 20 | .65 | .71** | 50 |
| Word choice | <0 | .13 | 30 | <0 | .00 | 30 | <0 | −.31 | 30 | .23 | .41 | 50 |
| Persuasion | <0 | .04 | 20 | <0 | .29 | 30 | .173 | .50 | 30 | .62 | .34 | 80 |

*Note*. ICC = intraclass correlation between raters; ρ = Spearman's rho; % Agree = percentage of exact agreement between raters.
**$p < .01$. *$p < .05$.
[a]Suggests that the correlation is weak or negative.

may be regression to the mean (i.e., the 10 videos were chosen because they happened to have high ICC in the original study, possibly by chance). We cannot prove or disprove this hypothesis because we did not replicate the original condition in the current study. Another reason could be rater selection and training. The raters in the current study were rated novice to performance, whereas those in the original study (Joe et al., 2015) were experienced essay raters. Last, the methodology could also have impacted the interrater reliability. The original scores were based on full (unlimited) viewing of the video performance, whereas the ratings in the current study were based on impoverished representations — transcripts (vs. audio and video) and thin slices (vs. complete videos). We note, however, that the distribution of ICCs in this study does not differ too much from those reported in Joe et al. (2015), which were in the .30 to .50 range. As we noted, the high ICCs for the original scores may be an artifact of the stimulus selection process.

## Consistency With Criterion (Original) Scores

### *Content-Related Scores: Transcript Versus Criterion*

The ICC, Spearman's rho correlation, and percentage of exact agreement were calculated between each rater's score in the transcript condition and the corresponding criterion score from the original condition. Because different raters rated the 10 videos, a one-way random model was used to calculate the ICC. Results (Table 8) show that while some dimensions and raters looked better than others, there was no consistent pattern to suggest that the scores based on speech transcripts were consistent with those obtained in the original condition. Specifically, the ICC was small or even negative between each transcript score and the corresponding criterion score, suggesting that the correlation between scores from the transcript and original condition is weak and unreliable.

In addition to evaluating the interrater reliability, we also evaluated whether there were any systematic biases for the transcript condition. The transcript bias was calculated as follows:

$$\text{Bias} = \text{Transcript score} - \text{Corresponding criterion score}. \quad (1)$$

The absolute bias of the transcript score was also calculated by taking the absolute value of the bias. The bias and the absolute bias for each content-related score in the transcript condition is shown in Table 9. Results show negative bias (i.e., underestimation of scores) across all raters, with scores from Rater C containing more error than other raters, as it had the largest negative bias and absolute bias. Rater D did better than others, because his or her scores showed relatively small bias. Across scores, organization and conclusion scores had the largest absolute bias. Overall, these results indicated that the content-related scores using the transcript method were generally harsher than the criterion scores. These results are supported by scatterplots, presented in Appendix C, that show the distribution of scores across videos for the criterion and transcript-based scores. As shown in Figures C1 – C5, raters in the transcript condition consistently underscored as compared to the criterion scores from the original condition.

Finally, to evaluate whether content-related scores across the original and transcript conditions were statistically different, a Wilcoxon signed-rank test was conducted. This test is the nonparametric version of analysis of variance (ANOVA) for small sample size. The results ($z$-values) are presented in Table 10. These results indicate that there were substantial individual differences among raters.

**Table 9** Bias and Absolute Bias of Rater's Content-Related (Transcript) Scores

| | Bias | | | | | Absolute bias | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Score | Rater A | Rater B | Rater C | Rater D | Sum | Rater A | Rater B | Rater C | Rater D | Sum |
| Introduction | −5.67 | −3.67 | −4.67 | −7.67 | *−21.67* | 9.33 | 9.33 | 10.33 | 9.33 | *38.33* |
| Organization | −4.67 | −7.67 | −15.67 | −8.67 | *−36.67* | 6.33 | 8.67 | 15.67 | 9.33 | *40.00* |
| Conclusion | −4.33 | −11.33 | −13.33 | −6.33 | *−35.33* | 10.67 | 11.33 | 13.33 | 7.67 | *43.00* |
| Word choice | −7.50 | −4.50 | −6.50 | −6.50 | *−25.00* | 8.50 | 7.50 | 9.50 | 6.50 | *32.00* |
| Persuasion | −6.50 | −6.50 | −6.50 | −2.50 | *−22.00* | 10.50 | 8.50 | 8.50 | 4.50 | *32.00* |
| Sum | *−28.67* | *−33.67* | *−46.67* | *−31.67* | | *45.33* | *45.33* | *57.33* | *37.33* | |

**Table 10** Wilcoxon Signed-Rank Test for Content-Related Scores (Criterion Versus Transcript Scores)

| Score | Rater A | Rater B | Rater C | Rater D |
|---|---|---|---|---|
| Introduction | −1.41 | −0.71 | −0.97 | −2.04* |
| Organization | −1.79 | −2.63** | −2.83** | −2.59** |
| Conclusion | −1.07 | −2.53* | −2.55* | −2.05* |
| Word choice | −2.49* | −1.81 | −1.87 | −2.26* |
| Persuasion | −1.75 | −2.01* | −2.08* | −0.96 |

**$p \leq .01$. *$p < .05$.

**Table 11** Summary of Confidence Scores for Transcript Condition

| Score | Rater A, *M* (*SD*) | Rater B, *M* (*SD*) | Rater C, *M* (*SD*) | Rater D, *M* (*SD*) | Average |
|---|---|---|---|---|---|
| Introduction | 41.00 (8.76) | 56.50 (6.69) | 81.00 (14.68) | 71.50 (7.84) | 62.50 |
| Organization | 35.00 (5.27) | 50.00 (10.54) | 79.00 (9.37) | 72.00 (5.87) | 59.00 |
| Conclusion | 38.00 (6.32) | 58.50 (8.83) | 83.00 (7.15) | 79.50 (7.62) | 64.75 |
| Word choice | 35.00 (8.50) | 56.50 (10.29) | 84.00 (3.94) | 71.50 (5.30) | 61.75 |
| Persuasion | 35.00 (7.07) | 51.00 (10.75) | 68.00 (20.30) | 73.00 (8.56) | 56.75 |

One reason for the differences in content scores using the transcription method could have been related to confidence ratings. As shown in Table 11, Rater A was the least confident in scoring performances using the transcription method, whereas Rater C was the most confident. These discrepancies in confidence ratings could provide some explanation for the low ICCs between raters and were likely due to the difficulty of reading through disfluencies in the transcription. That said, overall, results indicate that the transcript-based scoring of the content-related constructs differed substantially from the video-based scoring in the original condition.

### *Delivery-Related Scores: Video Thin-Slicing Versus Criterion*

In addition to evaluating content-related scores, we looked at the interrater reliability and bias for the delivery-related scores using the video thin-slicing technique. Combining the random- and sentence-slicing conditions, the ICC was calculated between each rater's slicing score and the corresponding criterion score (see the next section for results evaluating the two thin-slicing methods separately). Because different raters rated the 10 videos, a one-way random model was used to calculate the ICC.

