Brigham Young University
BYU ScholarsArchive

All Theses and Dissertations

2013-03-14

Investigating Prompt Difficulty in an Automatically Scored Speaking Performance Assessment

Troy L. Cox
Brigham Young University - Provo

Follow this and additional works at: https://scholarsarchive.byu.edu/etd

 Part of the Educational Psychology Commons

BYU ScholarsArchive Citation

Cox, Troy L., "Investigating Prompt Difficulty in an Automatically Scored Speaking Performance Assessment" (2013). *All Theses and Dissertations*. 3929.
https://scholarsarchive.byu.edu/etd/3929

This Dissertation is brought to you for free and open access by BYU ScholarsArchive. It has been accepted for inclusion in All Theses and Dissertations by an authorized administrator of BYU ScholarsArchive. For more information, please contact scholarsarchive@byu.edu, ellen_amatangelo@byu.edu.

Investigating Prompt Difficulty in an Automatically Scored Speaking Performance Assessment

Troy L. Cox

A dissertation submitted to the faculty of
Brigham Young University
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

Randall Spencer Davies, Chair
Dan P. Dewey
Richard R. Sudweeks
Ray Thomas Clifford
William G. Eggington

Department of Instructional Psychology and Technology

Brigham Young University

March 2013

Copyright © 2013 Troy Cox

All Rights Reserved

ABSTRACT

Investigating Prompt Difficulty in an Automatically Scored
Speaking Performance Assessment

Troy L. Cox
Department of Instructional Psychology and Technology
Doctor of Philosophy

Speaking assessments for second language learners have traditionally been expensive to administer because of the cost of rating the speech samples. To reduce the cost, many researchers are investigating the potential of using automatic speech recognition (ASR) as a means to score examinee responses to open-ended prompts. This study examined the potential of using ASR timing fluency features to predict speech ratings and the effect of prompt difficulty in that process. A speaking test with ten prompts representing five different intended difficulty levels was administered to 201 subjects. The speech samples obtained were then (a) rated by human raters holistically, (b) rated by human raters analytically at the item level, and (c) scored automatically using PRAAT to calculate ten different ASR timing fluency features. The ratings and scores of the speech samples were analyzed with Rasch measurement to evaluate the functionality of the scales and the separation reliability of the examinees, raters, and items.

There were three ASR timed fluency features that best predicted human speaking ratings: speech rate, mean syllables per run, and number of silent pauses. However, only 31% of the score variance was predicted by these features. The significance in this finding is that those fluency features alone likely provide insufficient information to predict human rated speaking ability accurately. Furthermore, neither the item difficulties calculated by the ASR nor those rated analytically by the human raters aligned with the intended item difficulty levels. The misalignment of the human raters with the intended difficulties led to a further analysis that found that it was problematic for raters to use a holistic scale at the item level. However, modifying the holistic scale to a scale that examined if the response to the prompt was at-level resulted in a significant correlation ($r = .98$, $p < .01$) between the item difficulties calculated analytically by the human raters and the intended difficulties. This result supports the hypothesis that item prompts are important when it comes to obtaining quality speech samples. As test developers seek to use ASR to score speaking assessments, caution is warranted to ensure that score differences are due to examinee ability and not the prompt composition of the test.

Keywords: Automatic Speech Recognition, second language oral proficiency, language testing and assessment, English as a second language tests, speech signal processing

ACKNOWLEDGMENTS

It would not have been possible to write this doctoral dissertation without the help and support of the kind people around me, to only some of whom it is possible to give particular mention here. For those whom I did not mention, I ask for your forgiveness in advance.

I would like to thank my advisor, Dr. Randall Davies, for his guidance and support throughout the production of this research and dissertation. I would also like to thank Dr. Richard Sudweeks, Dr. Ray Clifford, Dr. Dan Dewey and Dr. William Eggington for their helpful suggests and for serving on my dissertation committee.

I would also like to thank my colleagues both past and present at the English Language Center. In particular, I want to thank Robb McCollum and Ben McMurry for encouraging me to apply to the IP&T program, and for Neil Anderson, Norman Evans, James Hartshorn, Judson Hart and many others for supporting me as I pursued this degree. I would be remiss if I did not thank the students and faculty of the English Language Center for participating in the research and making this study possible.

I would like to thank my parents, Leigh and Joyce Cox, my extended family, and all my friends for their unwavering support and encouragement throughout the years. Finally, I would like to thank my family, my beautiful wife, Heidi, and my children, Cameron, Hannah and Camille for reminding me about what's really important, keeping me sane and making me laugh.

TABLE OF CONTENTS

LIST OF TABLES

TABLE OF FIGURES

Chapter 1

Introduction

As technology and travel make the world smaller, the need to assess speaking ability becomes increasingly important. Schools that teach foreign languages need some type of speaking assessment for placement, for exit exams, and for certifying the language proficiency of their graduates. Businesses want to certify that their employees have the language ability to participate in a global economy. Governments need a way to ensure their civil servants have the speaking ability needed to meet the needs of a linguistically diverse citizenry as well as the ability to adequately communicate with other governments. Language testers are under pressure to create tests that accomplish these functions faster, better and cheaper (Chun, 2010).

While there is a need to measure speaking ability, traditional methods of assessing speaking through oral interviews such as an OPI (Oral Proficiency Interview) are often impractical for a number of reasons (Luoma, 2004). First, obtaining a ratable speech sample is not a trivial matter (Buck, Byrnes, & Thompson, 1989). It requires a skilled and trained interviewer to prompt the type of speech needed to differentiate between levels. After the speech sample has been obtained, it must then be scored. To increase reliability in high stakes testing, the interview is recorded so a second trained rater can score it (Fulcher, 2003). Yet this increase in reliability can be cost-prohibitive and often greatly increases the turnaround time required to grade the performance and determine a score.

