Using Automatic Speech Recognition Technology with Elicited Oral Response Testing

TROY L. COX

RANDALL S. DAVIES

Brigham Young University

ABSTRACT

This study examined the use of automatic speech recognition (ASR) scored elicited oral response (EOR) tests to assess the speaking ability of English language learners. It also examined the relationship between ASR-scored EOR and other language proficiency measures and the ability of the ASR to rate speakers without bias to gender or native language. To that end, 179 subjects were given an ASR-scored EOR test with 60 items, followed by an oral proficiency interview (OPI) type assessment and a battery of other language tests. Findings suggest that ASR-scored EOR results could be used alone to predict speaking ability in specific situations and for limited purposes such as initial placement of students in language training situations. However, if more certainty is required, adding a listening component would improve the assessment. Analysis of the study results also suggests that while there were some differences in amount of variance explained in speaking scores based on gender and native language, there was no significant negative effect that would preclude the use of ASR-scoring. While EOR is not an authentic performance assessment of the speaking ability, it does correlate well with other assessments of this construct and has good content validity. The use of an ASR-scored EOR test seems to provide a practical estimate of speaking proficiency that could be used for initial placement of students in situations where assessments of speaking for the purpose of placement are not currently being used due to the cost of administering OPI type assessments.

KEYWORDS

Speaking Assessment, ASR, Elicited Imitation, Sentence Repetition

INTRODUCTION

In an increasingly global society, speaking proficiently in another language is one of the most important skills needed to interact effectively with individuals from countries where that language is spoken. As a result, universities around the world provide students with training in foreign languages. Because students typically start at different proficiency levels, it is desirable that trained personnel administer assessments designed to help place students appropriately. Unfortunately, assessing speaking can be challenging due to the time required to train raters and the cost of administering speaking exams (Coombe, Folse, & Hubley, 2007). This means that language students are commonly placed into cohorts without having their speaking skills evaluated adequately. Failure to assess students properly can result in skill misalignment, in which students with strong reading and writing skills yet weak oral skills are placed in a class in which they are unable to follow spoken directions effectively or participate in classroom discussions (Hudson & Clark, 2008).

The most common method for assessing speaking proficiency has been one-on-one face-to-face interviews (Luoma, 2004). To mitigate the tendency of single raters to follow their own

idiosyncratic pattern of scoring speech samples, best practice dictates that two raters should be involved (Fulcher, 2003). To decrease the time and labor needed to obtain and assess ratable speech samples, computers can be used to collect speech samples (Chapelle & Douglas, 2006); however, even with these advances in more efficient data collection, scoring can still be a time-consuming process (Brown, 2004).

A less expensive alternative to employing human raters is to use automatic speech recognition (ASR) technology. One limitation of ASR is that this technology still cannot reliably recognize spontaneous, natural speech from different speakers (O'Shaughnessy, 2008). However, reliability in ASR processing increases dramatically when limited to a single speaker or a narrow language domain. For example, some commercially available speech recognition programs have users read phonologically rich paragraphs to train the ASR to the individual user's voice. Improving the ASR by training it to a nonnative speaker's second language with their language learning idiosyncrasies would be inappropriate for language testing situations. However, narrowly defining the language to be recognized can also improve ASR's ability to process speech. Many cell phones do this when using ASR technology; they limit the language to digits in a telephone number or specific names in an internal address book. One way of delineating the language when assessing this construct is to use a process called elicited oral response (EOR) testing.

EOR has examinees listen to specific phrases in a foreign language, of varying sentence lengths, and then repeat what they hear. When the utterances are sufficiently long, the examinee is required to process the language, including grammar, vocabulary, and other linguistic features, to understand the meaning and then reconstruct the sentence to repeat it. The rationale is that examinees cannot process language that is beyond their proficiency level (Vinther, 2002). EOR test item difficulty can be varied by modifying those factors needed for comprehension; for example, the number of syllables in the sentence or the grammatical and lexical complexity (Bley-Vroman & Chaudron, 1994). Using ASR with EOR testing may be useful since programming the ASR technology with the specific words in the sentences to be recognized would enable it to score the examinee utterances more accurately.

BACKGROUND INFORMATION AND COMPARISONS OF SPEAKING ASSESSMENTS

Before evaluating the relative merits of any type of assessment, it is beneficial to have a set of criteria as the basis for the judgment. Bachman and Palmer's (1997) test usefulness model contends that the usefulness of a test is an interaction of its *reliability, construct validity, authenticity, interactiveness, impact*, and *practicality*. With that framework in mind, it is instructional to review one of the most commonly used and highly regarded assessments of speaking ability: the American Council of the Teaching Of Foreign Languages (ACTFL) Oral Proficiency Interview or OPI (Fulcher, 2003) and later compare it to ASR-scored EOR tests.

Oral Proficiency Interview Testing

The OPI is a structured interview between an examinee and a certified tester that lasts between 15 and 30 minutes. When conducting the interview, the tester has to adapt the topic and questions so they switch between establishing a proficiency level baseline and challenging the examinee to determine the upper level of their abilities. For quality purposes the interview is recorded and subsequently double-rated. If there is a discrepancy between the two ratings, additional certified testers resolve the dispute. To become a certified tester, an individual must attend a week-long training workshop, submit a practice round of interviews with different levels of language proficiency, receive feedback on that practice round, and then conduct interviews with different individuals for the final submission (OPI Tester Certification, 2012).

Applying the test usefulness framework, we can see that *reliability* is improved by the extensive training and multiple ratings of certified testers (Buck, Byrnes, & Thompson, 1989), though an inherent weakness is that there are fewer independent samples of speech to rate. With regards to the validity of the OPI, since the exam is an oral interview it could be argued that the score reflects the construct of speaking, and thus would be considered to have *construct validity*. However, if specific structures or types of speech need to be assessed, it can be difficult for the interviewer to elicit those forms and it is easy for an examinee to avoid them, thus the test could lack *content validity* if the interviewer is not careful. Since the nature of the interview is conversational, the assessment would be considered *authentic* and certainly the test allows high *interactivity* as the examinee moves between topics and different conversational strategies. One positive *impact* effect is that to prepare for this type of test, examinees would need to practice engaging in interviews. Unfortunately, this type of speaking assessment is only practical for institutions when they have the required resources. The *practicality* of institutionally testing large numbers of students with certified raters can be cost prohibitive. It is expensive to train interviewers, and the one-on-one nature of an interview procedure introduces time constraints that make this kind of testing difficult to use *en masse*. EOR testing with ASR scoring could improve the practicality of this type of assessment and allow it to be used more widely.