In contrast to the transcript-based results for the content scores, results indicate general consistency between delivery-related scores from the video thin-slicing conditions and the criterion scores from the original study. Overall, results found high average ICCs above .70 for the visual aids and holistic scores at Time 3 (Table 12). Verbal expression and nonverbal behaviors also showed high average ICCs at Time 1 and Time 2, respectively. Across time points, verbal expression accuracy tended to decrease; however, additional replications are needed to confirm this result.

Bias results were negative, indicating that raters tended to underestimate delivery scores (Table 13). Among the raters, Raters B and C contained the most error in their scores when compared to the corresponding criterion scores. Of the delivery-related scores, nonverbal behaviors had the most error across raters, followed by verbal expression. These results

**Table 12** Intraclass Correlation Between Criterion and Delivery-Related (Thin-Sliced) Scores

| | Rater A | | | Rater B | | | Rater C | | | Rater D | | | Average | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Score | T1 | T2 | T3 | T1 | T2 | T3 | T1 | T2 | T3 | T1 | T2 | T3 | T1 | T2 | T3 |
| Verbal expression | .84 | .76 | .62 | .81 | .76 | .77 | .58 | .52 | .50 | .82 | .62 | .73 | .76 | .66 | .65 |
| Nonverbal behavior | .85 | .86 | .74 | .78 | .67 | .55 | .33 | .65 | .69 | .78 | .86 | .79 | .68 | .76 | .69 |
| Adaptation to audience | .55 | .57 | .26 | .72 | .57 | .72 | .09 | .31 | .38 | .89 | .89 | .89 | .56 | .58 | .56 |
| Visual aids | .49 | .79 | .75 | .48 | .39 | .39 | .74 | .75 | .93 | .56 | .79 | .79 | .57 | .68 | .72 |
| Holistic | | | .57 | | | .74 | | | .78 | | | .85 | | | .73 |

*Note*. T = time.

**Table 13** Bias and Absolute Bias of Rater's Delivery-Related (Thin-Sliced) Scores

| | Rater A | | | Rater B | | | Rater C | | | Rater D | | | Sum | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Score | T1 | T2 | T3 | T1 | T2 | T3 | T1 | T2 | T3 | T1 | T2 | T3 | T1 | T2 | T3 |
| Bias | | | | | | | | | | | | | | | |
| Verbal expression | −0.5 | 0.5 | 1.5 | −4.5 | −7.5 | −7.5 | −1.5 | 0.5 | −1.5 | 0.5 | 0.5 | −0.5 | *−6.0* | *−6.0* | *−8.0* |
| Nonverbal behavior | 0.0 | 2.0 | 3.0 | −5.0 | −10.0 | −9.0 | −6.0 | −4.0 | −3.0 | −3.0 | 0.0 | −2.0 | *−14.0* | *−12.0* | *−11.0* |
| Adaptation to audience | −5.0 | −4.0 | −3.0 | −5.0 | −4.0 | −5.0 | −9.0 | −9.0 | −7.0 | 0.0 | 0.0 | 0.0 | *−19.0* | *−17.0* | *−15.0* |
| Visual aids | −2.5 | 0.5 | 1.5 | −5.5 | −6.5 | −6.5 | −0.5 | 1.5 | 0.5 | −4.5 | −2.5 | −3.5 | *−13.0* | *−7.0* | *−8.0* |
| Holistic | | | −1.5 | | | −7.5 | | | −2.5 | | | −2.5 | | | |
| Sum | *−8.0* | *−1.0* | *1.5* | *−20.0* | *−28.0* | *−35.5* | *−17.0* | *−11.0* | *−13.5* | *−7.0* | *−2.0* | *−8.5* | *−52.0* | *−42.0* | *−42.0* |
| Absolute bias | | | | | | | | | | | | | | | |
| Verbal expression | 4.5 | 5.5 | 6.5 | 5.5 | 7.5 | 7.5 | 6.5 | 6.5 | 7.5 | 4.5 | 6.5 | 5.5 | *21.0* | *26.0* | *27.0* |
| Nonverbal behavior | 5.0 | 4.0 | 5.0 | 7.0 | 10.0 | 10.0 | 9.0 | 7.0 | 7.0 | 6.0 | 5.0 | 7.0 | *27.0* | *26.0* | *29.0* |
| Adaptation to audience | 7.0 | 6.0 | 7.0 | 5.0 | 6.0 | 5.0 | 9.0 | 9.0 | 7.0 | 2.0 | 2.0 | 2.0 | *23.0* | *23.0* | *21.0* |
| Visual aids | 5.5 | 2.5 | 2.5 | 5.5 | 6.5 | 6.5 | 2.5 | 2.5 | 1.5 | 4.5 | 3.5 | 3.5 | *18.0* | *15.0* | *14.0* |
| Holistic | | | 6.5 | | | 7.5 | | | 4.5 | | | 4.5 | | | |
| Sum | *22.0* | *18.0* | *27.5* | *23.0* | *30.0* | *36.5* | *27.0* | *25.0* | *27.5* | *17.0* | *17.0* | *22.5* | *89.0* | *90.0* | *91.0* |

*Note*. T = Time.

are supported by scatterplots presented in Appendix D, which show the distribution of scores across videos for the criterion and thin-sliced scores. Similar to the transcript condition, raters in the thin-slicing conditions tended to underscore as compared to the criterion scores from the original condition (see Figures D1 – D5).

### *Sentence- Versus Random-Slicing Condition*

In this study's design, two methods for thin-slicing video performances were implemented: (a) random slicing, which likely resulted in fragmented utterances in each slice, and (b) sentence slicing, which preserved the semantic integrity of the utterance. The following analyses compared the two approaches.

The ICC was calculated between the random- or sentence-slicing delivery score and the corresponding criterion score. Because different raters rated the 10 videos, a one-way random model was used to calculate the ICC. The ICCs for the random- and sentence-slicing conditions are presented in Table 14. Both methods achieved quite high consistency with the criterion scores. For the holistic score, the average ICC was .77 and .72 for the random- and sentence-slicing method, respectively. In addition, the high ICCs were fairly consistent across three scoring times, again reinforcing the conclusion that the video thin-slicing technique is a promising solution for scoring the delivery-related constructs. The differences between random- and sentence-slicing ICCs were minor and unsystematic, suggesting that they may be partly due to chance. However, random slicing was slightly better in adaptation to audience, visual aids, and holistic, whereas sentence slicing performed slightly better for the verbal expression and nonverbal behavior scores. Additionally, Rater 1 (consisting of Raters A and C) seemed to be the most inconsistent when scoring audience adaptation in both the random- and sentence-slicing conditions.