One way to reduce costs is to use technology in the assessment. For example, the SPEAK test from Educational Testing Service, the SOPI (Simulated Oral Proficiency Interview) and the COPI (Computer Oral Proficiency Interview) both from the Center for Applied Linguistics, as well as the OPIc (Oral Proficiency Interview-computerized) from Language

Testing International all use technology to administer their assessments without trained interviewers present (Chapelle & Douglas, 2006). While this has reduced the personnel costs in administering the tests, the costs associated with the human scoring of the speech samples is still an issue.

Automatic Speech Recognition in Language Testing

One way that technology could reduce the cost associated with rating speaking abilities would be the implementation of Automatic Speech Recognition (ASR). ASR is based on pattern recognition and has been available for years in dictation software (O'Shaughnessy, 2008). ASR has been found to be a promising tool to facilitate the scoring of speaking tests, making scoring faster and cheaper, but is it better? If not better, at a minimum, can it retain the same quality level of human rating? ASR is clearly limited in its ability to recognize both speaker-independent and context-independent speech samples (O'Shaughnessy, 2008), but it can recognize various timing features of speech believed to be associated with fluency. Since an individual's ability to speak fluently (i.e., their rate of speak, number and length of pauses, etc.) is related to speaking ability, ASR-scored assessments of speaking fluency may provide a reasonable indicator of speaking ability.

The validity of the assessment results is based on the assumption that the speaking samples being rated adequately represent the examinees ability (i.e., the prompts used to obtain the speech sample were designed to elicit a response in which the full range of the examinees ability are evident). Many speaking assessments contain a number of *prompts* (also called items or task) that elicit the speech samples to be rated. What has yet to be examined is how the prompts used in an assessment affect the quality of the sample and thus the timing feature scores obtained through ASR. In high stake's testing situations (e.g. tests used to make important

decisions with real world implications), multiple test forms must be created and those forms need to be of equal difficulty to be fair to the examinees. If there is misalignment between the human rating and ASR item difficulty of the prompts, the test design could negatively impact the examinees. Item prompts on an assessment are typically designed to elicit specific levels of responses. For example, a test of overall proficiency would have prompts at various levels of difficulty. Consider the following prompts that might be asked on a test: (a) Describe your family, and (b) Compare and contrast the role of the nuclear and extended family in your country and in the US. In this example, most language educators would predict that the ability to successfully describe one's family is a prerequisite skill to the ability of comparing and contrasting family structures; thus the second prompt should be more difficult than the first. For the purposes of assessing speaking ability, the second prompt is expected to provide a situation where the respondent could demonstrate evidence of a higher level speaking ability whereas the first prompt would not elicit the full range of the examinee's ability to speak at a higher level. A natural extension of this assumption is that the ASR timing features for an examinee's response to the first prompt would be more fluent (i.e., faster rate of speech, fewer pauses, etc.) than that of the second prompt.

In analyzing the ASR timing features of fluency for both prompts, it would be expected that the ASR scores would differ. If the features are the same, then either the individual is extremely fluent or the timing fluency features are invariant and any prompt could be used. In this case, regardless of the item prompt used, the ASR timing features alone would be sufficient to determine the examinee's speaking ability. If, however, the ASR features vary depending on the intended difficulty of the prompt, then the way prompts are designed is important. If the

timing features vary across prompts in this way then it would be appropriate and necessary to use item prompts pre-calibrated at a variety of ability levels to accurately assess speaking fluency.

If every test taker were able to take the exact same test, the impact of individual prompt fluency variance would not be issue. Since it is not possible to administer the same test with the same prompts in perpetuity, equivalent test forms need to be created and equated to one another (Livingston, 2004). In pursuing that goal test developers need to be careful that their actions are fair to the test takers and those that are making decisions based on the scores. Peterson (2007) noted:

> Perhaps the psychometrician's oath should be essentially the Hippocratic Oath: Do no harm. That is the most important goal of equating. Not only have we produced the best equating possible for all possible test forms or subgroups, but, given all of the problems we might have encountered, we have also produced the best linking the client can afford with minimal negative impact on any subset of examinees. (p. 70-71)

The media ecologist, Neil Postman (1990) declared, "Technology always has unforeseen consequences, and it is not always clear, at the beginning, who or what will win, and who or what will lose" (p. 3). Thus, while we might welcome the promise of ASR technology to facilitate the efficient scoring of speaking tests, we should also be wary of any unanticipated consequences of replacing human rating with machine scoring. Since it is possible to take any speaking sample that exists electronically and score it with ASR, there is the temptation to use ASR to score speech samples and assume the result is a valid indicator of speaking ability regardless of the specific item prompts used to obtain the measure. While there may be some timing fluency features that are invariant across all speech samples that assumption needs to be

verified. The quest for faster and cheaper assessments could come at the expense of their inherent quality.

Research Purpose and Questions

The purpose of this study was twofold: (a) to replicate previous studies that examined how well ASR timing features predicted examinees' speaking ability at the test level and (b) evaluate the extent to which intended prompt difficulty was ordered by the empirical prompt difficulty as rated by humans and scored with ASR timing features.

Based on the purposes of the study, the research focused on the following questions:

1. What combination of ASR timing features best predicts speaking proficiency as measured holistically by human raters?
    a. Which potential predictors were deleted from the model because of multicollinearity or other reasons?
    b. What proportion of the variability in the model is explained by this optimum set of predictors?
2. To what extent is the rater predicted difficulty of the speech prompts ordered as expected for both (a) analytic human rating for each prompt and (b) empirical ASR scoring for each prompt?

Chapter 2

Literature Review

To address the research questions of this study, several characteristics of ASR need to be explored. They include how the ASR timing features could be used to predict overall speaking ability and the relationship of prompt difficulty between ASR scoring and human rating. To that end, this section provides an overview of (a) ASR and its current state, (b) the manner ASR is being used in language testing (c) the role of prompt difficulty in speaking exams and (d) the choice of measurement theory and how it can facilitate the development of equivalent measurements.