Elicited Oral Response Testing

In simple terms, elicited oral response (EOR) requires examinees to listen to a sentence and then repeat what they hear, but this definition does not do justice to the theory supporting the technique. The fundamental theory behind EOR is based on a well-established psycholinguistic research technique often referred to as *elicited imitation* (Berry, 1976; Erlam, 2006; Gallimore & Tharp, 1981; Hamayan, Saegart, & Larudee, 1977; Markman, Spilka, & Tucker, 1975; Naiman, 1974; Slobin & Welsh, 1968; Tomita, Suzuki, & Jessop, 2009; Vinther, 2002). We have chosen to use the term EOR to differentiate the use of this technique as a testing method rather than a research protocol. We make this distinction to emphasize the fact that more than mere rote imitation and repetition are occurring. Furthermore, we contend that some of the concerns about the use of elicited imitation as a research protocol are of less importance when it is viewed as a testing procedure.

Use of EOR as a speaking assessment is based on two concepts. First, second language learners have a transitional, variable, and systematic *interlanguage* that is implicit (Selinker, 1972). The structures of this interlanguage are influenced by many factors, including the person's native language, as well as some universal stages of grammar acquisition that all language learners pass through, regardless of their native language (Ellis & Barkhuizen, 2005).

The second fundamental concept of EOR relevant to its use in testing situations is that short-term or working memory has limits. Miller (1956) posited that the capacity of working memory is seven plus or minus two pieces of information; however more recent research indicates that the amount of information an individual can process and immediately recall might be closer to four (Cowan, 2001). The amount of information that can be stored in working memory is directly related to the ability of the examinees to access long-term memory and the capacity of their interlanguage skills to deconstruct the content into meaningful chunks (i.e., usable units of information). An examinee that has listened to the utterance to be repeated must make sense of the phrase then reconstruct the sentence. The degree to which the examinee can reproduce the sentence depends on an interaction of working memory and long-term memory. Thus the ability to repeat longer sentences depends on the examinees' skill with and knowledge of the language, not just an individual's ability to parrot what is heard (Okura & Lonsdale, 2012).

Nonnative language speakers' working memory capacity in their second language is affected by their proficiency in that language, since novice language learners can hold fewer items in working memory than advanced learners (Scott, 1994). As second language learners' proficiency in the new language advances, becoming more native-like, their working memory capacity advances as well. The more proficient second language learners become, the more likely they are to be able to chunk the language into meaningful units, thus improving their ability to repeat phrases (van den Noort, Bosch, & Hugdahl, 2006). In repeating the EOR utterance, examinees would need to deconstruct what they heard by accessing long-term memory and processing the sentence into meaningful chunks of information. They then have to reconstruct the chunks in order to reproduce the sentence. The more proficient nonnative speakers are with the new language, the more accurate they should be at repeating a phrase they hear in that language.

Automatic Speech Recognition

Automatic speech recognition (ASR) is the process of transferring spoken words to text. To accomplish this, the sound waves of speech are processed and the patterns are analyzed and are first matched with the sounds of the language via the acoustic model and are later matched with patterns of known words via the language model. The functionality of ASR software is not trivial, as a number of factors affect its ability to process speech (Benzeghiba et al., 2007). The ASR software first differentiates between sounds produced by the human vocal chords and all other possible sounds. Once the ASR identifies the human voice, a number of factors affect the acoustic signal the ASR processes, including vocal characteristics that vary systematically between groups of speakers, as well as individual variations within groups of speakers. After recognizing sounds, it parses those sounds into words and sentences.

Once the ASR software identifies a human voice it must consider factors involving the vocal features that vary systematically based on speaker characteristics such as gender and native language. With gender, the length of the vocal tract of men tends to be longer than that of women, resulting in men's voices having a lower pitch (Pickett & Morris, 2000). Concerning native language, the *voice quality setting* refers to the long-term postures of the vocal tract that are language specific (Derwing, 2008). For example, native English speakers tend to keep their lips spread far apart with a more open jaw and the tongue more in the palate. In contrast, French speakers keep their lips more closed and rounded with a fronted tongue (Esling & Wong, 1983). These voice quality settings affect the sound patterns that are produced, and they are often transferred to the second language being learned. Thus the French accent that is detected from French speakers learning English is based to some degree on the voice quality settings of French. The accuracy of the ASR software to recognize speech may be impacted by these systematic variations.

In addition to recognizing the vocal characteristics on a group level, an ASR must have the capability of processing variations within any group, including the unique physical variations in the length and shape of the pharynx, larynx, oral cavity, and articulators that can affect pitch, tone quality, and timbre of any individual speaker's voice. Even with individuals whose vocal tracts are physiologically similar, speech mannerisms such as speed, expressiveness, and volume may impact the acoustic signal of any given speaker (O'Shaughnessy, 2008).

In addition to making discriminating decisions involving voice characteristic factors, ASR software must be able to identify words in context. ASR software does this using a natural language processor. This procedure is complicated as the ASR software moves from processing individual sounds to longer utterances. First, the ASR software must determine when one word ends and another begins (e.g., Does the sound /aiskrim/ refer to "I scream" or the compound word "ice cream?"). The ASR needs to take word boundaries into account.

Beyond that, though, the ASR software needs to recognize enough context to know which word a homonym refers to (e.g., Does the sound /nait/ refers to *night* or *knight?)*. These examples illustrate the difficulty in achieving error-free recognition (Chiu, Liou, & Yeh, 2007). In order for ASR to function well, constraints need to be made by limiting the input to either specific speakers or specific words and contexts (Wachowicz & Scott, 1999).

Potential of ASR-scoring and EOR Tests

In applying Bachman and Palmer's (1997) test usefulness framework to evaluate the potential for using EOR when testing speaking ability even without the use of ASR, there are different strengths and weaknesses compared to the OPI method of testing. First, reliability can be established as it is possible to consistently administer independent items to all examinees (Coombe et al., 2007). Using EOR in a speaking assessment may improve test-retest reliability because it can target and elicit specific grammar and vocabulary in multiple instances that examinees might not utter spontaneously (Henning, 1983). Using EOR by itself would not eliminate the need for raters to score the recorded responses, but as the responses are narrowly defined, raters would not require as much training to be able to score whether the utterance is correct or incorrect.

In terms of validity, using EOR may have some benefits. Given that the EOR can be written to prompt examinees to say several specific phrases in a short time frame, test developers can increase content validity by intentionally including a wide range of topics, vocabulary and structures to be sampled. However, since EOR is an indirect test of speaking, it would have lower construct validity for testing conversational skills, as successfully repeating a sentence in a controlled environment might not indicate that the structure would be reproduced in natural speech (Erlam, 2006). A test using EOR would not reveal whether the individuals know when to use a specific grammatical structure, only whether they are capable of doing so.