The Wilcoxon signed-rank test was also conducted to test if there were any differences between random- and sentence-slicing delivery-related scores and the corresponding criterion score (Table 15). Results suggest some significant differences between the adaptation to audience scores in the random-slicing condition and the corresponding criterion scores

**Table 14** Intraclass Correlation Between Criterion and Raters' Delivery-Related (Thin-Sliced) Scores

| | Random slicing | | | | | | | | | Sentence slicing | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Score | R1/T1 | R1/T2 | R1/T3 | R2/T1 | R2/T2 | R2/T3 | Avg/T1 | Avg/T2 | Avg/T3 | R1/T1 | R1/T2 | R1/T3 | R2/T1 | R2/T2 | R2/T3 | Avg/T1 | Avg/T2 | Avg/T3 |
| Verbal expression | .55 | .45 | .47 | .73 | .76 | .88 | .64 | .61 | .67 | .85 | .77 | .66 | .88 | .68 | .64 | .86 | .73 | .65 |
| Nonverbal behavior | .44 | .70 | .70 | .74 | .71 | .65 | .59 | .71 | .67 | .76 | .81 | .74 | .81 | .80 | .71 | .79 | .81 | .73 |
| Adaptation to audience | .52 | .54 | .54 | .88 | .73 | .88 | .70 | .64 | .71 | .09 | .34 | <0 | .76 | .76 | .76 | .43 | .55 | |
| Visual aids | .93 | .93 | .93 | .48 | .62 | .62 | .70 | .78 | .78 | .31 | .61 | .75 | .56 | .56 | .56 | .44 | .59 | .66 |
| Holistic | | | .75 | | | .80 | | | .77 | | | .64 | | | .80 | | | .72 |

*Note*. Less than 0 suggests that the correlation is negative. Avg = average; R = rater; T = time.

**Table 15** Wilcoxon Signed-Rank Test for Delivery-Related Scores (Criterion Versus Thin-Sliced Scores)

| | Random slicing | | | | | | Sentence slicing | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Score | R1/T1 | R1/T2 | R1/T3 | R2/T1 | R2/T2 | R2/T3 | R1/T1 | R1/T2 | R1/T3 | R2/T1 | R2/T2 | R2/T3 |
| Verbal expression | −0.11 | −0.86 | −0.17 | −0.21 | −1.59 | −1.84 | −1.39 | −0.78 | −0.29 | −1.82 | −0.69 | −1.16 |
| Nonverbal behavior | 0.00 | −0.42 | −0.42 | −1.36 | −1.71 | −1.71 | −2.21* | −1.19 | −0.32 | −1.55 | −1.63 | −1.56 |
| Adaptation to audience | −2.33* | −2.12* | −2.12* | −1.41 | −0.58 | −1.41 | −1.93 | −2.11* | −1.41 | −1.34 | −1.34 | −1.34 |
| Visual aids | −0.58 | −0.58 | −0.58 | −2.04* | −2.04* | −2.04* | −0.92 | −0.74 | −0.82 | −2.04* | −1.63 | −2.04* |
| Holistic | | | −0.85 | | | −2.33* | | | −0.90 | | | −1.72 |

*Note*. R = rater; T = time.

** $p < .05$.

for Rater 1 and on the visual aids and holistic scores for Rater 2. For the sentence-slicing condition, significant differences were also found on adaptation to audience and visual aids scores, as well as the nonverbal behaviors score across Raters 1 and 2; however, these differences were directly related to the time point of the score.

## Time Spent on Scoring

One of the objectives of this study was to obtain an estimate of the scoring time necessary to achieve a reasonably reliable rating because shorter scoring time would not only mean less of a burden on the raters but also more affordable scoring costs. It should be noted that there were no time restrictions for transcript scoring because we hoped to evaluate the method under the most favorable conditions. Consequently, we do not know what the effect would be for an abbreviated time condition. On the other hand, video thin-slicing is inherently constrained in time: The raters were given a fixed amount of time to view the performances (i.e., 40 seconds for each random-sliced video, and approximately 40 seconds for each sentence-sliced video) and were only permitted to watch through the clips once; however, we did not limit the scoring time afterward. The estimates here are likely to be the upper limits and can be reduced in operational settings.

Contrary to our hypothesis, unconstrained transcript scoring took approximately the length of the video performance, though with much variation (Table 16). Several factors may have contributed to the result:

- We did not limit rater scoring time; an explicit guideline could have reduced the time.
- The raters in this study had minimal training scoring written texts; more experienced essay raters may be more efficient.
- Speech transcripts contain disfluencies and may be more difficult to comprehend, compared to written essays and the speech in audio form.
- Rating based on speech transcripts may be difficult without the support of the vocal expressions and/or nonverbal behaviors (e.g., gestures and eye contact).

The latter two would suggest that speech transcript scoring is inherently problematic, whereas the first two could be remedied with training. Given that we only have information on time duration, we cannot determine the reason based on the current study.

**Table 16** Summary of Scoring Duration Across Raters and Conditions

| | Time 1 | | | Time 2 | | | Time 3 | | |
|---|---|---|---|---|---|---|---|---|---|
| Rater | *M* (*SD*) | Min. | Max. | *M* (*SD*) | Min. | Max. | *M* (*SD*) | Min. | Max. |
| Transcript | | | | | | | | | |
| 1 | 5.30 (2.31) | 2 | 10 | – | – | – | – | – | – |
| 2 | 7.00 (2.67) | 4 | 14 | – | – | – | – | – | – |
| 3 | 4.00 (1.41) | 2 | 6 | – | – | – | – | – | – |
| 4 | 6.56 (3.21) | 3 | 13 | – | – | – | – | – | – |
| Random slicing | | | | | | | | | |
| 1 | 2.20 (1.81) | 1 | 7 | 1.00 (0.47) | 0 | 2 | 1.30 (0.68) | 0 | 2 |
| 2 | 1.70 (0.68) | 1 | 3 | 1.80 (0.92) | 1 | 4 | 1.80 (0.79) | 1 | 3 |
| Sentence slicing | | | | | | | | | |
| 1 | 2.22 (0.97) | 1 | 4 | 1.50 (0.71) | 1 | 3 | 1.30 (0.68) | 1 | 3 |
| 2 | 2.10 (0.57) | 1 | 3 | 1.40 (0.70) | 1 | 3 | 1.67 (0.50) | 1 | 2 |

*Note*. Times are in minutes.