Issues in Automatic Speech Recognition (ASR)

To better understand the application of ASR is speaking assessments, it is helpful to examine how ASR functions and the computer programming needed to accomplish those functions. This section will present an overview of ASR technology with some of the implications in its use for speaking assessment. The two broad categories of speech recognition, signal processing (or signal modeling) and audio transcription, will then be explored. That discussion will be followed by a description of different software packages available to conduct the acoustic analysis.

Overview of ASR. ASR is based on pattern recognition and the most widely used application has been with dictation software (O'Shaughnessy, 2008). Human speech is so varied and the individual sounds used to produce words are so context-dependent that the success rate of ASR dictation is highly dependent on either restricting the speaker or the context. For example, ASR recognition rates increase when they have been trained to an individual's voice (Wachowicz & Scott, 1999). Dictation software will have the user read a phonologically rich

paragraph in the initial set-up of the program to help calibrate how the speaker pronounces different sounds. While this increases word recognition rates, the ASR still can have difficulty distinguishing words that are phonologically similar such as *then* and *than*. For ASR dictation to work accurately with different speakers in speaker independent situations, the context is restricted (Wachowicz & Scott, 1999). For examples, companies that use ASR on customer support lines can do so when the context is highly restricted and the response choices are phonologically distinct. It is fairly easy for an ASR to differentiate between *yes, yeah, yup* and other positive response variants and the opposite variants of *no*, *nah, nope*, etc.

The only successful implementations of ASR for use in language testing have been in the cases where the context was highly restricted and the text being recognized was known beforehand. For example, the Pearson Test of English uses ASR to score the responses, but the item types are limited to (a) reading sentences aloud, (b) sentence repetition, (c) saying opposite words, (d) oral short answer responses, and (e) retelling spoken passages (Bernstein, Van Moere & Cheng, 2010).

Other speaking tests, though, use open-ended responses that would be difficult to define *a priori*. This is especially true with questions at the more advanced levels in which examinees could use a wide range of vocabulary domains to answer the question at hand. With this freedom, it might seem impossible to use ASR to recognize the meaning of spoken utterances, yet using an ASR as to process the acoustic signal would allow recognition of *how* the utterance is being said. An ASR can recognize silence, pauses, long pauses, and duration of answer, as well as how closely the individual sounds (or phones) spoken match the sounds (phonemes) of the target language (Muller, 2010). So, if some of these fluency features could predict speaking ability, then it might also be possible to use ASR to rate spontaneous speech, even if it is unable to

recognize the content words that are used. In the ETS Speechrater™ program, researchers used ASR timing features to represent the construct of fluency. In the multiple regression model used to predict speaking scores, the timing features in their equation included (a) articulation rate, (b) number of silent pauses, (c) mean of silent pauses, (d) mean length of run, (e) relative frequency of long pauses and (f) mean duration of long pauses (Xi, Higgins, Zechner, & Williamson, 2012) Ginther, Dimova & Yang (2010) found (a) speech rate, (b) speech time ratio, (c) mean length of run, (d) number of silent pauses and (e) length of silent pauses to have strong ($r = .72$) to moderate ($r = .30$) correlations with overall speaking tests scores. They cautioned, however, that speaking ability consists of more than timing indicators of fluency.

Signal processing. Signal processing is the "process of converting sequences of speech samples to observation vectors representing events in probability space" (Anusuya & Katti, 2011, p. 105). These vectors identify speech from all the other possible sounds in the world. Signal processing includes modeling the vocal features that can later be analyzed. Speech signal characteristics include (a) having a bandwidth signal of 4 kHz, (b) being periodic (fundamental frequency between 80 Hz and 350 Hz), (c) having spectral distribution of energy peaks, and (d) decreasing the power spectrum envelope (-6 dB per octave) with increasing frequency (Anusuya & Katti, 2011). These characteristics are represented numerically by what are called *feature vectors* that are measured at predetermined intervals (e.g. every 10 milliseconds). Those vectors can be used in a variety of applications such as measuring the rate of speech (de Jong & Wempe, 2009) and detecting the individual phonemes. Linguists, phoneticians and others interested in the study of acoustics have conducted research with signal processing with applications including (a) acoustic research on phonetics (Owren, 2008) and prosody(Jeon & Liu, 2012), (b) speaker identification in forensic analyses (Alexander, Dessimoz, Botti, & Drygajlo, 2005;

Kinnunen & Li, 2010), (c) identification of the speakers' emotions (Koolagudi & Krothapalli, 2011; Wu, Falk, & Chan, 2011), and (d) prediction of dialogic responses through prosodic and temporal features (Ward, Vega, & Baumann, 2012). The vectors extracted from signal processing are also a prerequisite part of audio transcription.

Audio transcription. Audio transcription builds upon the work done in signal processing by taking the signals and transferring spoken words to text (Benzeghiba et al., 2007). Government entities, telephone companies and other commercial ventures are among the groups that have conducted research with audio transcription with the applications ranging from ASR for military pilots (Anusuya & Katti, 2011) to portable devices such as navigation systems (Raab, Gruhn, & Noth, 2011) and smart phones (Aron, 2011). Still, the process is fairly complicated. After the ASR has differentiated between sounds produced by the human vocal chords and all other possible sounds (O'Shaughnessy, 2008) and created feature vectors, it can then begin the process of pattern recognition and transcribe what was uttered. As the content the ASR tries to recognize progresses from individual sounds to longer utterances, it must be linked to a natural language processor (Manning & Schütze, 1999). These processors typically include an acoustic model of the language that represents all the phonemes the language contains and a language model that contains the target language vocabulary. Both of these models are based on corpora of the language that have been tagged and are generally searched for through the use of Hidden Marcov Model (HMM) statistical procedures (Huang, Ariki, & Jack, 1990). Through this kind of processing, the ASR first must determine when one word ends and another begins. For example, /aiskrim/ could be the sentence *I scream* or the compound word *ice cream*. The ASR needs to take into account where the word boundaries might be. Beyond that, it needs to

recognize enough context to know if the sound /nait/ refers to *night* or *knight*. These examples illustrate the difficulty in achieving error-free recognition (Chiu, Liou, & Yeh, 2007).