Considering other characteristics of test usefulness, the authenticity of this test type is low, as speakers are rarely required to repeat verbatim what they hear. The interactivity of this test type is limited, as the students would need to use their background knowledge to understand and reconstruct vocabulary and structures, but they would not be using higher order thinking skills in their second language. A potential negative impact (i.e., washback) that might occur is that students might practice listening and repeating sentences to prepare for a test rather than engaging in conversations.

The greatest benefit of using EOR testing is practicality. EOR is relatively inexpensive to administer and rate (Matsushita, Lonsdale, & Dewey, 2010). If the purpose of the assessment is low stakes, EOR could be a viable, reliable and practical way to get a basic assessment of speaking ability where this skill might not otherwise be assessed. Since the language of EOR is narrowly defined, it could be even more practical if the rating could occur using ASR to score the assessment.

Using ASR technology with EOR testing would likely improve both the practicality and reliability of the EOR testing procedure over its use alone. First, using ASR eliminates the need for human raters. Second, since the nature of EOR is to repeat specific sentences, the language model used in a specific ASR application can be restricted to a simple dictionary that meets ASRs criterion of using narrowly defined language sets. Furthermore, the EOR items can be structured so that individual words in the sentence are phonologically distinct enough that the ASR should be able process them better. The ASR can then be programmed to recognize the words in each sentence as separate items and rate how many words were uttered correctly.

Many researchers have explored the technological possibility of using ASR to score speaking ability. Eskenazi (1999) discussed the use of the Carnegie Mellon's ASR FLUENCY system to provide pronunciation training for foreign language students. Rypa and Price (1999) described a prototype of the Voice Interactive Training System (VILTS) that used ASR to help students improve oral communication. Cucchiarini, Neri, and Strik (2009) found the use of ASR in giving Dutch students feedback on their pronunciation to be beneficial. They found that while the system did not achieve 100% accuracy in detecting errors, the students enjoyed using it and their pronunciation improved. Zechner, Higgins, Xi, and Williamson (2009) reported on the use of the program SpeechRater to rate the speech samples of the Test of English as a Foreign Language (TOEFL) Practice Online (TPO). The TPO samples consisted of open-ended topics no more than 45 seconds in length. They were able to find moderate correlations that concluded that ASR could be used in a low stakes practice environment. Bernstein, Van Moere, and Cheng (2010) examined the validity of using automated speaking tests in the assessment of Spanish, Dutch, Arabic, and English. They found that a combination of item types including reading sentences aloud, sentence repetition, saying opposite words, oral short answer responses, and retelling spoken passages were strongly correlated with the scores received during oral interviews. While studies such as these have been conducted, there is still a call for additional research that more fully explores the potential of ASR and natural language processing (Chapelle & Chung, 2010; Xi, 2010).

Some researchers have been specifically looking at the combination of elicited imitation and ASR. Graham, Lonsdale, Kennington, Johnson, and McGhee (2008) detailed the development of an ASR-scored elicited imitation engine for English language learners. They were able to achieve a correlation of .66 of human-scored elicited imitation and OPIs with a subset of participants (n=40). In refining the settings on the ASR engine, they were able to achieve a correlation of .90 between human and ASR scoring. Other researchers have looked at the potential use of ASR-scored elicited imitation in other languages, including French (Millard & Lonsdale, 2011), Spanish (Graham, McGhee, Sanchez-Tenney, & LeGare, 2011), and Japanese (Matsushita et al., 2010), and have found similar, promising results.

RESEARCH PURPOSE AND QUESTIONS

This study explored the use of ASR-scored EOR as a means of assessing speaking proficiency. The designers of this project built on work from Graham et al. (2008). The purpose for this study was to use an existing data set to determine whether this assessment process could be used to reliably place students studying English as a second language.

The following research questions guided the study:

1. To what degree can ASR-scored EOR tests predict speaking ability?
2. What is the relationship between ASR-scored EOR and other measures of language proficiency?
3. Are there any other tests or combinations of automatically scored tests that more accurately predict speaking ability?
4. Can ASR technology be used to rate EOR tests of speaking ability without bias related to gender or native language?

METHODS

To determine the degree to which ASR-scored EOR testing predicts speaking ability, an ASR-scored EOR test was administered to the students of an intensive English program associated with a large university, in conjunction with a battery of additional placement tests. The purpose of conducting this analysis was to determine whether the ASR-scored EOR results might supplement or even replace speaking proficiency interviews and what, if any, other language assessments could contribute to predicting speaking ability. The study

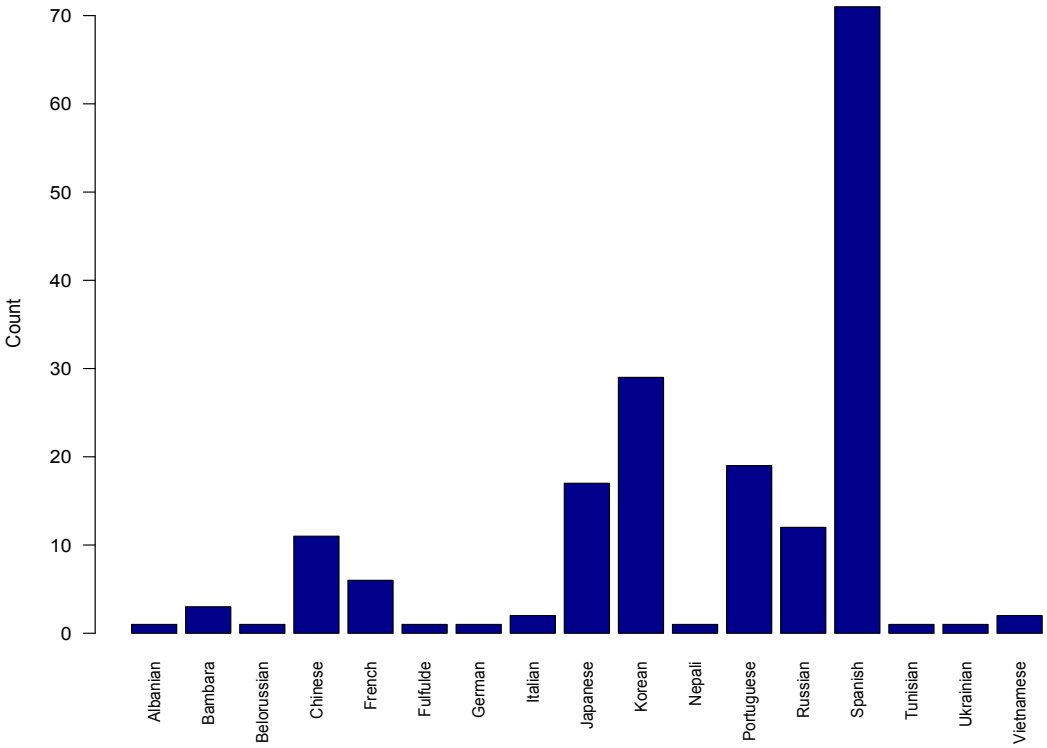
also examined the ASR's ability to rate EOR tests without bias in relation to gender and native language. If the ASR-scored EOR tests had statistically similar results regardless of these factors, then more confidence could be placed in the results as generalizable across populations.