**Table 17** Summary of Confidence Scores for Thin-Sliced Conditions

| | Rater 1, *M* (*SD*) | | | Rater 2, *M* (*SD*) | | | Average | | |
|---|---|---|---|---|---|---|---|---|---|
| Random slicing | T1 | T2 | T3 | T1 | T2 | T3 | T1 | T2 | T3 |
| Verbal expression | 59.0 (29.8) | 65.5 (28.1) | 61.0 (27.2) | 60.0 (13.9) | 64.5 (13.4) | 71.5 (11.8) | 59.0 | 65.0 | 66.3 |
| Nonverbal behavior | 54.0 (23.1) | 65.0 (21.3) | 67.5 (19.5) | 59.5 (16.1) | 63.0 (16.4) | 66.5 (15.1) | 54.0 | 64.0 | 67.0 |
| Adaptation to audience | 43.5 (21.0) | 47.5 (17.2) | 50.5 (17.7) | 57.0 (14.9) | 62.5 (13.2) | 65.0 (12.7) | 43.5 | 55.0 | 57.8 |
| Visual aids | 53.0 (25.9) | 64.0 (21.6) | 70.0 (24.2) | 56.0 (15.2) | 63.0 (11.0) | 69.0 (11.4) | 53.0 | 63.5 | 69.5 |
| Holistic | | | 60.9 (17.1) | | | 70.5 (9.3) | | | 65.7 |
| Sentence slicing | | | | | | | | | |
| Verbal expression | 58.5 (26.0) | 65.0 (24.8) | 66.5 (19.6) | 63.0 (12.3) | 62.0 (14.9) | 69.5 (10.7) | 58.5 | 63.5 | 68.0 |
| Nonverbal behavior | 57.0 (19.3) | 59.0 (22.2) | 62.5 (23.4) | 59.0 (13.1) | 63.0 (10.9) | 69.5 (11.2) | 57.0 | 61.0 | 66.0 |
| Adaptation to audience | 42.0 (29.3) | 48.5 (24.5) | 58.3 (17.7) | 52.5 (14.6) | 58.0 (15.3) | 64.5 (12.1) | 42.0 | 53.3 | 61.4 |
| Visual aids | 50.0 (18.7) | 55.0 (24.0) | 62.0 (17.9) | 54.0 (11.4) | 62.0 (9.1) | 65.0 (11.2) | 50.0 | 58.5 | 63.5 |
| Holistic | | | 64.0 (20.3) | | | 66.5 (10.6) | | | 65.3 |

*Note*. T = time.

For the two video thin-slicing methods, results show that the amount of time spent on scoring decreased from Time 1 to Time 2 but stayed fairly stable from Time 2 to Time 3. It was estimated that the scoring of the delivery dimensions took approximately 30 – 60 seconds. There was little difference between the scoring duration between the random- and sentence-slicing conditions.

## Confidence Scores Over Time for Thin-Slicing Scoring

In addition to providing scores across dimensions, raters were also asked to record their level of confidence in their score (on a scale of 1 – 100) in the two thin-sliced conditions. It was hypothesized that the confidence levels would increase with the additional exposure to the performance in subsequent viewing time blocks. Results found that while the confidence rating did go up, it seemed the additional viewing time blocks after the first only gave diminishing returns (Table 17). Combined with the finding that the quality of the rating increased only very modestly over the three viewing time blocks (Table 14), our tentative conclusion is that having one 40-second thin-slicing block gave the biggest bang for the buck. Additional blocks helped, but only modestly. More specifically, confidence scores increased on average between 6% and 11% from Time 1 to Time 2 in the random-slicing condition and from 1% to 6% from Time 2 to Time 3. Similar trends were found in the sentence-slicing condition with an average increase of 4% to 11% from Time 1 to Time 2 and from 5% to 8% from Time 2 to Time 3. Results also show that raters on average had similar confidence ratings across the two thin-slicing conditions at Time 3 (around 68%). Among the various delivery-related scores, raters showed the least amount of confidence when scoring adaptation to audience.

## Discussion and Conclusions

This proof-of-concept study tested the feasibility of two new methods for scoring video-based public speaking performance, with the goal of balancing the cost of scoring and the validity and reliability of the scores. The study was designed with two existing models of video-performance rating in mind: (a) unconstrained viewing and (b) watching once. The unconstrained viewing method was used in Joe et al. (2015) to provide the most favorable condition for reliability and validity and requires 3 to 4 times real time for scoring. As a result, scores from Joe et al. (2015) served as the criterion scores for this study. The watching-once method is used in some operational scoring of video performances where raters are required to watch the performance exactly once without pausing, skipping, or rewinding and then to provide scores. Unlike the unconstrained viewing method, this method controls the time duration to 1 times real time plus scoring time. While it is effective for holistic scoring, it may be susceptible to issues with attention, working memory, and bleeding across dimensions if independent subscores (e.g., analytical rubrics) are needed.

Considering the two existing models of video-performance scoring, this study proposed to score (a) content-related constructs based on speech transcripts and (b) delivery-related constructs using thin slices of videos. We hypothesized that the use of these two methods (transcript plus thin-slicing) will have the following consequences:

- It will reduce the working memory load by concentrating on a smaller number of constructs at a time (five content-related and five delivery-related constructs).
- It will optimize the presentation of information for the dimensions to be scored. We predicted that written transcripts are better than the fleeting audio or video presentations in evaluating the structure of an oral presentation (i.e., content-related dimensions); similarly, video thin-slicing can focus raters' attention on delivery features by removing much of linguistic context (i.e., delivery-related dimensions).
- It will reduce the scoring time. This is obvious in the case of thin-slicing. With regard to speech transcript scoring, our prediction was that the faster speed of reading would reduce the scoring time.
- It will allow for crowd-sourced scoring. By decomposing the scoring for different dimensions into separate scoring tasks, each of which may be done with a fraction of the time to score the whole performance, we hope this will open the door to high-quality crowd-sourced scoring, where each dimension of the performance may be scored by multiple raters (as opposed to single or double scoring by the same rater or raters), thus reducing rater variance.

### Transcript-Based Scoring

Results were mixed when using the speech transcription for scoring content-related dimensions. On one hand, raters were able to agree on the structural aspects of the speech (i.e., introduction, organization, and conclusion), though not so much on word choice or persuasion. On the other hand, the scores based on the transcript only differ systematically from the criterion scores, as shown in Table 6 as well as the figures in Appendix C. Raters also took about 1 times real time to read and score these dimensions (i.e., the time saving was not significant). There are several, not necessarily exclusive, reasons behind the finding. First, the speech transcripts may have been so impoverished that they are not appropriate as the basis for scoring. Second, the verbatim speech transcripts may have contained many speech disfluencies and other stylistic differences from a typical written essay, such that the (novice) raters may have had difficulties reading and rating. Last, the speech transcripts actually eliminated contaminations from delivery channels (in the original scoring condition), and therefore the scores were (largely) consistent but differ from the criterion scores from the original study.

Although we cannot rule out any of the interpretations in the current study, we can suggest several ideas for follow-up research and development. Specifically, to address the concern about impoverished speech transcripts, future prototypes could integrate the speech audio with the transcript in a scoring interface that presents the transcript text and the ability to play back the audio in synchrony. A close analogy is the read-aloud with synchronized word highlighting found in many electronic book reader software programs today. In addition, we expect that the audio can be played at 1.5 – 2 times real time using VSP technologies, which will speed up the scoring. Having both the transcript and the audio could be better than having either alone. Meanwhile, not showing the video of the performance helps to isolate content-related constructs from delivery dimensions.

To clean up speech transcripts with disfluencies and other obvious impediments to reading comprehension, natural language processing technologies may be applied. This could increase the readability of the speech transcripts, making

it easier to score the various content-related dimensions. Finally, to test whether there is potential contamination from delivery dimensions to content-related constructs, a well-designed experimental study may be needed.

