The IALLT Journal
A publication of the International Association for Language Learning Technology

EFFECTS OF TECHNOLOGY MODES ON RATINGS OF LEARNER RECORDINGS

Elizabeth (Betsy) Lavolette
Michigan State University

Abstract

While research has investigated the effect of visuals in tests of listening comprehension (e.g., Suvorov, 2009; Wagner, 2008, 2010), student-recorded video for oral formative assessment is relatively unexplored. In this study, I examined 15 teachers' ratings of speech recorded by 39 ESL learners to see if teachers assess speech differently depending on whether it is presented with visuals. The learners recorded 4 speech samples: 2 with webcams, 2 with microphones only. A third speech condition was created by removing the video track from the webcam recordings, resulting in 3 conditions and 6 samples for each individual. The teachers rated all 6 samples. I used repeated-measures ANOVAs to determine whether the teachers assigned significantly different scores based on the speech conditions. The results showed that the teachers rated the audio stripped from the video significantly higher than the video/audio recordings ($p = .004$, $d = .38$). This suggests that teachers may be biased in favor of audio-only recordings and that teachers should not give students an option of making either an audio or video recording for a given formative assessment. Further analyses examined how the students' and teachers' preferences for audio-only or video recordings were related to the ratings.

Vol. 43 (2) 2013

INTRODUCTION

Teachers use web-based, oral-skills assessment tools for a number of reasons, including formative and summative assessment. Summative assessments are used to measure what students have learned and often are used after instruction is completed (Brown, 2004). In contrast, formative assessments are learning activities that provide feedback on students' performances as part of instruction (Black & Wiliam, 1998; Winke, 2010) and are more likely to be informal (Brown, 2004). A student's in-class or homework assignment to record a short audio or video response to a prompt may thus be considered a type of formative assessment of oral skills. Given the already informal nature, teachers may be more likely in formative assessments of oral skills to allow students the freedom to choose whether they want to record their speech using a microphone only or a webcam that includes a microphone.

When teachers do allow students the option to record with a webcam or with a microphone only, the construct of speaking may differ depending on the student's choice in recording. To determine whether the practice of letting students choose their assessment mode is fair, in this study I investigate whether the modality of oral test recordings (audio-only or video) affects assessment scores. I also investigate whether learner and teacher preferences are related to the ratings given to recordings, that is, whether students who prefer making video recordings receive higher ratings on video as compared to audio recordings and whether teachers who prefer rating video recordings give higher ratings on video as compared to audio recordings.

Before explaining the current study in depth, I first explore the formative assessment of oral skills and argue for the appropriateness of using video for this purpose. Then I examine the literature on teachers' ratings of audio and video recordings. Next, I discuss learner and teacher preferences for audio and video recordings. Finally, I outline the need for the current research.

Formative Assessment of Oral Skills

In addition to the comments and suggestions that teachers often given learners during formative assessment (Brown, 2004), the formative assessment of oral skills may also include ratings. These ratings can be assigned using either holistic or analytic scales. Holistic scales sum up an impression of the test-taker's ability in a single score (Luoma, 2004, p. 60). Some commonly used holistic

scales are those used to score ACTFL Oral Proficiency Interviews and the rubrics used for the TOEFL independent and integrated speaking sections (Educational Testing Service, 2008a). Analytic scales, on the other hand, give a separate score for each of several criteria, resulting in multiple scores (Luoma, 2004, p. 68). An analytic rubric could be designed for rating a speaker's performance on such criteria as content, pronunciation, and syntactic complexity.

Bachman and Palmer (2010) argued that unlike analytic scales, holistic rating scales are difficult to interpret because "many rating scales that are called 'global' or 'holistic' include multiple components, with little or no indication as to how these different components are to be considered in either arriving at a single rating or in interpreting it" (p. 341). However, holistic scales may be faster and simpler for teachers to use, making them more practical for use in formative assessments.

Most classroom rubrics for assessing speaking are likely to omit nonverbal behavior. For example, in the descriptions of the proficiency levels for speaking, neither the ACTFL guidelines (American Council on the Teaching of Foreign Languages, 2012) nor the Common European Framework of References for Languages (Council of Europe, 2001) mention nonverbal behavior. Therefore, when assigning a formative assessment of oral skills as homework, teachers may think that audio recordings are a more appropriate assessment instrument than video recordings. However, I argue that video recordings are also appropriate for the same purpose because of their ability to capture nonverbal behavior.

Why Video is Appropriate for Assessing Oral Skills

Given that most classroom speaking rubrics omit nonverbal behavior, an appropriate way to assess speaking seems to be via audio recordings. On the other hand, assessing communicative competence in speaking includes assessing nonverbal behaviors and thus requires that the evaluator be able to see the learner.

The notion of communicative competence, as developed by Canale and Swain (e.g., Canale, 1983; Canale & Swain, 1980), includes grammatical competence, sociolinguistic competence, and strategic competence. Grammatical competence covers linguistic knowledge, and sociolinguistic competence covers sociocultural rules of speech. Strategic competence, which is of the greatest interest in the current study, is explained as follows:

> This component is composed of mastery of verbal and non-verbal communication strategies that may be called into action for two main reasons: (a) to compensate for breakdowns in communication due to limiting conditions in actual communication (e.g. momentary inability to recall an idea or grammatical form) or due to insufficient competence in one or more of the other areas above; and (b) to enhance the effectiveness of communication (e.g. deliberately slow and soft speech for rhetorical effect). (Canale, 1983, pp. 10–11)

This definition of strategic competence includes not only verbal strategies for repairing breakdowns in communication, but also nonverbal strategies for enhancing the effectiveness of communication. Examples of these nonverbal strategies include eye contact, facial expressions, and gestures. In agreement with Canale (1983), Pennycook (1985) and Neu (1990) also suggested that nonverbal behavior be included as a primary component of communicative competence.