Subjects

The study focused on students enrolled in an intensive English program to study English in preparation for university study. This population was self-selecting in that they had chosen to apply to a language school. To be admitted, they also had to have shown academic aptitude in their previous schooling. There was no minimum language proficiency requirement and the ability of the students ranged from beginner to advanced.

Participants included 179 students from various countries around the world speaking 17 different languages (see Figure 1). This sample included data from 68 males (38%) and 111 females (62%). Participants' ages ranged from 17 to 58, with a mean age of 24.5 and a standard deviation of 6.6 years.

Figure 1
Native Language Frequency of Test Subjects



Data Collection Instruments

The study consisted of six instruments: an ASR-scored EOR test, a Speaking Proficiency Interview (SpPrI), a Writing Placement Exam (WPE), and a series of ESL Computer Adaptive Placement Exams (ESL-CAPE) which consisted of listening, reading, and grammar (see Table 1). The variable of speaking ability was explored using both the SpPrI and the ASR-scored EOR test.

Table 1
Skill and Operational Variables

Test	Skill	Operational Variable
Listening-CAPE	Listening	Listening CAPE score
Grammar-CAPE	Grammar	Grammar CAPE score
Reading-CAPE	Reading	Reading CAPE score
Writing Placement Exam (WPE)	Writing	Writing level
Speaking Proficiency Interview (SpPrI)	Speaking	Speaking level
ASR-scored EOR	Speaking	ASR-scored EOR result

ASR-scored EOR Test

The administered EOR test consisted of 60 items with sentences ranging from 5 to 23 syllables. The test items had been previously validated by Graham et al. (2008) and a detailed description of the test can be found in their paper. This test was administered on the first day of testing, and required approximately 15 minutes to complete. The students were directed to repeat the sentence exactly as it was heard. Student responses to the EOR test were recorded and batch-scored using Sphinx ASR software, an open source speech recognition toolkit from Carnegie Mellon University (see Lamere et al., 2003). The acoustic model used was the Wall Street Journal corpus and the language model was restricted to a simple dictionary that only included the words uttered in each different EOR sentence. The ASR software rated the repeated sentence by determining if each word repeated correctly or not. The overall score awarded each student was a proportion of the number of correct words divided by the total number of words that were recognized. Each sentence or item had a ratio score between 0 and 1 indicating the number of correct words in the sentence divided by the total number of words. A score of 1, for example, indicated 100% recognition of the repeated phrase; a score of .45 meant that 45% of the words were recognized; and a score of 0 meant that none of the words were repeated correctly.

Speaking Proficiency Interview (SpPrI)

Students' ability to speak English was assessed by the SpPrI, an in-house speaking interview protocol. The SpPrI lasted between 5 to 10 minutes and was conducted by an experienced teacher. Each of the SpPrI interviewers had taught at the institution for more than three years and had gone through calibration training. Teachers conducting the interviews followed standard oral proficiency interview protocols, which included warming up with the student, establishing a baseline of what the student could consistently do, probing to find at what level the student's language broke down, and concluding with a wind-down to put the student at ease (Brown, 2004). Due to budget and time constraints, the interviews were not recorded and were single-rated. The teacher conducted the interview without knowing the scores of the students' other tests to ensure that the interviewer assessed only speaking ability. After the teacher concluded the interview, a level was assigned based on the 7-point scale that corresponded with the program levels (see Table 2).

Writing Placement Exam (WPE)

Writing was assessed by the WPE, an in-house writing placement test consisting of two prompts: a pictorial description and an essay. The five-minute pictorial description presented the students a scene and asked them to describe it, and was targeted at the lower end of the proficiency range. The thirty-minute essay asked the students to respond to a question in an essay-length (multiple paragraph) format and was targeted at the intermediate to high proficiency range. Both writing tasks were double-rated by experienced raters on a 7-point scale that corresponded with the program levels (see Table 2). The

scores of the two raters were averaged. If there was a discrepancy of greater than one level, a third rater was consulted.

Table 2
Program Level and Rubric Scale Scores for Speaking and Writing

Program Level	OPI equivalence	Level Number
Foundations Prep	Novice Low	0
Foundations A	Novice Mid	1
Foundations B	Novice High	2
Foundations C	Intermediate Low	3
Academic A	Intermediate Mid	4
Academic B	Intermediate High	5
Academic C	Advanced low	6

The ESL-CAPE (listening, reading, and grammar)

These tests were part of an in-house developed computer adaptive placement exam battery. The tests were developed in the early 1990s by administering items to a large group of students, calibrating the responses through item response theory (Rasch modeling) and then programming the computer adaptive test. When students take the ESL-CAPE, they receive an ability estimate with a standard error for each skill being tested. As students answer more items, the ability estimate is refined and the standard error diminishes until it reaches the test's stopping mechanism, which was predetermined to be 0.4. Person ability estimates typically range between -3 and 3, but the ESL-CAPE transforms the scores so the range is between 0 and 1200. As the test was adaptive, the time for each test and the number of items varied depending on the student's ability to consistently answer items of similar difficulty. Furthermore, since the student scores are actually measures derived from item response theory ability estimates, the data can be treated as true interval level data (Bond & Fox, 2001). The students received a score for each of the three skills tested.

Procedure

On the first day of testing, the students took all the computerized tests based on the schedule shown in Table 3. For the CAPE exams (listening, reading, and grammar), the scoring took place as the students completed the tests. The EOR and WPE were rated later in the day. On the second day of testing, speaking was assessed with the SpPrI.