While great progress is being made in ASR technology, the ability to transcribe unrestricted speech from any speaker is still often wanting. The reason for this difficulty is due to the amount of variance in the feature vectors that exist independently of the actual words that are uttered. Individuals that belong to the same group (e.g. gender, regional etc.) can have wide variations in their vocal features. Group differences based on gender, age, regional accent, and native language can all affect the spoken acoustic features. To be successful with unrestricted speech ASR needs to be able to process vocal characteristics that take into account individual variations that occur within groups of speakers and vary systematically between groups of speakers (Kinnunen & Li, 2010). After recognizing sounds, it must then parse those sounds into words and sentences.

Individual variations. The popular sitcom from the 90's, *Seinfeld*, poked fun at the notion that there can be a wide amount of acoustic individual variation of people in the same group. The show introduced characters referred to as the *long talker* (Mehlman, 1994) who did not pause very often, the *low talker* (David, 1993) who spoke quietly, and the *high talker* (Gammill & Pross, 1994) who was a male but on the phone sounded like a female. For an ASR to accurately transcribe test, it must take into account all 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 (Ghosh & Narayanan, 2011). Even individuals whose vocal tracts are physiologically similar, have other speech mannerisms that impact their acoustic signal such as speed, expressiveness, and volume (O'Shaughnessy, 2008). The same individual can have very different vocal characteristics that can impact word

recognition including the individual's emotions (Wu et al., 2011) and vocal effort such as whispering or shouting (Zelinka, Sigmund, & Schimmel, 2012).

Group variations. ASR is a developing technology and the word recognition rates have various degrees of success depending on the contexts that it used. To understand the complexity involved, consider the following vocal features that vary systematically based on speaker characteristics such as gender, age and native language. With gender, the length of the vocal tract of men tends to be longer than that of women resulting in a lower pitch than those of women (Pickett & Morris, 2000). Age affects the voice in two ways. As children grow, the length and shape of their vocal tract fluctuates until they reach maturity. Once maturity is reached, the physical characteristics stabilize. However, as the physiological changes of aging progress, a gradual shift in vocal characteristics occurs (Pickett & Morris, 2000). In the transitional period from middle-aged to elderly, the differences occur more rapidly and can be much more pronounced as the fundamental frequency shifts and there is an increase in vocal tremors (Xue & Deliyski, 2001).

While there are a number of characteristics that differentiate different languages, *voice quality setting* and *rhythm* can aptly illustrate the complexity. The voice quality setting refers to the long-term postures of the vocal tract that are language specific (Esling & Wong, 1983). 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. French speakers, on the other hand, 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 are often transferred to a second language. Thus, the French accent that is detected from French speakers learning English is based to some degree on the voice quality settings of French. With rhythm, each language has it's own unique pattern

stressing words and syllables and the duration of those stresses. For example, Japanese has been classified as a *syllable-timed* language because the length of the syllables is fairly consistent and is not changed by stress (Tajima, Zawaydeh, & Kitahara, 1999). While both English and Arabic are classified as *stress-timed* languages in which the syllable length does change, the timing of those changes has been found to differ. These rhythmic variations can impact a generic ASR's ability to discern some of the language features unique to the languages (Loukina, Kochanski, Rosner, Keane, & Shih, 2011). As language learners retain those rhythmic features to the language they are learning, the accuracy of the ASR will be impacted. While there has been some success in training ASRs to process speech from nonnative English speakers from a single language background like Chinese (He & Zhao, 2007; Sangwan & Hansen, 2012), it is has been far more problematic to program ASRs to process speech samples from diverse native languages as "it is extremely difficult to capture the rather diffuse pattern of variation." (van Doremalen, Strik, & Cucchiarini, 2009, p. 595).

ASR software packages. Different ASR software packages are available to use depending on the needs of the end users. When the speech recognition is limited to an individual, speaker-dependent ASRs have been developed. When speech recognition must function with different, speaker-independent ASRs have been developed. The following section will discuss speaker-dependent and speaker-independent ASRs as well as second language research that has been conducted with them.

Speaker-dependent ASRs. Speaker-dependent ASRs are programmed to a single individual's voice (Kolar, Liu, & Shriberg, 2010; Wachowicz & Scott, 1999). Typically, the ASR has the individual self-select what group he or she belongs and then the individual trains his or her voice to the ASR by reading a series of phonemically rich and varied sentences and

phrases that are known (Kolar et al., 2010). For example, the MacSpeech Dictate software asks first time users to select their accent as (a) American, (b) American-Inland Northern, American Southern, American-Teens, Australian, British, Indian, Latino or South East Asian. Then users are asked to speak at their normal conversational volume and pace as the software calibrates to their voices prior to reading the sentences (*MacSpeech Dictate,* 2010). The ASR can then store the sounds that the users produce as a reference key when transcribing the speech.

This individualized training can reduce the word error rates and can result in the most successful application of ASR for dictation purposes. Unfortunately, this type of individual training is impractical in a testing situation and can also be somewhat undesirable. If an individual examinee were unable to differentiate between /d/ and /th/ in natural speech training the ASR would bias it to the individuals idiosyncratic pronunciation. An examinee saying /duh/ could train the ASR to transcribe the word as *the* when native English speakers would recognize it as the slang word *duh*.

Nonstandard pronunciation can be further complicated by speakers with a shared native language background trying to speak English. For example, Japanese speakers often have difficulty pronouncing /l/ and /r/ (Goto, 1971) and Spanish speakers often have difficulty with /i/ and /ɪ/ (Delattre, 1964). If the ASR were fine-tuned to the Japanese speakers, it might not work as well for the Spanish speakers. More troubling, though, is that customizing the ASR for the language background might increase transcription accuracy when the Japanese speaker states *lice* /lajs/ but intends *rice* /rajs/ or the Spanish speaker states *sheep* /ʃɪp/ but intends *ship* /ʃip/ to the detriment of the language learner. The ASR may recognize what they intend even though a native speaker would not. As a language-learning tool, it could reinforce pronunciation pattern

remnants from their native language. As an assessment tool, it might rate the speech sample as correct when a human rater would not.