To understand why nonverbal strategies are an important part of communicative competence, consider that the listener can see the speaker in most speaking situations (except, for example, when talking on the phone) and that the speaker's nonverbal behavior can contribute to the message's meaning. Gullberg (2006) provided an example of a heavily meaningful gesture that was naturally incorporated into speech: "He went [ ] and everybody laughed" (p. 106). The meaning is opaque without being able to see the gesture. Skilled communicators can use nonverbal behavior to add to what they say, and incongruence between what speakers say and do can undermine the verbal message from the perspective of the listener (Neu, 1990). Tellier (2006), as cited by Gullberg (2008), found that French children who were taught vocabulary using gestures along with explanations performed better if they reproduced the gestures, which indicates that the production of appropriate gestures may aid second language acquisition. Stam (2007) went so far as to claim that learners' acquisition of their L2 can only be judged if both speech and gestures are examined. In addition, von Raffler-Engel (1980) argued that verbal language and its corresponding nonverbal behavior should be taught together, rendering obsolete the system of teaching verbal language in isolation.

Thus, video-based speaking assessments may be preferred to audio-only ones. A further argument for using face-to-face or video assessments comes from drawing a parallel with the assessment of listening skills. The ACTFL Guidelines (American Council on the Teaching of Foreign Languages, 2012) at the novice

level mention, "They [novice learners] rely heavily on extralinguistic support to derive meaning" (p. 19). If extralinguistic support (e.g., nonverbal behavior) is part of the listening guidelines, it follows that it should also be produced by learners when speaking. That is, educators should expect learners to use nonverbal behaviors that complement their speaking. In addition, as Kellerman (1992), Coniam (2001), and Wagner (2007, 2008, 2010) have argued, video texts are more appropriate for use in testing listening comprehension because most listening occurs in a situation in which the speaker is visible. If we extend this argument to include productive skills, using video or face-to-face conditions is appropriate to test speaking, allowing learners to display not only their oral language skills, but complete communicative competence, including nonverbal behaviors.

Influences on Ratings of Learners' Speech

Beyond the differences in modality between audio and video recordings, many factors may influence the ratings of learners' recordings, including the learners' nonverbal behavior, which is only visible in video recordings; learners' preferences for making recordings in either mode; raters' preferences for rating in either mode; and ethnic characteristics of learners, which may be more salient in video versus audio recordings.

In language testing, researchers have examined the influence of nonverbal behavior on rating. Nambiar and Goon (1993) found that the nonverbal behavior of examinees during face-to-face interviews positively influenced the ratings compared to the ratings of the audio recordings of the same interviews. Kenyon and Malabonga (2001) investigated the differences in L2 Spanish students' opinions on a face-to-face oral proficiency interview (OPI) and two versions of recorded OPIs: a simulated OPI (SOPI) recorded on tape and a computerized OPI (COPI) recorded on a computer. They found that the students thought that the face-to-face OPI allowed them a better opportunity to demonstrate their language abilities and was more accurate and fair. The ratings on the face-to-face OPI were lower than on either the SOPI or COPI. Although the authors did not discuss the fact that nonverbal behavior of the test takers was visible in the face-to-face OPIs but not the SOPI or COPI, it is possible that this behavior had an influence on these ratings. Jenkins (2003) found that test takers' nonverbal behavior in an interview influenced raters' comprehensibility judgments. The test takers who exhibited nonverbal behaviors similar to those of North Americans were judged more comprehensible than the test takers who did not. While these studies have

begun to explore the influence of nonverbal behaviors on ratings, more studies are needed to improve our understanding.

Little research has been conducted on learners' and teachers' preferences for or use of audio versus video recordings for formative or proficiency-based oral skills assessments. A few studies have asked learners whether they prefer to use audio or video texts for listening comprehension tests, and some of these studies have looked at students' performance. Progosh (1996) found that Japanese college students overwhelmingly (92%) preferred video over audio texts. Sueyoshi and Hardison (2005) investigated the effects of three conditions (audio only, audio with visuals of a face, and audio with visuals of a face and gestures) on ESL learners' scores on a listening comprehension test. They found that the learners preferred to see visual cues during listening comprehension tasks and that the scores were significantly higher for both conditions that included visuals.

Londe (2009) examined the effects of audio and two different types of videos on ESL students' comprehension of a lecture. One video was a close-up of the instructor's face, while the other was shot from a distance that showed the instructor's body, a blackboard, and three students listening. No differences were found between the results of a comprehension test after students saw or listened to the lecture in the three conditions. Wagner (2010), on the other hand, found that ESL students scored higher on a listening comprehension test when the text was provided as video rather than audio. Suvorov (2009) obtained yet different results: ESL learners scored significantly lower on the portion of a listening test that used a video text compared to the portions that used an audio text and photos with audio.