Table 3
Placement Testing Schedule

Day 1	
Computer Adaptive Placement Exams	Listening-CAPE Reading-CAPE Grammar-CAPE
Writing Placement Exam (WPE)	Description of a picture Essay
ASR-scored EOR	EOR - 60 items
Day 2	
Speaking Proficiency Interview (SpPrI)	

Data Analysis

To answer the first question regarding the degree to which ASR-scored EOR could predict speaking ability, a simple linear regression was used. The dependent variable was the SpPrI, and the independent variable was the ASR-scored EOR results. The purpose of this analysis was to determine how well the ASR-scored EOR results alone predicted the results of the speaking proficiency assessment.

To answer the second question that examined the relationship between ASR-scored EOR and other language assessments, a Pearson product-moment correlation was used on all of the tests in the placement battery. The null hypothesis was that there would be no relationship between or among any of the tests. Since the tests measured related but different constructs, researchers expected that the correlation would need to be greater than r=.3 in order to reject null hypotheses (Hatch & Lazaraton, 1991).

To answer the third question and determine the degree to which a combination of other automatically scored assessments could be used to predict the SpPrI, a multiple regression was run on the ASR-scored EOR results and the scores on the Grammar-CAPE, the Listening-CAPE, and the Reading-CAPE. Through analyzing these different measures, the researchers would be able to determine which combination of results accounted for the most variance in predicting speaking ability.

To answer the fourth question and determine if the extraneous factors of gender or native language might cause bias in the ASR ratings, two separate one-way ANOVAs were run to determine if there was a difference in the mean of each of the subgroups. For this analysis, the dependent variable was the ASR-scored EOR test results, and the independent variables were gender and native language respectively. These variables were operationalized as follows. Gender was coded as nominal data into two categories: male and female. For native language, only those languages native to more than 15 examinees were considered, and data were coded nominally into the languages that were spoken. This was done to determine whether the ASR software scores were systematically different based on gender or native language groups. To see how well the groups correlated with the SpPrI, correlations were run disaggregated by each of the subgroups.

Limitations

For this study, a number of limitations should be acknowledged. First, the scale used to measure speaking and writing ability was treated as producing interval data, even though it had not been validated accordingly. Similarly, the scores reported for the EOR test were also treated as interval level data. Parametric statistics have been found to be robust enough to allow violations to some of these assumptions without negating the insight that can be gleaned from their use (Knapp, 1990; Norman, 2010).

Other weaknesses relate to the quality of the data gathered and the generalizability. The oral interviews used to measure speaking ability were only single-rated; thus the reliability of the ratings cannot be verified. Furthermore, the subjects in the study were a convenience sample of students who had the financial means and inclination to study abroad. This also affected the balance of the native languages represented; in the native language subset over half were Spanish speakers. These factors may impact the generalizability of the findings to other populations.

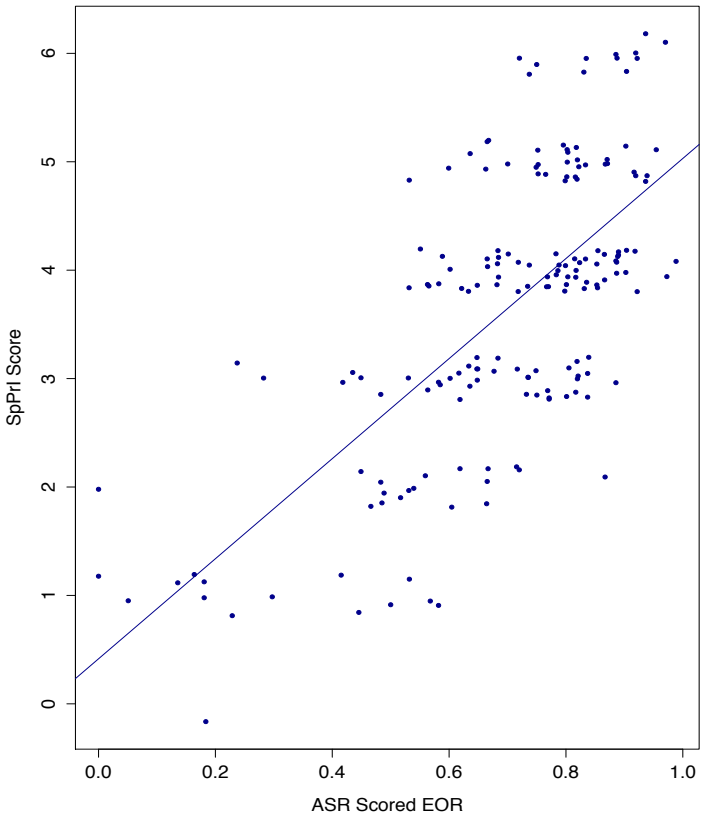
For this particular experiment, one of the 60 items on the EOR assessment was not scored due to an unknown technical issue; thus while researchers had anticipated using 60 items on the ASR-scored EOR assessment, only 59 were used.

RESULTS

Use of ASR-scored EOR Tests to Predict Speaking Ability

The ASR-scored EOR tests had an internal reliability of .94 as measured by a Cronbach Alpha calculation. To answer the first question and test the degree to which the ASR-scored EOR test results could be used to predict speaking level results, a simple regression was run. This regression was found to be significant at the α=.05 level, $F(1, 174)=154.74$, $p<.001$, adjusted $r^2=.47$, indicating that about 47% of the variance in the SpPrI could be explained by the results of the ASR-scored EOR test results (see Figure 2).

Figure 2
Relationship between ASR-Scored EOR Test and Speaking Score Level



Relationship between Speaking Ability and Other Test Scores

To answer the second question and examine the relationship between ASR-scored EOR and other language assessments, Pearson correlations were run between ASR-scored EOR and the Placement Battery Tests. The correlations between all the tests in the placement battery (see Table 4) were found to be significant, with the effect size ranging from moderate to large (see Cohen, 1988). The highest correlation was between the Listening-CAPE and the ASR-scored EOR tests (r=.74). This result seems to suggest that the students with good listening skills tended also to be able to repeat phrases they hear more accurately. This finding also seems to suggest that listening via the medium of a computer has similarities to listening and interacting with a human in face-to-face interviews, as evidenced by the high correlation (r=.74) between listening skills and both the ASR-scored EOR and the SpPrI results. The high correlations between assessments warranted a closer look at which assessments might be used to predict speaking ability as measured by the SpPrI.