Speaker-independent ASRs. Speaker-independent ASRs are designed to recognize any speaker of the language (Ghosh & Narayanan, 2011; Wachowicz & Scott, 1999) . For these applications a wide range of voice types and accents are used to train the ASR. The added complexity will increase the error rate in recognizing words, but it also makes it more useful in L2 learning. With the added complexity of recognizing multiple speakers, an increased success rate is dependent on reducing what the recognizer is processing. For example, it is easier for a speaker-independent ASR to process the words *yes* and *no* when those are the only two options available. Furthermore since those words are phonologically very different, it is easier to differentiate between the two (Anusuya & Katti, 2011). Words that are phonologically similar such as homonyms create more difficulty for the ASR to process. If the ASR processes the word /nait/ and the only option available in the ASR is *night*, then it is easy for the ASR to recognize that word. If both variants were present (i.e. knight and night) then the ASR would have to have more robust programming to determine if the context is the time of day or someone who wears armor. As more words are added to the ASR's possibilities and as the length of what needs to be recognized increases, the error rate in recognizing the utterances will increase.

Most of the research used in the second language research has been speaker-independent ASRs. One program used in a number of second language studies (Cox & Davies, 2012; Millard & Lonsdale, 2011; Okura & Lonsdale, 2012) is Sphinx-4 that was developed at Carnegie Mellon (Lee, 1989). This package is amenable to second language research as it is modular and able to support different languages, their unique grammars, acoustic models and language models. Another program developed specifically for large vocabulary continuous speech recognition is

Julius. It was originally programmed to recognize spoken Japanese, but other acoustic and language models for other languages are being developed (Kawahara & Lee, 2005), and it has been used in speaking assessments of Japanese as a second language (Matsushita, Lonsdale, & Dewey, 2010). In calculating timing features, these programs rely on post-processing the words that were recognized and the time it took to say the words. A weakness of this approach is that often the word recognition rates are calculated on the words that are recognized which can often be quite low (Zechner, Higgins, Xi, & Williamson, 2009). Word accuracy recognition rates could be increased through creating custom dictionaries for the ASR, but that process is expensive and time-consuming. When the calculated timing features are based on words that may or may not be accurate, the reliability of those timing features is suspect.

If audio transcription is not practical, many fluency features, such as timing, can be measured with signal processing software. For example, PRAAT has been used to measure timing features with a script that was used to detect the syllables in an utterance through analyzing the peaks and dips in acoustic intensity (de Jong & Wempe, 2009). This script has been used in a number of studies to extract the data when audio transcriptions were not practical (Christensen, 2012; Ginther, Dimova, & Yang, 2010; de Jong & Wempe, 2009) .

ASR scoring in speaking assessments. Many researchers have explored the technological possibility of using ASR in language pedagogy and assessment. Some of the more common applications include using ASR to provide feedback on pronunciation, using it to score restricted speech such as elicited oral response and incorporating it into speaking test practice software.

Pronunciation tests. A number of researchers have examined the use of ASR in scoring pronunciation. Eskanzi (1999) discussed the use of the Carnegie Mellon's ASR FLUENCY

system to provide pronunciation training for foreign language students. Others have found that accuracy in detecting pronunciation errors increases when the native language of the learners is built into the feedback (Moustroufas & Digalakis, 2007). Price and Rypa (1999) described a prototype of the Voice Interactive Language Training System (VILTS) that used ASR to help students improve oral communication. They found that while the system did not achieve 100% accuracy in detecting errors, the students enjoyed using it and their pronunciation improved. Cucchiarini, Neri, and Strik (2009), building on earlier research (Cucchiarini, Strik, & Boves, 2000), found the use of ASR to give Dutch students feedback on their pronunciation beneficial. Cincarek, Gruhn, Hacker, Nöth & Nakamura (2009) found that English learners from different language backgrounds could be given automatic feedback on word level pronunciation when using tagged corpus that had annotated language errors. SRI International's Eduspeak® provides phone-level feedback on learner pronunciation and has been found to have reliability rates of transcription similar to those of human raters (Franco et al., 2010).

Elicited oral response tests. Some researchers have been specifically looking at the combination of elicited oral response or sentence repetition 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-rated 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 rating and ASR scoring. Other researchers found ASR-scored elicited oral responses to be a suitable speaking assessment for low stakes testing in which the consequence of the outcome will have a minimal lasting impact on the examinee, such as student placement (Cox & Davies, 2012). Furthermore, the use of ASR-scored elicited imitation in other languages including French

(Millard & Lonsdale, 2011), Spanish (Graham et al., 2008), and Japanese (Matsushita et al., 2010) have been found to have high correlations with human rated speaking tests.

Mixed response speaking tests. Others have examined the use of ASR to score speaking tests with mixed item types. Van der Walt, de Wet, & Neisler (van der Walt, de Wet, & Niesler, 2008) developed a speaking test with some restricted responses (e.g. read aloud, sentence repetition) and some open-ended responses (for examinees who were non-standard speakers of English). While they had a small number of participants and had difficulties in receiving reliable ratings from the human raters, they found initial ASR results promising (van der Walt, et al., 2008). In a test of Japanese as a second language, Matsushita (2011) was able to find promising results using machine learning to combine elicited oral response and open-ended speaking prompts to predict the class level of the students. 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 (Bernstein et al., 2010). This combination of task types was first used in Ordinate's PhonePass and is currently used in the Pearson Test of English Academic.

Open response speaking tests. 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. The ASR engine that was used employed previously transcribed responses to train the language model though the word recognition rate was only 53%. The feedback algorithm used a multiple regression equation that found metrics that

represented pronunciation, vocabulary and fluency. They were able to find moderate correlations that concluded that ASR could be used in a low stakes practice environment.