Coniam (2001) tested English language teachers in Hong Kong using audio and video versions of the same text and found no significant differences between the scores of the two groups on a test of short- and extended-answer listening comprehension questions. The researcher also found that test takers who had taken video tests noted that they would have preferred audio, while some audio test takers indicated that they would have preferred video.

Outside of testing, Tian (2011) compared the oral proficiency of learners of Mandarin Chinese who participated in 12 half-hour audio or video conferences with native Mandarin speakers during a month. The learners who used video chat improved their proficiency significantly more than those who used audio chat. This indicates that visual elements in the video chat helped the learners improve their speaking and perhaps their listening skills. The researcher also found that

88% of the Mandarin Chinese learners who participated in the video conferencing thought that seeing body language was useful in understanding their native-speaker partners. Among the learners who participated in the audio conferencing, 68% thought that being unable to see their partners' body language hindered their comprehension. Further, the researcher speculated that the presence of web cameras increases students' fears of making mistakes or being embarrassed.

Finally, rating conditions may differ between audio and video modes due to teachers' ethnic or racial biases. For example, Rubin (1992) and Rubin and Smith (1990) found that simply seeing a face that was identified as Asian caused undergraduate students to identify a speaker as having a foreign accent, even when none existed. Raters may be more readily able to recognize ethnicities when presented with a video than when listening to audio only. Teachers may therefore be more strongly reminded of the ethnicities of their students when they watch videos than when they listen to audio clips, which suggests that using video over audio-only recordings may change the rating conditions in oral assessment.

THE CURRENT STUDY

Research is lacking on classroom use of audio and video recording modes for formative oral speech assessment. Therefore, the current study is designed to give teachers insight into their possible biases in rating video and audio formative speaking assessments for their classrooms. The first research question is as follows:

1. Do teachers rate audio and video recordings differently?

The outcome of this research question cannot be predicted based on previous research. If the ratings do differ between audio and video modes, the next question to be addressed is why they differ. Students' preferences for recording audio or video and the teacher's preference for rating audio or video may affect ratings. For example, learners who prefer recording audio may receive higher ratings in that mode. It is also possible that teachers who prefer rating videos will give higher ratings to videos. Thus, two further research questions are as follows:

2. Do students with preferences for recording audio or video receive differing ratings on the two types of recordings?

3. Do teachers with preferences for rating audio or video rate the two types of recordings differently?

A reasonable prediction is that students will perform better on the mode that they prefer and thus get higher ratings in that mode and that teachers will assign higher ratings to the mode that they prefer.

Finally, if student and teacher preferences are related to the ratings, it is possible that a combination of these preferences could compound the difference. Therefore, the final research question is the following:

4. Do teachers with preferences for rating audio or video differently rate the types of recordings made by students who share their preferences?

I predict that teacher and student preferences will have an additive effect and that ratings in the condition where teachers and students have the same preference will be higher in the preferred mode.

METHOD

Participants

Speech samples were recorded by 39 English-language learners enrolled in an English language center at a large Midwestern university. Nineteen of the learners were at the intermediate level and 14 were at the upper-intermediate level in intensive English courses; 3 were at the advanced level in English-for-academic-purposes courses (already enrolled in the university); and the levels of 3 were not reported. The average age was 20, with a range from 18 to 43. The majority were male (27 male, 12 female), and the largest L1 groups were Chinese (16) and Arabic (15), with the rest natively speaking Korean (4), Japanese (1), and Kurdish (1). Three participants did not specify an L1, and one participant indicated two L1s (Chinese and Japanese).

Twenty students in a graduate language-assessment class rated the speech samples. These students were pursuing PhD degrees in second language studies, MA degrees in TESOL, or were practicing teachers in K-12 schools pursuing higher degrees within education. All of them had experience teaching languages. Fifteen of the raters completed the ratings, and 14 of these raters completed an exit questionnaire. Of these 14 raters, the mean age was 28. Eight were male, and six were female. Seven of them had English as an L1, and one rater each reported

an L1 of Afrikaans, Arabic, French, German, Japanese, Spanish, and Ukrainian (one rater reported both English and French as L1s).

Materials

I used TOEFL iBT Test Independent Speaking prompts (Educational Testing Service, n.d., 2006) to elicit audio and video recordings from the English-language learners. For the purposes of rating, I created a third type of recording by removing the video track from the video recordings to test the effects of the audio alone on the teachers' ratings.

The rubric that was used to score the recordings was the iBT TOEFL Independent Speaking rubric (Educational Testing Service, 2008b), which is used to assign holistic ratings from 0 to 4. Both the learners and the raters filled out demographic questionnaires.

Procedure

Collection of learner data

The learners came to a computer lab outside of their normal class time. I explained the study and demonstrated the technology that was used. Then, the learners recorded three audio and three video speech samples (45 seconds each) in response to the TOEFL prompts, with the order of the prompts and technology counterbalanced. They used the Audio and Video Dropboxes applications developed by the Center for Language Education and Research at Michigan State University (www.clear.msu.edu) to make the recordings. The first two samples (one audio and one video) were used for the learners to practice using the technology and were not rated. The learners read the prompts on the computer screen, then were allowed to make as many practice recording attempts as they liked. The time for a response was limited to 45 seconds, as in the TOEFL test. After completing the recordings, the learners completed a questionnaire on their opinions of the two modes of recording and a background questionnaire.