Table 4
Correlations between ASR-Scored EOR and the Placement Battery Tests

	ASR-scored EOR	SpPrI	Listening-CAPE	Grammar-CAPE	Reading-CAPE	WPE
ASR-scored EOR	1	.69	.74	.58	.60	.62
SpPrI		1	.74	.57	.65	.67
Listening-CAPE			1	.68	.77	.68
Grammar-CAPE				1	.66	.64
Reading-CAPE					1	.65
WPE						1

All correlations were significant at α=.05 level (two-tail)

The third question concerns determining the degree to which a combination of other automatically scored assessments could be used to predict the SpPrI. A multiple regression using stepwise regression methods found that the only two measures that significantly predicted the SpPrI levels were ASR-scored EOR and the Listening-CAPE score, α=.05 level, $F(2, 175)=124.27$, $p<.001$, with the adjusted $r^2=.59$. About 59% of the variance in the speaking score levels could be explained by the results of the ASR-scored EOR and the Listening-CAPE scores together (See Table 5).

Table 5
Regression Results for Predicting Speaking Ability as Measured by SpPrI

	Unstandardized Coefficients		Standardized Coefficients		
Model	B	St. Error	Beta	*t*	*p*-value
1 (Constant)	.413	.269		1.536	.126
ASR-Scored EOR	4.63	.372	.686	12.44	.000
2 (Constant)	-1.055	.315		-3.347	.001
Listening-CAPE	.005	.001	.512	4.341	.000
ASR-Scored EOR	2.073	.488	.307	4.244	.000

Effect of Gender and Native Language on ASR-Scored EOR Results

To answer the fourth question and determine if ASR technology could be used to rate EOR tests of speaking ability without bias related to gender or native language, ANOVAs and correlations between the SpPrI and the disaggregated subgroups were conducted. This is important to verify as the accuracy of ASR is believed to be affected by the vocal characteristics of the individual speaker; thus significant rating bias could affect the validity of the scoring technique.

Gender bias

Since the vocal differences between males and females could affect the way the ASR scores the EOR, gender bias rating was examined. Using a one-way ANOVA, researchers found no significant difference between males and females in the scores they received from the ASR-scored EOR: α=.05 level, $F(1, 174)=.18$, $p=.68$ (see Table 6). To see how well the ASR-scored EOR correlated with the SpPrI based on gender, separate correlations were run for males and females. The relationship between the ASR-scored EOR and the SpPrI was strongly correlated for both males ($r=.58$) and females ($r=.74$), though the females had a higher correlation (see Table 7). The ASR-scored EOR accounted for 33% of the variance of the SpPrI scores of males, 54% of the variance of the SpPrI scores of females was explained. So while ASR-scored EOR predicted speaking ability adequately for both genders, it did seem to do a slightly better job predicting female voices.

Table 6
The Effect of Gender on ASR-scored EOR Test Results

		Descriptive Statistics		
		N	*Mean*	*SD*
ASR-scored EOR Test	Female	109	.70	.20
	Male	67	.69	.19
	Total	176	.70	.20

Table 7
Correlation of ASR-Scored EOR by Gender to SpPrI

	N	*r*	*p*-value
Female	109	.74	<.001
Male	67	.58	<.001
Combined	176	.69	<.001

Native language

Since differences in voice quality settings based on native language could affect the way the ASR scores the EOR, bias on native language rating was examined. Only languages spoken by more than 15 participants were analyzed. This restricted the analysis to four language groups: Spanish, Korean, Portuguese, and Japanese. While the means of the Korean and Japanese speakers were slightly lower than the means for the Spanish and Portuguese speakers, a one-way ANOVA found no significant difference in scores at the α=.05 level, $F(3, 129)=.82$, $p=.49$. This mirrors the results of the SpPrI in which the Korean and Japanese subjects were lower than the other two groups (see Table 8).

Table 8
The Effect of Native Language on ASR-scored EOR Tests Compared to Speaking Level

		Descriptive Statistics		
		N	*Mean*	*SD*
ASR-scored EOR	Spanish	69	.70	.19
	Korean	29	.65	.21
	Portuguese	18	.72	.19
	Japanese	17	.65	.22
	Total	133	.69	.20
SpPrI	Spanish	69	3.62	1.38
	Korean	29	3.45	1.33
	Portuguese	18	4.22	.81
	Japanese	17	3.00	1.22
	Total	133	3.58	1.32

To see how well each ASR-scored EOR correlated with the SpPrI results based on native language, separate correlations were run for the Spanish, Korean, Portuguese, and Japanese speakers. The relationship between the ASR-scored EOR and the SpPrI was found to be significant at the α=.05 for all languages but the strength of the relationship varied. The correlations ranged from a low of $r=.50$ with the Portuguese speakers to a high of $r=.79$ with the Japanese speakers (see Table 9). The amount of variance in the SpPrI scores

predicted from the ASR-scored EOR ranged from 25% with the Portuguese speakers up to 64% with the Japanese speakers. Based on this result the ASR-scored EOR test predicted speaking ability adequately for each of the languages, however, it did a slightly better job with Spanish and Japanese students.

Table 9
Correlation of ASR-scored EOR by Native Language to SpPrI

	N	*r*	*p*-value
Spanish	69	.76	<.001
Korean	29	.55	<.001
Portuguese	18	.50	.033
Japanese	17	.79	<.001
Total	133	.68	<.001

RESULTS SUMMARY

Based on these results it appears that ASR-scored EOR testing could be used as an alternative to speaking proficiency interviews to measure speaking ability. A simple linear regression showed that ASR-scored EOR results predicted SpPrI speaking scores fairly well. However, Listening-CAPE scores were also strongly related to the speaking scores produced by both the ASR-scored EOR test and the SpPrI speaking scores. This supports the obvious conclusion that listening ability is an important component of both conversational speaking and the ability to process and repeat phrases in a second language. Further investigation suggests that a better prediction of speaking ability might be obtained by considering both the Listening-CAPE assessment and the ASR-scored EOR results.

Analysis of these results also suggests that ASR technology can produce reliable results largely unbiased by gender or native language differences. ASR-scored EOR results were adequate predictors of speaking ability but did predict slightly better for females, as well as Spanish and Japanese students. Still, subgroup correlations between the SpPrI and the ASR-scored EOR results showed a statistically significant moderate to strong relationship for all disaggregated subgroups.

DISCUSSION AND CONCLUSIONS

This study examined how well ASR-scored EOR tests might predict speaking ability as measured by speaking proficiency interviews. It also looked at how well ASR technology could rate EOR tests without bias resulting from gender and native language differences.

The evidence suggests that ASR-scored EOR tests could be used to predict speaking ability, but if more certainty were warranted, adding a listening component might improve the assessment. In terms of ASR's ability to process student results without bias due to gender or native language, analysis of the results suggested that these factors had little effect. We therefore conclude that for purposes of making low-stakes decisions like that of initial student placement, the ASR-scored EOR testing method seems to show great potential as a cost-effective alternative to conducting costly face-to-face speaking proficiency interviews.