## Video Thin-Slicing

Overall, results of this study suggest that the video thin-slicing method is promising. Even with the small sample of raters ($n = 4$), reasonable interrater reliability and correlations with the criterion score were found across most of the delivery-related dimensions. This method also shows promise in achieving the design goal of controlling and reducing the scoring time. Results suggest that a single time block of approximately 40 seconds of viewing of thin slices was sufficient to achieve reasonable confidence and reliability. Additional viewing times helped, but the payoff diminished compared to the first time block. Additional time may or may not be necessary if this method were to be used operationally. Given the concern about individual differences among raters, it may be more profitable to assign additional raters than to ask the same rater to view more of the video slices. In other words, the thin-slicing scoring approach is consistent with a crowd-sourcing model, where the average of many relatively "shallow" ratings may be better than a few very "deep" ratings.

On the basis of the results from this proof-of-concept, the video thin-slicing method shows promise and has the potential to substantially reduce the time for scoring delivery-related dimensions in a public speaking assessment. The cost savings in scoring time may be used to increase the number of raters and thus the quality of the scores.

## Technical Feasibility

For this proof-of-concept study, we chose to implement the key technology to support video thin-slicing; for transcript-based scoring, we used the low-tech paper-based scoring. This experiment indicated that video thin-slicing can be done effectively when using the standard-compliant SMIL approach. Specifically, virtual slices were defined in an SMIL file that can be edited in any text editor. The SMIL player seeks and rewinds the video playhead automatically according to the SMIL file, leaving the original performance video recording file intact. There is no need to create individual video slices (i.e., no need for costly video processing or the headache of managing of hundreds of video vignette files). Using the standard-compliant SMIL approach, slicing can be automatically generated. This was easy for the random-slicing condition, where a simple script calculated the starting times of the slices according to the total duration of the video and the length of each slice. Sentence-based slicing required the onset and offset times for each sentence, which were obtained using force-alignment algorithms. That said, results showed little evidence favoring the more complex sentence-based slicing, suggesting that the simple random slicing may work just as well.

For simplicity, we chose to present the thin-slicing stimuli using the AMBULANT SMIL player. SMIL can also be easily implemented in the browser — numerous JavaScript libraries support SMIL playback. Integrating SMIL-based thin-slicing in operational scoring interfaces seems technically feasible. Last, we suggested that transcript-only scoring may need to be augmented with audio playback in the future. While we did not test this setup, it is feasible to implement this functionality in HTML5 to play back the audio (potentially in variable speed) while displaying synchronized word highlighting of the transcript text using, again, SMIL.

To summarize, despite the minimalistic approach of this study, this study demonstrates some promise toward a reliable and cost-efficient method for video thin-slicing scoring. However, because this study only used criterion scores with high ICCs, it is unclear whether the same results would be found using criterion scores with lower ICCs. In the future, the SMIL-based approach should also support additional functionalities, such as scoring of transcript and audio.

## Study Design Limitations

The scoring method prototype was developed based on video stimuli from Joe et al. (2015). The method was tested with four novice raters, who underwent a training session before scoring the 10 video performances using the new scoring methods. This is a preliminary study of an early prototype of the scoring method, based on a very small ($n = 4$) sample size with a convenience sample. Because this was a proof-of-concept study, it is critical to identify the study limitations

to determine how to further investigate this method for video-based scoring of public speaking performance in future studies. Three important limitations of the study are highlighted.

First, the sample of raters was very small, nonrandom, and undertrained. A larger sample was not available at the time of the data collection. Meanwhile, we also wanted to see if novice raters could do the job with minimal training. Second, the stimuli (10 videos) were selected based on high criterion scores in the original condition. As discussed, this may have introduced sampling biases. The higher criterion scores can potentially introduce a regression-to-the-mean effect, which can lead to underestimation of the criterion-based ICC. These limitations underscore the different design priorities between a proof-of-concept and a validity study — numeric results from this study cannot be used as estimates of how the method will perform in operational settings.

Last, the design of the transcript-based method was not optimized. The original plan was to use transcript-based scoring with synchronized audio/video playback; however, time and resource constraints meant that we had to choose between focusing the technology development on the video thin-slicing technique or transcript-based scoring. We decided to compromise the transcript-based method by doing low-tech paper-based scoring. If the minimal technology approach for transcript-based scoring was successful, it would have meant good news for cost reduction. The fact that it did not work as well as video thin-slicing should not be taken as evidence that the idea of transcript-based scoring should be abandoned. As previously discussed, we have identified ways to improve the design and implementation in the next iteration.

We stress that a proof-of-concept study necessarily implies less-than-ideal implementations and design trade-offs. The conditions for operational scoring will be different, which should be taken into account when evaluating the results.

## Future Directions for Scoring Video Performances

Looking beyond the current technology-assisted scoring proof-of-concept, we believe automated scoring will ultimately be part of the solution for a valid, reliable, and cost-effective oral communication skills assessment for higher education. Although there has been some development of automated multimodal scoring technologies, more research and test-specific training are needed for these approaches to become operational in a large-scale test setting. The transition toward automated scoring may begin with a period of "technology-assisted" scoring, where human raters are assisted with multimodal technologies, followed by a period of "technology-on-training-wheels," where human raters monitor and correct automated scoring models, before automated scoring technologies may be used as the primary score under human monitoring. This proof-of-concept study suggests that innovations in human scoring methods can control the cost of scoring, potentially maintaining or even increasing the quality of the scores.

For future research regarding scoring of video-based public speaking performances, we recommend a small-scale prototyping study to improve the scoring of content-related constructs using both transcript and audio and further prototyping and fine-tuning the video thin-slicing method using authentic assessment tasks. For instance, future research should investigate whether the method in which clips are selected and presented (i.e., at random or systematically) has any particular advantages or disadvantages. Additionally, future research should investigate the reliability of content and delivery scores separately based on watching the full video and based on using the two separate scoring methods (transcription and thin-slicing). It could be that using a method that separates the scoring of content and delivery can impact the relationship between these two dimensions and reduce any cross-trait bleeding that may occur when using a method where the scoring of the two dimensions occurs simultaneously.

To conclude, results of this study show promise that video thin-slicing is a viable option to improve human scoring of delivery-related constructs that reduces scoring costs. That said, another round of prototyping is necessary to improve the scoring of the content-related constructs.

## Acknowledgments

Jilliam Joe is currently with Leap Innovations, Chicago, IL. Christopher Kitchen is now at the University of Maryland, Baltimore, School of Pharmacy. Liyang Mao is currently at IXL Learning, San Mateo, CA. Lei Chen is now at LAIX Silicon Valley AI Lab, San Mateo, CA.

## Notes

1 http://www.fon.hum.uva.nl/praat/
2 http://www.w3.org/TR/SMIL/
3 http://www.ambulantplayer.org/

## References

Albrechtsen, J. S., Meissner, C. A., & Susa, K. J. (2009). Can intuition improve deception detection performance? *Journal of Experimental Social Psychology*, *45*, 1052 – 1055.