Beigi (2008) analyzed OPI (Oral Proficiency Interview) and OPIc (Oral Proficiency Interview-Computer) data to see if *verbosity*, or the amount of speech uttered in a response, could predict ACTFL (American Council on the Teaching of Foreign Languages) proficiency levels. In an OPI, two interlocutors are present, the interviewer and the examinee. In order to rate the examinee, the ASR was programmed to identify the interviewer's speech and extract it prior to running the analysis. The ASR then transcribed the speech, though the word accuracy rates were not reported. From that data, the verbosity was calculated as well as the rareness of the vocabulary used. With the OPIc, since the only speaker is the examinee, verbosity was calculated by extracting the segments in which audio was present. With both studies, it was found that combing verbosity and rareness of vocabulary used "provide very promising capabilities for the automatic rating of candidates taking the OPI exams" (Beigi, 2008, p. 8).

Others have similarly examined the impact of temporal fluency features on speaking test scores. Ginther, Dimova & Yang (2010) examined the responses of 150 students to a single item on a speaking test. The speakers came from three different first language backgrounds (Chinese, Hindi and English). Their dependent variable was the speaking score on an oral proficiency test and the independent variables were different timing fluency features that were extracted with the phonological software PRAAT. They found strong to moderate correlations between the scores on an in-house speaking test and speech rate, speech time ratio, mean length of run, and the number and length of silent pauses. However, the timing features alone were not enough to distinguish between adjacent levels of the speaking test that they used. In another study Ginther, Dimova & Park (2012) examined the effect of three task types (read aloud, structured compare

and contrast and unstructured response on news item) on the mean length of run. They found that the read aloud task types resulted in longer runs but the other two task types were indistinguishable. Data on individual item difficulty levels, however, were not reported.

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). None of the studies have looked explicitly at the role that item or prompt difficulty had on the variance of the ASR scoring.

Issues in Speaking Assessments

Assessment theory tells us that not all test questions are created equal (Raykov & Marcoulides, 2010). For a variety of reasons some test questions are more difficult than others. This is generally true for any assessment from multiple-choice tests to complex performance assessments. Yet in many assessment contexts questions are treated as if they were equivalent. An examinee is given a series of questions and then each question is added up to create a total score. The underlying assumption is that each question contributed an equal amount of information to the total score. That score is supposed to represent the amount of knowledge, understanding, or ability that an individual taking the test has (Bond & Fox, 2007). When every examinee receives the same set of questions, the variability in the scores is most likely due to differences in the ability level of the examinees. However, when examinees receive different sets of questions, it is difficult to know if the variability is due to the examinee ability or extraneous factors associated with the specific items used (Livingston, 2004).

Performance assessments have an additional layer of complexity as they use human judges in the scoring process. In a speaking performance test, the questions examinees answer are called prompts. While it is hoped that the examinee's score solely reflects ability level, the

score could reflect the prompt difficulty, the subjective criteria of the rater, and any rater bias involved (Eckes, 2011). For example, if two students had the same ability level and the same raters scored their performance, ideally the two examinees would receive the same score. However, if one student were asked to respond to prompts that were more difficult, his inability to adequately respond to the prompts would likely result in a lower score than the student who had easier prompts. In this case the difference in scores would not be due to the examinee's ability, but to the sampling variability of the prompts (Shavelson, Baxter, & Gao, 1993).

Equating tests. One way to minimize the effect of these construct irrelevant factors would be to have examinees take the same test with the same prompts and have the same raters (Eckes, 2011). While this may be practical in small-scale assessments such as classroom tests, it is impractical for large scale testing. First, speaking tests have relatively few prompts, thus test security is a concern as examinees can remember the prompts from the test and share them with others. Furthermore, it is impractical for the same raters to rate the performance of every examinee. This is one advantage that ASR offers—the capacity to score every test.

Prompt difficulty. In order to administer tests with different prompts, conscientious test developer must perform rigorous test equating (Petersen, 2007). This can be done through selecting prompts that have similar item difficulty and discrimination statistics (Livingston, 2004). Another way is to create a test bank of speaking prompts that have been calibrated using Rasch scaling or other Item Response Theory (IRT) models from which items can be drawn to create unique tests (Carr, 2011). Both of these equating methodologies require some mechanism to estimate the difficulty of each prompt or item. Item statistics are calculated in different ways depending on whether the analysis is being done with classical test theory or IRT, but one

prerequisite for either procedure is that each prompt needs to be rated independently (Bachman, 2004).

Holistic tests. When tests are graded holistically, item statistics cannot be estimated. For example, ACTFL OPIs and OPIcs are given a holistic score (Taylor, 2011) based on the criteria described in ACTFL's major proficiency levels (Novice, Intermediate, Advanced or Superior). To be rated at a major level, examinees must show conjoint mastery of all parts of the descriptive rubric for that level. For example, ACTFL's definition of an advanced speaker is someone who can (a) narrate in all major time frames and handle complicated situations or transactions (b) using the text type of paragraph level speech with its attendant cohesion markers, (c) on a wide range of topic/vocabulary domains and (d) do so accurately enough in pronunciation and grammar that the speaker can be easily understood by someone unaccustomed to interacting with non-native speakers. Thus, a person with a strong accent that interferes with their ability to be understood by someone unaccustomed to interacting with non-native speakers would not get a rating of Advanced even if all the other features were present.

For computerized tests like the OPIc, items are developed that target a specific major level and raters make a holistic determination if the examinee is able to sustain performance across all of the areas of the rubric at the major level. In this situation, there is an underlying assumption that each task targeted at a specific major level is equivalent to all the other tasks at that level. There is a further assumption that tasks at higher proficiency levels are more difficult than tasks at the lower levels. For example, a task designed to represent the advanced category would be more difficult than one written to represent the intermediate category.

To create equivalent forms for those tests, expert judgments of item writers are used instead of empirical item statistics to predict an examinee's performance on any given prompt.