The strengths of an ASR-scored EOR test are content validity, reliability, and practicality. EOR allows test designers to sample a wide range of speech structures and can require examinees to respond to items they might otherwise avoid, thus improving content validity. Reliability is increased as multiple samples of the same objective can be tested. Reliability in rating can also be improved, since it is easier to consistently score set items (for both human and ASR ratings). The greatest advantage of ASR–scored EOR tests is practicality. Once the technological infrastructure is in place and calibrated, this rating method is of

limited cost. ASR technology also seems to function well regardless of gender and native language, alleviating the concern over bias related to these matters.

The weaknesses of using ASR-scored EOR testing include issues of construct validity, authenticity, and a potential for a washback effect. Students might practice repeating phrases in order to do better on this type of test rather than practicing conversation skills. Furthermore, confidence in the capability of EOR test results to predict conversational speaking ability rather than a more generic speaking ability may be overly enthusiastic. EOR testing simulates speaking situations but does not directly measure conversation proficiency. Certainly, being able to process and repeat a phrase is not the same as knowing when to use various forms of speech in an authentic situation such as an interview. The disconnect between the EOR task and real world language situations might result in a reduction in construct validity (Vinther, 2002), causing some stakeholders (e.g., students, teachers, administrators) to feel that students are not being tested fairly or accurately. Yet when required resources are greater than available resources, speaking assessments may not occur. When testing for placement, students are often misplaced because no speaking assessment was included. Given the compelling need to ensure speaking is assessed, ASR-scored EOR testing may be a viable alternative. If the limitations of this type of assessment can be mitigated, than the benefits of using ASR-scored EOR testing as a simple yet practical assessment of speaking ability may be worth exploring further.

FUTURE RESEARCH

EOR presents a number of interesting questions when used as a testing technique, and the use of ASR introduces even more questions. As mentioned previously, EOR is based on a psycholinguistic research tool, elicited imitation. While the original research using elicited imitation examined the roles of grammatical complexity and syllable length in item difficulty, little has been done to examine the interaction between listening comprehension and EOR. For example, to what extent does the speech rate of the prompt impact the item difficulty? Are there other acoustic differences such as monotonicity or fundamental frequency that impact the item difficulty?

Scoring for EOR responses also needs to be examined, as automated scoring has great potential. Should the sentences be scored dichotomously as right or wrong? Should partial credit be awarded? If so, should it be at the word or syllable level? By using item response theory (IRT), computer adaptive speaking tests could be developed. A natural extension then is to see what IRT model should be used and how cut scores should be established. Furthermore, ASR ratings might not respond as sensitively as human raters to some of the issues, so further study comparing the scores assigned by ASR and those awarded by human raters must be evaluated. If systematic differences are found between the ASR and human raters, what safeguards need to be in place to ensure the ASR is functioning without bias?

An impact that needs to be examined is the effect of practicing EOR. If EOR is a skill that can be learned independent of speaking proficiency, unintended negative consequences might result from students practicing and increasing their ability to parrot sentences at the expense of more interactive speaking practice. In addition, more complete examination of the relationship between working memory capacity and EOR would be desirable.

It is possible that because the acoustic characteristics of voices change quite dramatically as humans progress from childhood to adulthood, age may influence the ability of ASR to accurately recognize speech. ASR systems may have difficulty processing the speech of children, and they may have difficulty processing the speech of elderly individuals due to tremors that have developed. Before making broad generalizations on the use of ASR-

scored EOR with those populations, it is important to see if the ASR might have bias in rating the speech of those on extreme ends of the age continuum.

REFERENCES

Bachman, L. F., & Palmer, A. S. (1997). *Language testing in practice.* New York, NY: Oxford University Press.

Benzeghiba, M., De Mori, R., Deroo, O., Dupont, S., Erbes, T., Jouvet, D., … Wellekens, C. (2007). Automatic speech recognition and speech variability: A review. *Speech Communication*, *49*(10-11), 763-786.

Bernstein, J., Van Moere, A., & Cheng, J. (2010). Validating automated speaking tests. *Language Testing*, *27*(3), 355-377.

Berry, P. B. (1976). Elicited imitation of language: Some ESNS population characteristics. *Language and Speech*, *19*(4), 350-63.

Bley-Vroman, R., & Chaudron, C. (1994). Elicited imitation as a measure of second-language competence. In M. Tarone, S. Gass, & A. Cohen (Eds.), *Research methodology in second-language acquisition* (pp. 245-261). New York, NY: Routledge

Bond, T. G., & Fox, C. M. (2001). *Applying the Rasch model: Fundamental measurement in the human sciences.* Mahwah, NJ: Lawrence Erlbaum.

Brown, D. H. (2004). *Language assessment: Principles and classroom practices.* White Plains, NY: Pearson/Longman.

Buck, K., Byrnes, H., & Thompson, I. (1989). The ACTFL oral proficiency interview tester training manual. *Yonkers, NY: ACTFL.*

Chapelle, C. A., & Chung, Y. R. (2010). The promise of NLP and speech processing technologies in language assessment. *Language Testing*, *27*(3), 301-315.

Chapelle, C. A., & Douglas, D. (2006). *Assessing language through computer technology.* Cambridge, UK: Cambridge University Press.

Chiu, T. L., Liou, H. C., & Yeh, Y. (2007). A study of web-based oral activities enhanced by automatic speech recognition for EFL college learning. *Computer Assisted Language Learning*, *20*(3), 209-233.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences.* Mahwah, NJ: Lawrence Erlbaum.

Coombe, C. A., Folse, K. S., & Hubley, N. J. (2007). *A practical guide to assessing English language learners.* Ann Arbor, MI: University of Michigan.

Cowan, N. (2001). The magical number 4 in short-term memory: A reconsideration of mental storage capacity. *The Behavioral and Brain Sciences*, *24*(1), 87-114; discussion 114-85.

Cucchiarini, C., Neri, A., & Strik, H. (2009). Oral proficiency training in Dutch L2: The contribution of ASR-based corrective feedback. *Speech Communication*, *51*(10), 853-863.

Derwing, T. (2008). *Curriculum issues in teaching pronunciation. Phonology and second language acquisition*. Amsterdam, The Netherlands: John Benjamins Publishing Company.

Ellis, R., & Barkhuizen, G. (2005). *Analysing learner language.* New York, NY: Oxford University Press.