Collection of ratings

The teachers were trained using normed samples of learner speech and were retrained on three new samples at the beginning of each rating session. They

rated the audio recordings, video recordings, and the audio portions of the video recordings (with the video track removed) using the ETS rubric. Each rater rated each recording. They also filled out a background questionnaire and a questionnaire on their opinions of rating the audio and video files. Most of the rating took place over the course of a month during the language assessment class; three of the raters finished the rating after the class had ended.

Analysis

I calculated intrarater reliabilities as a preliminary assessment of whether the teachers were rating the video files and their corresponding audio files in the same way. To do this, I analyzed the mean scores that the teachers gave on the video versus audio modes using analysis of variance (ANOVA) to see if they assigned significantly different scores based on modality and if so, how much of an effect on the scores the modality had; that is, if one mode resulted in higher scores, how *much* higher were those scores? I also used ANOVA to analyze separately the mean scores for the students who indicated that they preferred to record audio or video to see if their preferences had an effect on their scores, and if so, how much of an effect. Similarly, I used ANOVA to analyze separately the mean scores assigned by the raters who indicated that they preferred to rate audio or video files to see if their preferences had an effect on the scores they gave, and if so, how much of an effect.

RESULTS

Do Teachers Rate Audio and Video Recordings Differently?

To begin to address this question, I calculated the intrarater reliability (Table 1, third column) by correlating each teacher's rating of a video file with his or her rating of the corresponding audio file, which was created by removing the video picture from the video recording. These should be rated the same if the raters are consistent and if video-based, visual information (nonverbal behavior) is not considered part of the speaking construct; and note that nonverbal behavior was not included on the rubric used for this study. The intrarater reliabilities ranged from .120 to .728.

Rater	Interrater reliability	Intrarater reliability
R1	.555	.499
R2	.691	.496
R3†	.706	.635
R4	.664	.613
R5	.690	.521
R8	.474	.120 (n.s.)
R9†	.717	.660
R10†	.751	.524
R11	.696	.728
R14†	.730	.676
R16†	.719	.711
R17†	.732	.577
R18	.694	.418
R19	.571	.566
R20	.675	.616

Table 1: Interrater/Intrarater Reliabilities*

I used repeated-measures ANOVAs to determine whether the teachers assigned significantly different scores based on the condition. The means and standard deviations are shown in Table 2. For all ANOVAs reported below (Table 3), Mauchly's test indicated that the assumption of sphericity had been violated, so the results of multivariate tests (Pillai's trace) are reported (Field,

* Note. All reliabilities in the table are significant at the .01 level except for the intrarater reliability of R8, which is not significant.

† More reliable raters (reliability above .7).

2009). All post-hoc pairwise comparisons were computed using the conservative Bonferroni adjustment.

	All raters		More reliable raters		Less reliable raters	
	Mean	*SD*	Mean	*SD*	Mean	*SD*
Audio	2.49	0.47	2.52	0.54	2.47	0.42
Audio track from video	2.35	0.64	2.49	0.72	2.24	0.59
Video	2.28	0.60	2.31	0.69	2.25	0.55

Table 2: Mean Ratings and Standard Deviations for All Raters, More Reliable Raters, and Less Reliable Raters

	All raters	More reliable raters	Less reliable raters
V	.35	.47	.16
F(2,32)	8.43	14.27	3.01
p	.001	<.001	.063

Table 3: Results of ANOVAs for All Raters, More Reliable Raters, and Less Reliable Raters

The ANOVA for the 15 teachers was significant, $V = .35$, $F(2, 32) = 8.43$, $p = .001$. Pairwise comparisons (Table 4) showed a significant difference between the video and the audio track from the video, $p = .004$, with the ratings for the audio being higher, $d = .38$ (a small effect size).

	All raters		More reliable raters		Less reliable raters	
	Audio track from video	Video	Audio track from video	Video	Audio track from video	Video
Audio	$p = .49$	$p = .077$	$p = 1.00$	$p = .14$	$p = .066$	$p = .054$
Audio track from video	-	$p = .004$	-	$p < .001$	-	$p = 1.00$

Table 4: Results of Pairwise Comparisons for All Raters, More Reliable Raters, and Less Reliable Raters

To rule out the possibility that the significant difference found in the ANOVA analysis above was due to the raters being unreliable, I separated the raters into two groups: more and less reliable. First, I calculated interrater reliabilities by correlating each rater's ratings with a common set of ratings, which was composed of the average ratings given by all of the raters. I used that information to classify the teachers into the more reliable (Spearman's rho > .7) and less reliable (< .7) groups. The more reliable raters are indicated with a dagger in Table 1.

The ANOVA for the more reliable raters was significant, $V = .47$, $F(2,32) = 14.27$, $p < .001$. Similarly to the raters overall, these teachers rated the audio stripped from the video significantly higher than the video, $p < .001$, $d = .67$ (a medium effect size). No other comparisons were significant.

The ANOVA for the less reliable raters approached significance, $V = .16$, $F(2, 32) = 3.1$, $p = .063$. The pairwise comparisons also did not reach significance, and the differences that approached significance were not the same as those of the more reliable raters and for the raters overall. For the less reliable raters, the ratings for the audio-only recordings (when only a microphone was used) were higher than the ratings for both the audio stripped from video, $p = .066$, $d = .50$ (a medium effect size) and the video recordings, $p = .054$, $d = .49$ (a medium effect size), with the differences approaching significance.