This approach has its weaknesses. One study found that by using an information processing approach, test-developers were not able to predict prompt difficulty (Iwashita, McNamara, & Elder, 2002). If a single item does not function well, its effect on the test as a whole is minimized because a trained human rater can use expert judgment to minimize the impact of spurious items. For example, imagine a simple test with two prompts. The first requires examinees to identify the objects in a room and the second requires them to describe their function. In a holistically-scored test, instead of assigning point values to the two prompts and totaling them up to get a total score, raters judge the performance as a whole. Based on a scoring rubric the rater assigns a single value for the entire performance. An examinee may perform poorly on the first prompt and do exceptionally well on the second prompt, but the rater, using expert judgment, might minimize the effect of the poor performance on the first part of the performance awarding a high overall rating to the examinee. With the above example, while the first prompt might be easier, if the test is scored holistically, there is no empirical way to prove it because the details of the scoring for each prompt are not recorded as part of the assessment.

Analytic tests. Item level statistics can only be calculated when each item on the test is either rated or scored (Bachman, 2004). For example, Educational Testing Service's (ETS) TOEFL exam presents a student with 6 speaking tasks that are each rated on a scale of 0 to 4. The sum of the scores is then converted to a scaled score of 0 to 30. As the prompts are summed, there is an underlying assumption that each task is equivalent in difficulty level.

Suppose several examinees took the two-item test described previously and information on how well each individual performed for each prompt were recorded. In this scenario we could empirically calculate the difficulty of each prompt based on the assumption that students would do less well on more difficult items. With item statistics available, it is then possible to

create item banks that include details regarding the difficulty of a specific question prompt. Test developers could then create equivalent forms of an assessment based on item difficulty (i.e., having a similar number of items with specific difficulty ratings on each test).

Speaking prompt difficulty, however, has been relatively unexplored (Fulcher & Reiter, 2003). One study examined if test developers could predict the prompt difficulty through the use of an information processing approach (Iwashita et al., 2002). They examined different task dimensions including the number of elements in the prompt, the abstractness of the information, the type of information, the nature of the operation and the familiarity of the task and found that predicted item difficulty did not align with the calculated item difficulties. Another study looked at the effect of the native language, cultural background, and pragmatic task features on the prompt difficulty and found that there was a significant interaction between the language background, social power, and the degree of imposition embedded in the tasks (Fulcher & Reiter, 2003). In this study, they cautioned test-developers to be sensitive on how examinees' language and cultural background could be a source of Differential Item Functioning (DIFF). In a separate study involving speaking test validation, prompts were found to have a separation reliability of .93 indicating that the items could be separated into different levels (Kim, 2006). However, there was no attempt to predict the prompt difficulty before hand or to align the prompts based on difficult for other forms of the test.

Human rating vs. ASR scoring of prompts. There are important differences in human rating and ASR scoring of prompts. With human rating, the rater can judge the whole performance including content, organization, lexical diversity, fluency and other factors. The ability to judge the whole instead of being limited to individual variables is one of greatest strengths of human raters. The weakness is that single raters tend to follow their own

idiosyncratic patterns of scoring and this can lead to unreliability in scoring (Fulcher, 2003). To compensate for this weakness, many performance tests have multiple human raters, but even then there is still the potential for error variance in the ratings (Eckes, 2011). Extensive rater training can minimize some of the random error associated with having multiple raters, but even training cannot ensure expert judges will agree in all instances (McNamara, 1996).

One advantage of using ASR is the possible elimination of one source of error in the assessment—the variability that comes from different human raters. Because a computer scores every item on every test, the rating of the speaking tests should be more consistent (i.e., reliable). However, because of technological limitations, ASR is highly unlikely to measure every essential element that comprises spontaneous speech. ASR is currently limited to measuring proxy variables designed to assess various aspects of speak that might be used to provide a reasonable estimations of the performance's quality. A human rater can take into account semantics and meaning as they score while current ASR technology simply measures the related proxy variables. If those ASR features systematically change from one prompt to another, in ways different to that of humans scoring the prompts, then the differences in alignment based on the item statistics could affect the validity of comparing test results comprised of different items.

Imagine an item bank with two prompts that are estimated by experts to be equivalent. The prompts are intended to elicit the same kind of language and the learning objective is considering the function of past narration. The first prompt— a personal narration—requires examinees to narrate in the past and describe *their* first day at school. The second prompt—a picture narration— gives the examinees a series of cartoon pictures that illustrate a young lady's first day at school and requires them to narrate in the past and describe *her* first day at school. Successful completion of both of these prompts should produce the same type of language in that

each should elicit a response that requires the student to demonstrate the ability to narrate a past experience using verbs in past tenses while using vocabulary associated with schools. This would need to be done using an appropriate rate of speech, accurate pronunciation and a paucity of unnatural pauses so that a native speaker could easily understand the response. Note that the fluency features, while necessary to answer the response, may provide insufficient evidence, on their own, that the person has responded in an acceptable manner.

Examinees might respond to the two prompts and receive the same overall holistic score from a human rater (based on content, grammar, fluency, etc.), but when looking at fluency alone, the individual performances could differ. The expected outcome is that the fluency timing features would be the same regardless of the prompt used. For example, if the examinees had low fluency for the Personal Narration prompt, they would have low fluency for the Picture Narration Prompt. Conversely; and if the examinees had high fluency for the Personal Narration prompt, they would have high fluency for the Picture Narration Prompt. In this case the two prompts could be used interchangeably in an item bank that is scored via ASR. However, it is also possible that the results of the timing features could be different for the two prompts. Examinees could exhibit high fluency in the Personal Narration Prompt (e.g. they could easily recall their first day at school and speak quickly with few pauses), but low fluency in the Picture Prompt (e.g. they spoke more slowly and hesitated because of the cognitive load needed to interpret the cartoon). Or vice versa, examinees could exhibit low fluency in the Personal Narration Prompt (e.g. they had difficulty remembering their first day of school) and high fluency with the Picture Prompt (e.g. they could speak quickly because the content was provided).