Ambady, N., Krabbenhoft, M. A., & Hogan, D. (2006). The 30-sec sale: Using thin-slice judgments to evaluate sales effectiveness. *Journal of Consumer Psychology*, *16*(1), 4 – 13.

Ambady, N., & Rosenthal, R. (1993). Half a minute: Predicting teacher evaluations from thin slices of nonverbal behavior and physical attractiveness. *Journal of Personality and Social Psychology*, *64*, 431 – 441.

Ambady, N. E., & Skowronski, J. J. (2008). *First impressions*. New York, NY: Guilford Press.

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

Carver, R. P. (1992). Reading rate: Theory, research, and practical implications. *Journal of Reading, 36*(2), 84 – 95.

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 Inc., Partnership for 21st Century Skills, Corporate Voices for Working Families, & Society for Human Resource Management.

Chen, L., Feng, G., Joe, J., Leong, C. W., Kitchen, C., & Lee, C. M. (2013, November). *Toward automated assessment of public speaking skills using multimodal cues*. Paper presented at the 16th International Conference on Multimodal Interaction, Istanbul, Turkey.

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

Educational Testing Service. (2013). *Quantitative market research*. Unpublished PowerPoint slides.

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

Leung, C., & Mohan, B. (2004). Teacher formative assessment and talk in classroom contexts: Assessment as discourse and assessment of discourse. *Language Testing*, *21*, 335 – 359. https://doi.org/10.1191/0265532204lt287oa

Mason, A. E., Sbarra, D. A., & Mehl, M. R. (2010). Thin-slicing divorce: Thirty seconds of information predict changes in psychological adjustment over 90 days. *Psychological Science*, *21*, 1420 – 1422.

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

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

Richeson, J. A., & Shelton, J. N. (2005). Brief report: Thin slices of racial bias. *Journal of Nonverbal Behavior*, *29*, 75 – 86. https://doi.org/10.1007/s10919-004-0890-2

Ritzhaupt, A. (2008). *Effects of time-compressed audio and adjunct images on learner recall, recognition, and satisfaction* (Unpublished doctoral dissertation). University of South Florida, Tampa. Retrieved from http://scholarcommons.usf.edu/etd/477

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

Williams, R. G., Klamen, D. A., & McGaghie, W. C. (2003). Cognitive, social and environmental sources of bias in clinical performance ratings. *Teaching and Learning in Medicine*, *15*, 270 – 292.

## Appendix A: Rating Forms

| **File Name**: | | | |
|---|---|---|---|
| | | Score | Confidence Rating |
| Condition | Performance Standard | 0 to 4 | 0 to 100% |
| Transcript-only | 1. Formulates an introduction that orients audience to topic and speaker | | |
| Transcript-only | 2. Uses an effective organizational pattern | | |
| Transcript-only | 3. Develops a conclusion that reinforces the thesis and provides psychological closure | | |
| Transcript-only | 4. Demonstrates a careful choice of words | | |
| Transcript-only | 5. Constructs an effectual persuasive message with credible evidence and sound reasoning | | |

**Figure A1** Transcript condition rating form for content-related dimensions.

| **File Name**: | | | | | | | |
|---|---|---|---|---|---|---|---|
| | | After Clip 1 | | After Clips 1 and 2 | | After Clips 1, 2, & 3 | |
| | | Score | Confidence Rating | Score | Confidence Rating | Score | Confidence Rating |
| Video-only | Performance Standard | 0 to 4 | 0 to 100% | 0 to 4 | 0 to 100% | 0 to 4 | 0 to 100% |
| Condition 1 | 1. Effectively uses vocal expression and paralanguage to engage the audience | | | | | | |
| Condition 1 | 2. Demonstrates nonverbal behavior that supports the verbal message | | | | | | |
| Condition 1 | 3. Successfully adapts the presentation to the audience | | | | | | |
| Condition 1 | 4. Skillfully makes use of visual aids | | | | | | |
| Condition 1 | Holistic Score | | | | | | |

**Figure A2** Thin-slicing Condition 1 (random or sentence slicing) rating form for delivery-related dimensions.

| **File Name**: | | | | | | | |
|---|---|---|---|---|---|---|---|
| | | After Clip 1 | | After Clips 1 and 2 | | After Clips 1, 2, & 3 | |
| | | Score | Confidence Rating | Score | Confidence Rating | Score | Confidence Rating |
| Video-only | Performance Standard | 0 to 4 | 0 to 100% | 0 to 4 | 0 to 100% | 0 to 4 | 0 to 100% |
| Condition 2 | 1. Effectively uses vocal expression and paralanguage to engage the audience | | | | | | |
| Condition 2 | 2. Demonstrates nonverbal behavior that supports the verbal message | | | | | | |
| Condition 2 | 3. Successfully adapts the presentation to the audience | | | | | | |
| Condition 2 | 4. Skillfully makes use of visual aids | | | | | | |
| Condition 2 | Holistic Score | | | | | | |