Do Students With Preferences for Recording Audio or Video Receive Differing Ratings?

I split the students into two groups based on their responses to a questionnaire item that asked them which recording mode they liked better, audio

or video. Of the 39 students, 24 (61.5%) preferred audio, 19 of whom provided both audio and video recordings and are thus included in the analysis below. Fifteen (38.5%) of the students preferred video, and 13 of these provided both audio and video recordings and are included in the analysis below as well.

I used repeated-measures ANOVAs to determine whether the students who preferred making a certain type of recording received significantly different ratings based on the recording condition. The hypothesis I wanted to test was that if a student preferred using a microphone only to record, he or she would more likely receive a higher rating on audio-only recordings. Alternatively, if the student preferred using a webcam, he or she should receive a better test score when using a webcam. The means and standard deviations are shown in Table 5.

	Prefer audio		Prefer video	
	Mean	*SD*	Mean	*SD*
Audio	2.60	0.37	2.36	0.55
Audio track from video	2.47	0.74	2.22	0.49
Video	2.35	0.71	2.19	0.43

Table 5: Mean Ratings and Standard Deviations for Students Who Preferred Recoding Audio or Video

The ANOVA for the students who preferred audio was significant, $V = .642$, $F(2, 17) = 15.27$, $p < .001$ (Table 6). However, the pairwise comparisons (Table 7) showed a significant difference between only the ratings for video and audio stripped from video, $p = .001$, $d = 0.73$ (a medium effect size), with the audio tracks minus video being rated higher. No other comparisons were significant, and thus there is no evidence that those who preferred using a microphone only receive higher scores when recording with a microphone versus a webcam.

	Prefer audio	Prefer video
V	.642	.115
F	$F(2,17) = 15.27$	$F(2,13) = 0.84$
p	<.001	.45

Table 6: Results of ANOVAs for Students Who Preferred Recording Audio or Video

	Audio track from video	Video
Audio	$p = .97$	$p = .17$
Audio track from video	-	$p = .001$

Table 7: Results of Pairwise Comparisons for Students Who Preferred Recording Audio

The ANOVA for the students who preferred video was not significant, $V = .115$, $F(2, 13) = 0.84$, $p = .45$ (Table 6), so no further comparisons were calculated. This again demonstrates that preference for a certain recording condition was *not* associated with an underlying ability to achieve a higher score in that recording condition.

Do Teachers With Preferences for Rating Audio or Video Rate the Two Types of Recordings Differently?

I divided the teachers into two groups based on their responses to a questionnaire item that asked them which mode they preferred to rate (audio- or video-based). Of the 14 teachers who filled out the exit questionnaire, 7 each preferred audio and video. Thus, the analyses below are with those 14 individuals.

I used repeated-measures ANOVAs (Table 9) to determine whether the two groups of teachers assigned significantly different scores based on the mode. I was investigating whether the teachers gave higher scores to the mode they

preferred. The means and standard deviations are shown in Table 8. The results of the post-hoc pairwise comparisons are shown in Table 10.

	Prefer audio		Prefer video	
	Mean	*SD*	Mean	*SD*
Audio	2.50	0.41	2.50	0.57
Audio track from video	2.45	0.65	2.35	0.65
Video	2.37	0.59	2.26	0.62

Table 8: Mean Ratings and Standard Deviations for Teachers Who Preferred Rating Audio or Video

	Prefer audio	Prefer video
V	.254	.216
$F(2,32)$	5.44	4.40
p	.009	.02

Table 9: Results of ANOVAs for Teachers Who Preferred Rating Audio or Video

	Prefer audio		Prefer video	
	Audio track from video	Video	Audio track from video	Video
Audio	$p = 1.00$	$p = .53$	$p = .46$	$p = .073$
Audio track from video	-	$p = .011$	-	$p = .13$

Table 10: Results of Pairwise Comparisons for Teachers Who Preferred Rating Audio or Video

The ANOVA for the teachers who preferred rating audio was significant, $V = .254$, $F(2, 32) = 5.44$, $p = .009$. Pairwise comparisons (Table 10) showed a significant difference between the ratings for video and audio stripped from video, $p = .011$, $d = 0.71$ (a medium effect size), with the audio tracks being rated higher. However, no other comparisons were significant; they did not rate the microphone-only audio mode higher than the video mode.

The ANOVA for the teachers who preferred rating video was also significant, $V = .216$, $F(2, 32) = 4.40$, $p = .02$. However, the pairwise comparisons (Table 10) showed no significant differences.

Do Teachers With Preferences for Rating Audio or Video Differently Rate the Two Types of Recordings Made by Students Who Share Their Preference?

In the same way as described above, I divided the teachers into two groups based on their preferences for rating audio or video, and I divided the students into two groups based on their preferences for recording audio or video. The numbers of students and teachers in each group are as indicated above. Based on these groups, I examined how the teachers who preferred audio rated the recordings of the students who preferred audio and how the teachers who preferred video rated the recordings of the students who preferred video.

I used repeated-measures ANOVAs to determine whether the two groups of teachers assigned significantly different scores to the two groups of students based on the mode. I was testing whether student and teacher preferences for the same types of recordings were related to higher ratings on those types of recordings. The means and standard deviations are shown in Table 11.