Erlam, R. (2006). Elicited imitation as a measure of L2 implicit knowledge: An empirical validation study. *Applied Linguistics*, *27*(3), 464-491.

Eskenazi, M. (1999). Using automatic speech processing for foreign language pronunciation tutoring: Some issues and a prototype. *Language Learning & Technology*, *2*(2), 62-76.

Flynn, S. (1986). Production vs. comprehension: Differences in underlying competencies. *Studies in Second Language Acquisition*, *8*(2), 135-164.

Fulcher, G. (2003). *Testing second language speaking.* New York, NY: Pearson Education.

Gallimore, R., & Tharp, R. G. (1981). The interpretation of elicited sentence imitation in a standardized context. *Language Learning*, *31*(2), 369-392.

Graham, C. R., Lonsdale, D., Kennington, C., Johnson, A., & McGhee, J. (2008). Elicited imitation as an oral proficiency measure with ASR scoring. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC 2008)* (pp. 1604-1610). European Language Resources Association.

Graham, C. R., McGhee, J., Sanchez-Tenney, M., & LeGare, M. (2011 June). Examining the validity of an elicited imitation instrument to test oral language in Spanish. Presentation at 33rd Annual Convention of the Language Testing Research Colloquium (LTRC 2011). Ann Arbor, MI, USA.

Hamayan, E., Saegert, J., & Larudee, P. (1977). Elicited imitation in second language learners. *Language and Speech*, *20*(1), 86-97.

Hatch, E., & Lazaraton, A. (1991). *The research manual: Design and statistics for applied linguistics.* New York, NY: Heinle & Heinle.

Henning, G. (1983). Oral proficiency testing: Comparative validities of interview, imitation, and completion methods. *Language Learning*, *33*(3), 315-332.

Hudson, T., & Clark, M. (2008). Designing sorting hats: Foreign language placement processes. *Case Studies in Foreign Language Placement: Practices and Possibilities*, *1*, 1-6.

Knapp, T. R. (1990). Treating ordinal scales as interval scales: An attempt to resolve the controversy. *Nursing Research*, *39*(2), 121-3.

Lamere, P., Kwok, P., Gouvêa, E., Raj, B., Singh, R., Walker, W., Warmuth, M., & Wolf, P. (2003). The CMU SPHINX-4 speech recognition system. In *IEEE international conference on acoustics, speech and signal processing.* Hong Kong.

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

Markman, B. R., Spilka, I. V., & Tucker, G. R. (1975). The use of elicited imitation in search of an interim French grammar. *Language Learning*, *25*(1), 31-41.

Matsushita, H., Lonsdale, D., & Dewey, D. (2010). Japanese elicited imitation: ASR-based oral proficiency test and optimal item creation. In G. R. S. Weir & S. Ishikawa (Eds.), *Corpus, ICT, and language education* (pp. 161-172). University of Strathclyde Publishing.

Millard, B., & Lonsdale, D. (2011). *French oral proficiency assessment: Elicited imitation with speech recognition; Selected proceedings from LSRL 2011*. Manuscript submitted for publication.

Miller, G. A. (1956). The magical number seven plus or minus two: Some limits on our capacity for processing information. *Psychological Review*, *63*(2), 81-97.

Naiman, N. (1974). The use of elicited imitation in second language acquisition research. *Working Papers on Bilingualism*, *2*, 1-37, 0319-5171.

Norman, G. (2010). Likert scales, levels of measurement and the "laws" of statistics. *Advances in Health Sciences Education: Theory and Practice*, *15*(5), 625-32.

O'Shaughnessy, D. (2008). Automatic speech recognition: History, methods and challenges [invited paper]. *Pattern Recognition*, *41*(10), 2965-2979.

Okura, E., & Lonsdale, D. (2012). Working memory's meager involvement in sentence repetition tests. In N. Miyake, D. Peebles, & R. P. Cooper (Eds.), *Proceedings of the 34th Annual Conference of the Cognitive Science Society* (pp. 2132-2137). Austin, TX: Cognitive Science Society.

OPI Tester Certification. (2012). Retreived March 14, 2012, from http://www.actfl.org/i4a/pages/index.cfm?pageid=3350

Pickett, J. M., & Morris, S. R. (2000). The acoustics of speech communication: Fundamentals, speech perception theory, and technology. *The Journal of the Acoustical Society of America*, *108*, 1373.

Rypa, M. E., & Price, P. (1999). VILTS: A tale of two technologies. *Calico Journal*, *16*(3), 385-404.

Scott, M. L. (1994). Auditory memory and perception in younger and older adult second language learners. *Studies in Second Language Acquisition*, *16*(3), 263-281.

Selinker, L. (1972). Interlanguage. *International Review of Applied Linguistics in Language Teaching, 10*(1-4), 209-232.

Slobin, D., & Welsh, C. (1968). Elicited imitation as a research tool in developmental psycholinguistics. Working Paper 10. Reprinted in C.A. Ferguson & D.I. Slobin (1973). *Studies of child language development* (pp. 485-497). New York: Holt, Rinehart, and Winston Inc.

Tomita, Y., Suzuki, W., & Jessop, L. (2009). Elicited imitation: Toward valid procedures to measure implicit second language grammatical knowledge. *TESOL Quarterly*, *43*(2), 345-350.

van den Noort, M. W. M. L., Bosch, P., & Hugdahl, K. (2006). Foreign language proficiency and working memory capacity. *European Psychologist*, *11*(4), 289-296.

Vinther, T. (2002). Elicited imitation: A brief overview. *International Journal of Applied Linguistics*, *12*(1), 54-73.

Wachowicz, K. A., & Scott, B. (1999). Software that listens: It's not a question of whether, it's a question of how. *CALICO Journal*, *16*, 253-276.

Xi, X. (2010). Automated scoring and feedback systems: Where are we and where are we heading? *Language Testing*, *27*(3), 291-300.

Zechner, K., Higgins, D., Xi, X., & Williamson, D. M. (2009). Automatic scoring of non-native spontaneous speech in tests of spoken English. *Speech Communication*, *51*(10), 883-895.

AUTHOR'S BIODATA

Troy L. Cox has been involved with language assessment and technology in teaching for over 16 years. He is a certified ACTFL oral proficiency tester and has used his testing expertise as a forensic linguist and in test development projects for various organizations. He is currently an assessment consultant for Brigham Young University's (BYU) Center for Language Studies and is the associate coordinator of technology and assessment at BYU's English Language Center.

Randall S. Davies is an Assistant Professor in the Department of Instructional Psychology and Technology at Brigham Young University.

AUTHOR'S ADDRESS

E-mail: troy_cox@byu.edu