**Figure A3** Thin-slicing Condition 2 (random or sentence slicing) rating form for delivery-related dimensions.

## Appendix B: A Sample Synchronized Multimedia Integration Language File

```
<?xml version="1.0"?>
<!DOCTYPE smil PUBLIC "-//W3C//DTD SMIL 3.0 Language//EN"
                      "http://www.w3.org/2008/SMIL30/SMIL30Language.dtd">
<smil xmlns="http://www.w3.org/ns/SMIL" version="3.0" baseProfile="Language">
 <head>
  <meta name="title" content="Video Tests"/>
  <layout>
   <root-layout xml:id="root-layout" backgroundColor="white" width="100%" height="100%"/>
   <region xml:id="Title" left="10%" width="80%" top="10" height="20"/>
   <region xml:id="Video" left="5%" top="35" width="90%" height="80%"  fit="meet"/>
   <region xml:id="Text" left="10%" top="70%" width="90%" height="30%"/>
  </layout>
 </head>
 <body>
  <par>
   <smilText region="Title" textColor="navy" textFontSize="16">
```

```
LabMMI_006_c_10-22-2013
   </smilText>
   <seq>
<seq>
      <par>
        <video xml:id="v1" region="Video" src="LabMMI_006_c_10-22-2013.mp4"
         clipBegin="10.728938s" dur="17.69225s"/>
        <smilText region="Text" textColor="blue" textFontSize="14" dur="17.69225s">
         Clip 1
        </smilText>
      </par>
      <par>
        <video xml:id="v2" region="Video" src="LabMMI_006_c_10-22-2013.mp4"
         clipBegin="157.471313s" dur="16.917062s"/>
        <smilText region="Text" textColor="blue" textFontSize="14" dur="16.917062s">
         Clip 2
        </smilText>
      </par>
      <par>
        <video xml:id="v3" region="Video" src="LabMMI_006_c_10-22-2013.mp4"
         clipBegin="273.3995s" dur="21.957813s"/>
        <smilText region="Text" textColor="blue" textFontSize="14" dur="21.957813s">
         Clip 3
        </smilText>
      </par>
      <par>
        <video xml:id="v4" region="Video" src="LabMMI_006_c_10-22-2013.mp4"
         clipBegin="230.027125s" dur="13.596125s"/>
        <smilText region="Text" textColor="blue" textFontSize="14" dur="13.596125s">
         Clip 4
        </smilText>
      </par>
</seq>
    <seq>
    <audio src="bell.wav"/>
    <smilText region="Text" textColor="red" textFontSize="14" dur="indefinite"
     repeatcount="4" end="activateEvent">
         Please rate the candidate's performance. Click here when done.
        </smilText>
    </seq>
<seq>
      <par>
        <video xml:id="v5" region="Video" src="LabMMI_006_c_10-22-2013.mp4"
         clipBegin="112.164375s" dur="10.666125s"/>
        <smilText region="Text" textColor="blue" textFontSize="14" dur="10.666125s">
         Clip 5
        </smilText>
      </par>
      <par>
        <video xml:id="v6" region="Video" src="LabMMI_006_c_10-22-2013.mp4"
         clipBegin="147.451938s" dur="8.18325s"/>
        <smilText region="Text" textColor="blue" textFontSize="14" dur="8.18325s">
         Clip 6
        </smilText>
      </par>
      <par>
        <video xml:id="v7" region="Video" src="LabMMI_006_c_10-22-2013.mp4"
         clipBegin="300.443438s" dur="7.626437s"/>
        <smilText region="Text" textColor="blue" textFontSize="14" dur="7.626437s">
         Clip 7
        </smilText>
      </par>
      <par>
        <video xml:id="v8" region="Video" src="LabMMI_006_c_10-22-2013.mp4"
         clipBegin="262.296625s" dur="8.857125s"/>
```

```
        <smilText region="Text" textColor="blue" textFontSize="14" dur="8.857125s">
         Clip 8
        </smilText>
      </par>
</seq>
    <seq>
    <audio src="bell.wav"/>
    <smilText region="Text" textColor="red" textFontSize="14" dur="indefinite"
    repeatcount="4" end="activateEvent">
         Please rate the candidate's performance. Click here when done.
        </smilText>
    </seq>
<seq>
      <par>
        <video xml:id="v9" region="Video" src="LabMMI_006_c_10-22-2013.mp4"
         clipBegin="105.543063s" dur="4.665687s"/>
        <smilText region="Text" textColor="blue" textFontSize="14" dur="4.665687s">
         Clip 9
        </smilText>
      </par>
      <par>
        <video xml:id="v10" region="Video" src="LabMMI_006_c_10-22-2013.mp4"
         clipBegin="206.729563s" dur="8.88625s"/>
        <smilText region="Text" textColor="blue" textFontSize="14" dur="8.88625s">
         Clip 10
        </smilText>
      </par>
      <par>
        <video xml:id="v11" region="Video" src="LabMMI_006_c_10-22-2013.mp4"
         clipBegin="216.821375s" dur="5.396438s"/>
        <smilText region="Text" textColor="blue" textFontSize="14" dur="5.396438s">
         Clip 11
        </smilText>
      </par>
      <par>
        <video xml:id="v12" region="Video" src="LabMMI_006_c_10-22-2013.mp4"
         clipBegin="1.3225s" dur="6.550751s"/>
        <smilText region="Text" textColor="blue" textFontSize="14" dur="6.550751s">
         Clip 12
        </smilText>
      </par>
</seq>
    <seq>
    <audio src="bell.wav"/>
    <smilText region="Text" textColor="red" textFontSize="14" dur="indefinite"
    repeatcount="4" end="activateEvent">
         Please rate the candidate's performance. Click here when done.
        </smilText>
    </seq>
    <seq>
    <smilText region="Text" textColor="red" textFontSize="10" dur="10s" repeatcount="4"
    end="activateEvent">
         Task complete. This message will disappear in 5 seconds.
        </smilText>
    <smilText region="Text" textColor="red" textFontSize="10" dur="10s" repeatcount="4"
    end="activateEvent">
         Please close the program.
        </smilText>
    </seq>
    </seq>
  </par>
 </body>
</smil>
```