	Prefer audio		Prefer video	
	Mean	*SD*	Mean	*SD*
Audio	2.54	0.44	2.36	0.66
Audio track from video	2.41	0.83	2.20	0.49
Video	2.37	0.79	2.19	0.47

Table 11: Mean Ratings and Standard Deviations for Teachers and Students Who Both Preferred Audio or Video

The ANOVA for the teachers and students who preferred audio was not significant, $V = .072$, $F(2, 17) = 0.66$, $p = .53$. Similarly, the ANOVA for the teachers and students who preferred video was not significant, $V = .080$, $F(2, 13) = 0.562$, $p = .58$, so no further comparisons were performed in either case.

DISCUSSION

Do Teachers Rate Audio and Video Recordings Differently?

Several analyses in this study indicate that the teachers rated audio and video recordings differently. First, the intrarater reliabilities shown in Table 1 are low (compared to the reliability that ETS reported for the iBT TOEFL speaking test of .88, ETS, 2011, p. 5) and vary widely among the teachers, which indicates that the teachers did not give the same ratings to the video files and the corresponding audio files created from those video files.

The three ANOVAs that were calculated using all of the raters, the more reliable raters only, and the less reliable raters only also each indicate possible differences in how the audio and video files were rated. The ANOVA using only the less reliable raters showed that they tended to rate the audio files higher than both the audio files stripped from the video files and the video files. This result is difficult to interpret, however. The raters may have been rating the types of files differently. Another possibility is that the learners performed differently in the two modes. An argument against this possibility, however, is that if it were the case, a similar difference would be expected for the other groupings of raters. That is, for the raters overall and the more reliable raters, we would expect that the microphone-only audio files would be rated higher than both the audio files stripped from the video files and the video files. That was not the case.

Both the raters overall and the more reliable raters rated the audio files stripped from the videos significantly higher than the video files. Because the audio files stripped from the video files contained exactly the same oral performance as the video files, these differences in the ratings can only be due to the teachers seeing the video picture.

Before I discuss how the students' and teachers' preferences were related to the ratings, I examine some alternative possibilities for why the raters were biased against the video recordings. First, the teachers (overall and in the more

reliable group) who rated video files lower than their corresponding audio files stripped from the video may have been distracted by the video, as some of the teachers reported, similarly to how students have reported being distracted by video texts in listening tests (e.g., Coniam, 2001; Ockey, 2007). However, if the raters were distracted, one would expect the ratings for the video files to less reliable, rather than consistently lower. However, the reliabilities of the three types of files showed no patterns of differing reliabilities among the three modes. Thus, this explanation of the biased ratings is unsatisfactory.

Another possible reason for the lower ratings is that the learners were not exhibiting the nonverbal behavior that the teachers expected or thought appropriate (Gullberg, 2008; Neu, 1990), as mentioned by one of the teachers in an informal interview. Further research is needed to investigate this possibility.

One final reason for the ratings to have differed by mode is that the raters were not familiar with the learners who made the recordings, so they may have been influenced by what they saw. That is, if a given rater was biased for or against people from a certain ethnic group, that rater may have been more readily able to recognize that ethnicity when presented with a video than when listening to audio only. Simply seeing a face that was identified as Asian has been shown to cause undergraduates to identify a speaker as having a foreign accent, even when none existed (Rubin & Smith, 1990; Rubin, 1992); the raters in the current study may have been similarly prone to stereotyping students as good or poor speakers based on their ethnicity, which may have been more readily identifiable in a video than an audio recording. Research targeting this speculation is needed.

The current result that the raters were biased against the video-recorded speech samples is similar to the findings of Kenyon and Malabonga (2001), who found that students were scored lower on a face-to-face OPI than on a SOPI or COPI. However, the two studies are not directly comparable because in Kenyon and Malabonga's (2001) study, the differences between the face-to-face and recorded conditions were not only the visibility of nonverbal behaviors; the tests themselves were different. The current results contrast with the findings of Nambiar and Goon (1993), who found that ratings were lower when audio recordings were assessed than when speech was assessed in a face-to-face context. Besides the fact that the current study used video recordings, rather than face-to-face testing, other discrepancies in the conditions may account for the different findings. In Nambiar and Goon's study, the tests were much longer, with each lasting up to 45 minutes, as compared to the four 45-second speech samples collected from each participant in the current study. This gave the

participants in the 1993 study more opportunities to display both their verbal and nonverbal communication skills. Another difference is that the test-takers in Nambiar and Goon's study interacted with either two interviewers in one section of the test or another student in the other section. This interaction may have increased the importance of nonverbal communication. Yet another difference between the studies is the raters. Nambiar and Goon reported little about their raters beyond the fact that they were experienced. For example, we do not know how familiar they were with the nonverbal behavior of native speakers of English. The raters in the current study had all lived in the United States for extended periods of time or were native speakers of American English.

Can Teacher and Student Preferences Explain the Bias in the Ratings?

Further analyses in the current study explored whether student and teacher preferences for audio or video were related to the ratings given. The students who preferred recording audio were rated significantly higher on the audio track stripped from the video than on the video, which is the same discrepancy seen in the ratings of all students' recordings, with a similar (medium) effect size. In addition, no significant difference was seen between the ratings of these students' microphone-only audio recordings and video recordings, which suggests that their preference for audio did not affect the ratings. The students who preferred recording video, on the other hand, were not rated significantly differently on the three recording modes. This suggests that students who prefer making video recordings may perform better in videos, counteracting the bias of raters against video recordings.

The results for the teachers who preferred rating audio or video recordings paralleled those of the students. That is, the teachers who preferred rating audio recordings rated the audio taken from the video files significantly higher than the video files, also with a medium effect size. This suggests that their preference for audio did affect the ratings, although surprisingly, only on one of the audio conditions. The teachers who preferred rating videos showed no significant differences in the pairwise comparisons, which suggests that their preference for video did not affect the ratings.

When both teachers and students preferred audio, there were no significant differences in how the three modes (audio, video, and audio stripped from video) were rated. This is interesting in light of the results above that although the student preferences for making audio recordings did not affect their ratings, the

teachers who preferred rating audio recordings rated the audio tracks stripped from the videos higher than the videos. This bias did not appear in the analysis of the teachers and students who both preferred audio.

When both teachers and students preferred video, there were no significant differences in how the three modes were rated. This fits well with the results that the ratings of students who preferred recording video were not significantly different by mode and that the ratings by teachers who preferred rating videos were not significantly different by mode.

So, can the rating bias be explained by student and teacher preferences? The answer is yes in some cases. When students, teachers, or both preferred recording or rating video, no biases were seen. However, when students or teachers preferred recording or rating audio files, the audio recordings stripped from the video recordings were rated significantly higher than the corresponding video recordings. Because both types of recordings contained the same oral performance, these teachers demonstrated a bias toward the audio mode. It is unclear why this bias did not appear when both students and teachers preferred audio files, but it may be due to the lower power associated with having fewer participants in the analysis.

CONCLUSIONS

Bias is apparent in the teachers' ratings of the three types of recordings. The current study cannot tell us if the learners performed differently depending on the recording mode, but the differences seen in the more reliable raters' ratings of the video tracks and the audio taken from those same video tracks reveals that for these raters, the type of recording itself influenced the ratings, with the audio taken from the videos being rated higher than the videos themselves.

Pedagogical Implications

The purpose of this study was to determine whether the common pedagogical practice of allowing students to make either audio or video recordings for formative oral assessments is fair to students. The results indicate that teachers should not let students choose between audio and video. That is, if the teacher assigns a recording, ratings may differ for students that record using webcams and students that record from computers that have microphones but no video cameras. The reason for the differing ratings may be due to a performance

difference on the part of the students in the differing modes, or the ratings may differ due to differences in how the teacher rates the recordings. Ultimately, audio and video recordings are different assignments, so teachers should not allow students to mix the two.

Despite varied student preferences for recording either audio or video, it is important not to allow them to follow these for a given assignment. Students' preferences for one mode or the other do not seem to be related to the bias. On the other hand, teachers' preferences may be related to the bias, with the results showing that only the teachers who preferred rating audio rated the audio tracks stripped from the video higher than the video recordings. Teachers who prefer rating video recordings may not be biased in their ratings. However, this finding should not be taken as license for teachers who prefer rating video to allow students to make either mode of recording. To completely eliminate the possibility of bias due to the recording mode, all students should make the same type of recording.

Several other factors are important in a teacher's choice to have students use audio or video recording tools for formative oral assessments. Perhaps most importantly, teachers should consider their own definitions of the construct of speaking and whether the nonverbal aspects of communicative competence are important for a particular assessment. If nonverbal aspects are important, they should be included on the rubric used to evaluate students' performance. The ultimate decision, however, is up to the teacher, and he or she should make an informed decision, taking into account the bias found in the current study.

ABOUT THE AUTHOR

Elizabeth (Betsy) Lavolette is a PhD candidate in Second Language Studies at Michigan State University. Her research interests include computer-assisted language learning, online learning, and assessment.

ACKNOWLEDGEMENTS

I am grateful to Dr. Paula Winke and Dr. Senta Goertler for their valuable comments on earlier versions of this manuscript. I also thank Dr. Dennie Hoopingarner and the Center for Language Education and Research for creating the recording tools used in this study and invaluable assistance in processing the resulting files. Last but not least, I thank the students and teachers who participated in the recording and rating.

REFERENCES

American Council on the Teaching of Foreign Languages. (2012). *Proficiency guidelines*. Retrieved from http://actflproficiencyguidelines2012.org/

Bachman, L., & Palmer, A. (2010). *Language assessment in practice*. Oxford, England: Oxford University Press.

Black, P., & Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education*, *5*(1), 7–74. doi:10.1080/0969595980050102

Brown, H. D. (2004). Testing, assessing, and teaching. In *Language assessment: Principles and classroom practices* (pp. 1–18). White Plains, NY: Pearson/Longman.

Burgoon, J. K. (1994). Nonverbal signals. In M. Knapp & G. Miller (Eds.), *Handbook of interpersonal communication* (pp. 229–285). Thousand Oaks, CA: Sage.

Canale, M. (1983). From communicative competence to communicative language pedagogy. In J. C. Richards & R. W. Schmidt (Eds.), *Language and communication* (pp. 2–27). London, England: Longman.

Canale, M., & Swain, M. (1980). Theoretical bases of communicative approaches to second language teaching and testing. *Applied Linguistics*, *1*(1), 1–47.

Coniam, D. (2001). The use of audio or video comprehension as an assessment instrument in the certification of English language teachers: A case study. *System*, *29*, 1–14.