## Appendix C: Scatterplots for Content-Related Dimensions in the Transcript Condition

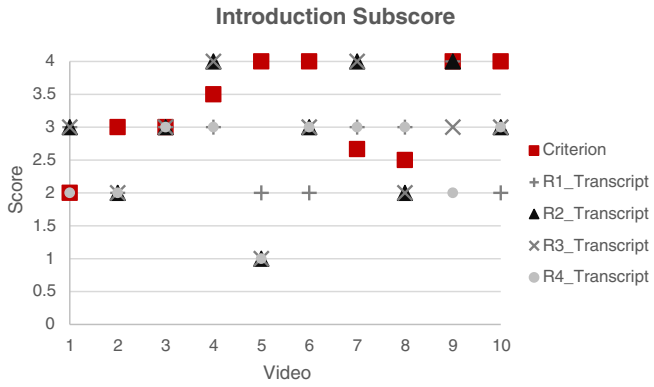


**Figure C1** Introduction criterion and transcript-based scores across raters for the 10 videos.

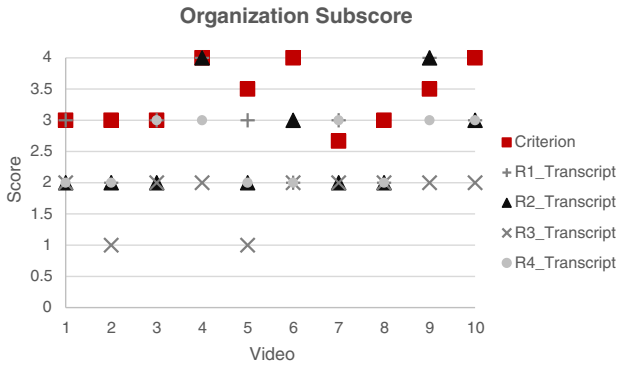


**Figure C2** Organization criterion and transcript-based scores across raters for the 10 videos.

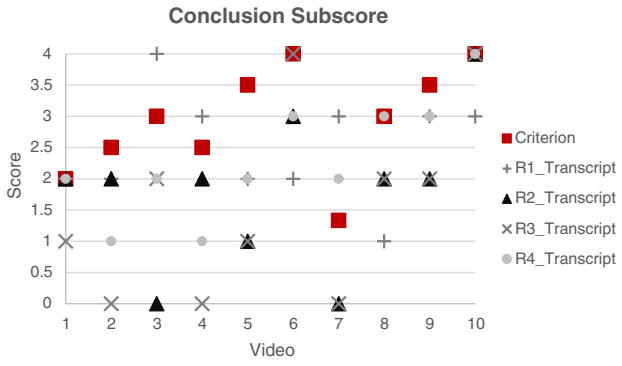


**Figure C3** Conclusion criterion and transcript-based scores across raters for the 10 videos.

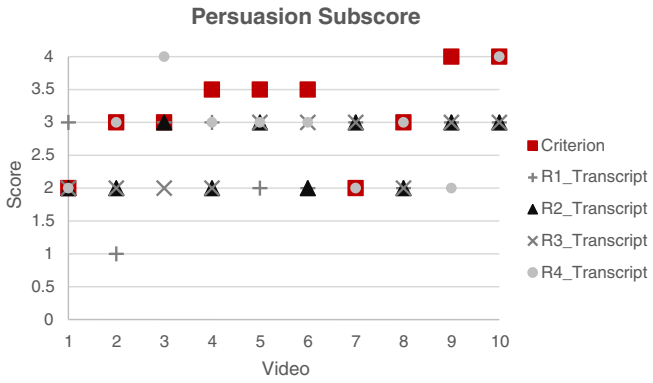


**Figure C4** Persuasion criterion and transcript-based scores across raters for the 10 videos.

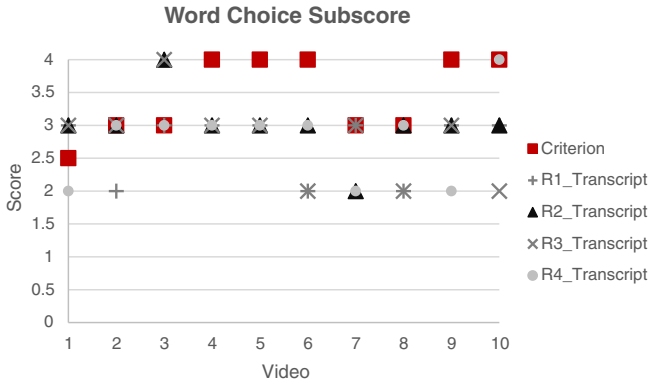


**Figure C5** Word choice criterion and transcript-based scores across raters for the 10 videos.

## Appendix D: Scatterplots for Delivery-Related Dimensions in the Thin-Sliced Conditions

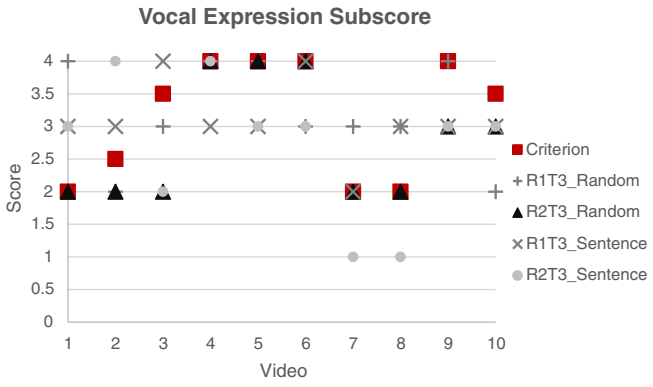


**Figure D1** Vocal expression criterion and thin-sliced (random and sentence) scores across raters for the 10 videos.

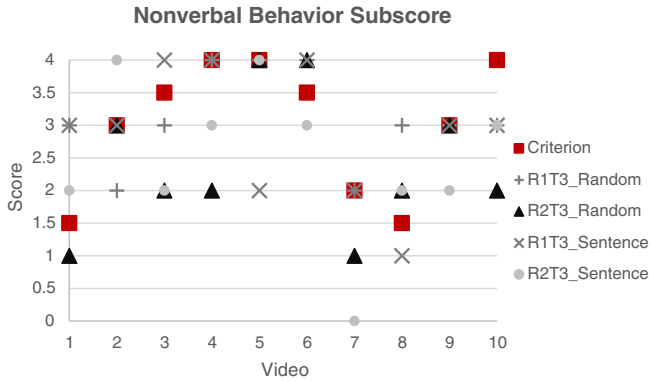


**Figure D2** Nonverbal behavior criterion and thin-sliced (random and sentence) scores across raters for the 10 videos.

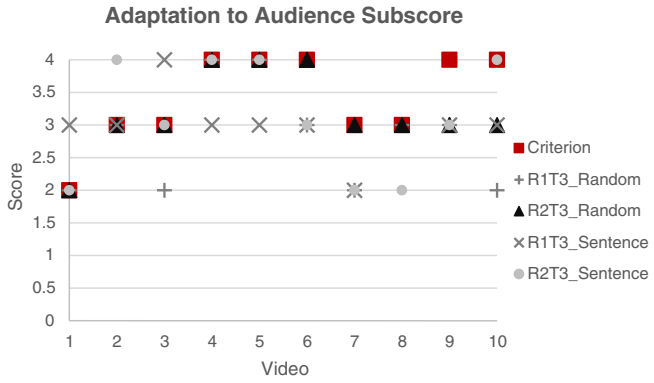


**Figure D3** Adaptation to audience criterion and thin-sliced (random and sentence) scores across raters for the 10 videos.

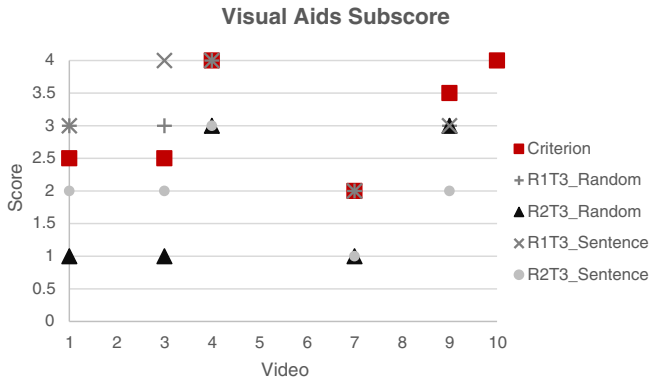


**Figure D4** Visual aids criterion and thin-sliced (random and sentence) scores across raters for the 10 videos. Note that some videos were not scored on this dimension.

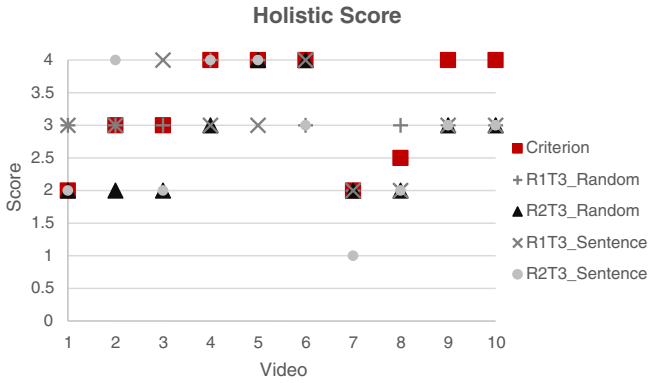


**Figure D5** Holistic criterion and thin-sliced (random and sentence) scores across raters for the 10 videos.

**Suggested citation:**

Feng, G., Joe, J., Kitchen, C., Mao, L., Roohr, K. C., & Chen, L. (2019). *A proof-of-concept study on scoring oral presentation videos in higher education* (Research Report No. RR-19-22). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/ets2.12256

**Action Editor:** Beata Beigman Klebanov

**Reviewers:** Aliaksei Ivanou and Blair Lehman