Council of Europe. (2001). *Common European Framework of Reference for languages: Learning, teaching, assessment*. Retrieved from http://www.coe.int/t/dg4/linguistic/cadre_en.asp

Educational Testing Service. (n.d.). TOEFL sample questions. Retrieved from http://www.ets.org/Media/Tests/TOEFL/pdf/SampleQuestions.pdf

Educational Testing Service. (2006). TOEFL iBT Speaking. In *Official Guide to the New TOEFL iBT with CD-ROM* (pp. 207–248). New York, NY: McGraw-Hill.

Educational Testing Service. (2008a). TOEFL iBT Tips: How to prepare for the TOEFL iBT.

Educational Testing Service. (2008b). TOEFL iBT Test: Independent speaking rubrics (scoring standards).

Educational Testing Service. (2011). Reliability and comparability of TOEFL iBT ™ scores. In *TOEFL iBT research insight* (Vol. 3). Princeton, NJ: ETS.

Field, A. (2009). *Discovering statistics using SPSS* (3rd ed.). London, England: Sage.

Gullberg, M. (2006). Some reasons for studying gesture and second language acquisition (Hommage à Adam Kendon). *International Review of Applied Linguistics in Language Teaching*, *44*, 103–124. doi:10.1515/IRAL.2006.004

Gullberg, M. (2008). Gestures and second language acquisition. In P. Robinson & N. C. Ellis (Eds.), *Handbook of cognitive linguistics and second language acquisition* (pp. 276–305). New York, NY: Routledge.

Jenkins, S., & Parra, I. (2003). Multiple layers of meaning in an oral proficiency test: The complementary roles of nonverbal, paralinguistics, and verbal behaviors. *The Modern Language Journal*, *87*(1), 90–107.

Kellerman, S. (1992). "I see what you mean": The role of kinesic behaviour in listening and implications for foreign and second language learning. *Applied Linguistics*, *13*, 239–258.

Kenyon, D. M., & Malabonga, V. (2001). Comparing examinee attitudes toward computer-assisted and other oral proficiency assessments. *Language Learning & Technology*, *5*(2), 60–83.

Londe, Z. C. (2009). The effects of video media in English as a second language listening comprehension tests. *Issues in Applied Linguistics*, *17*(1), 41–50.

Luoma, S. (2004). *Assessing speaking*. Cambridge, England: Cambridge University Press.

Nambiar, M. K., & Goon, C. (1993). Assessment of oral skills: A comparison of scores obtained through audio recordings to those obtained through face-

to-face evaluation. *RELC Journal*, *24*(1), 15–31. doi:10.1177/003368829302400102

Neu, J. (1990). Assessing the role of nonverbal communication in the acquisition of communicative competence in L2. In R. Scarcella, E. Andersen, & S. D. Krashen (Eds.), *Developing communicative competence in a second language* (pp. 121–138). New York, NY: Newbury House.

Ockey, G. J. (2007). Construct implications of including still image or video in computer-based listening tests. *Language Testing*, *24*(4), 517–537. doi:10.1177/0265532207080771

Pennycook, A. (1985). Actions speak louder than words: Paralanguage, communication, and education. *TESOL Quarterly*, *19*(2), 259–282.

Progosh, D. (1996). Using video for listening assessment: Opinions of test-takers. *TESL Canada Journal*, *14*(1), 34–44.

Rubin, D. L. (1992). Nonlanguage factors affecting undergraduates' judgments of nonnative English-speaking teaching assistants. *Research in Higher Education*, *33*(4), 511–531.

Rubin, D. L., & Smith, K. A. (1990). Effects of accent, ethnicity, and lecture topic on undergraduates' perceptions of nonnative English-speaking teaching assistants. *International Journal of Intercultural Relations*, *14*, 337–353.

Stam, G. (2007). Second language acquisition from a McNeillian perspective. In S. D. Duncan, J. Cassell, & E. T. Levy (Eds.), *Gesture and the dynamic dimension of language: Essays in honor of David McNeill* (pp. 117–124). Amsterdam, The Netherlands: John Benjamins.

Sueyoshi, A., & Hardison, D. M. (2005). The role of gestures and facial cues in second language listening comprehension. *Language Learning*, *55*(4), 661–699. doi:10.1111/j.0023-8333.2005.00320.x

Suvorov, R. (2009). Context visuals in L2 listening tests: The effects of photographs and video vs. audio-only format. In C. A. Chapelle, H. G. Jun, & I. Katz (Eds.), *Developing and evaluating language learning materials* (pp. 53–68). Ames, IA: Iowa State University.

Tian, J. (2011). Oral language performance of English-speaking learners of Mandarin in web-based audio and video conferencing. (Doctoral dissertation). Available from ProQuest Dissertations and Theses database. (UMI No. 3454491).

Von Raffler-Engel, W. (1980). Kinesics and paralinguistics: A neglected factor in second-language research and teaching. *Canadian Modern Language Review*, *36*(2), 225–237.

Wagner, E. (2010). The effect of the use of video texts on ESL listening test-taker performance. *Language Testing*, *27*(4), 493–513. doi:10.1177/0265532209355668

Winke, P. (2010). Using online tasks for formative language assessment. In A. Shehadeh & C. Coombe (Eds.), *Applications of task-based learning in TESOL* (pp. 173–185). Alexandria, VA: TESOL.