If the fluency timings for examinees systematically from what a human rater would have awarded, then there is a confounding factor that would affect the reliability and validity of the scoring when using ASR as a predictor of speaking ability. While a human judge might take the prompt into account and could weigh the entire content of the response when awarding a score, the ASR can only look at the fluency proxy variables and that may not be sufficient to establish a score. Thus it is important to examine how prompt difficulty (or intended difficulty) affects the ASR scoring of timing features.

Measuring speaking assessments. One criticism of scoring in the human sciences including speaking assessments is that the data are presumed to be interval when that presumption has not been tested empirically. Stevens (1946) in his seminal work on types of measurement scales noted that most of the scales used by researchers in the social sciences were actually ordinal, and that the parametric statistics used are in "error to the extent that the successive intervals are equal in size (p. 679)." The tendency to assign numbers to objects and then treat the numbers as interval data in doing statistical tests still persists (Bond & Fox, 2007; Crocker & Algina, 1986; Raykov & Marcoulides, 2010). One way to ensure that the data truly meet interval criterion is by converting the *raw scores* to *measures*. Raw scores are the observed counts in their original state with no statistical adjustment (Bond & Fox, 2007). Measures are derived by assigning numerals to objects based on rules (Stevens, 1946). For the measures to be interval, the rules require that the numerals are assigned in a linear manner based on a scale (Crocker & Algina, 1986).

Characteristics of interval data. In classical test theory, an examinee's ability is estimated by a score based on the total number of test items answered correctly (Brown, 1996). An item's difficulty is calculated by dividing the number of examinees who answer the item

correctly by the total number of examinees (Bachman, 2004). Both of these measures are dependent on the population of examinees who took the test and the items that were included on the test (Crocker & Algina, 1986), and there is no guarantee (and is in fact unlikely) that the resulting measures have the properties of interval data (Bond & Fox, 2007). The property of equal-intervalness requires that space between any two adjacent scores be equidistant (Stevens, 1946). Furthermore, the same distance between two points should demonstrate the same increase in ability regardless of where it falls. So, if an examinee takes a pretest and has a score of 10 and takes the posttest and has a score of 15, it would be assumed the examinee gained in skill by five points. To be interval data, that ability increase of five should have the same significance wherever the score increases, though most would find an increase of five from 2 to 7 to signify a different amount of growth than a score increase from 18 to 23. While most social scientists acknowledge that the data from their test scores is not truly interval in this sense, they still use parametric statistics (Wright & Linacre, 1989). While some might argue that parametric statistics are robust enough to use with ordinal data (Knapp, 1990; Norman, 2010), the use of Rasch scaling can make the criticism a moot point.

Rasch scaling. The Rasch procedure transforms person ability and item difficulty estimates into measures called logits (Baylor et al., 2011). Logits (or log odds ratios) are the natural logarithm of odds ratios of success and can be converted to and from probabilities. For example, if someone has a 0.6 probability of answering an item correctly, then the odds of them answering the item correctly is .6/.4 = 1.5 or 1.5 to 1. The odds ratio, as the name indicates, is on a ratio scale and is therefore constrained to multiplicative arithmetic (Linacre, 1991). By being transformed to a log odds ratio, the measures are now interval data and have additive properties.

Those logits can then be transformed back into probabilities. Georg Rasch, the Danish mathematician who developed the measurement model, stated

> In simple terms, the principle is that a person having a greater ability than another person should have the greater probability of solving any item of the type in question, and similarly, one item being more difficult than another means that for any person the probability of solving the second item is the greater one. (Rasch, 1960 p. 117)

There are two assumptions that must be met in order to perform Rasch scaling: (a) *local independence* and (b) *unidimensionality* (Bond & Fox, 2007; DeMars, 2010). Local independence assumes that the response of any item on a test is independent of the response of any other item. Language testers have been able to meet this assumption successfully in speaking tests through deliberate item creation (Adams, Griffin, & Martin, 1987; Griffin, 1985).

Unidimensionality assumes that the trait being measured "share a common primary construct" (DeMars, 2010, p. 38). This assumption has created great controversy in the language testing community as many find the complexity required to engage in any form of communicative competence is comprised of multiple dimensions (McNamara & Knoch, 2012). Henning (1992) argued that there was a difference between psychometric and psychological unidimensionality and the assumption that needed to be met was psychometric in nature. Even Wright & Linacre (1989) point out that no test is perfectly unidimensional and that it is a qualitative rather than a quantitative concept. Some have argued that when data is multidimensional, mathematical models that address that condition such as multidimensional item response theory (MIRT) should be used (Reckase, 2009). One researcher, Ip (2010), however, found in a theoretical investigation with multidimensional data with a dominant or essential dimension that "MIRT is empirically indistinguishable from a locally dependent

unidimensional model" (p. 407). Most investigations in language testing use unidimensional models (McNamara & Knoch, 2012). Stansfield & Kenyon (1996) were among the first to use Many Facets Rasch Measurement with speaking tests when they speaking prompts that were based on the multidimensional speaking rubric of the ACTFL speaking proficiency guidelines.

In Rasch scaling, person ability and item difficulty are measured conjointly so that an examinee with a person ability estimate of a given value will have a 0.50 probability of answering an item with a difficulty parameter of that same value correctly (Linacre, 1991). So if an examinee has a person ability estimate logit of 1.00 and a prompt has an item difficulty parameter logit of 1.00, the probability that examinee responding to that prompt correctly is 50-50 or odds of 1 to 1. If an examinee has a person ability estimate of 1.00 and the prompt has an item difficulty parameter of -1.00 for a distance of two logits between person ability and item difficulty, than the probability of the examinee responding to that prompt correctly is 0.88 or odds of 7.3 to 1.

Ensuring measurement invariance. Besides being interval data, another advantage of using Rasch scaling is that the parameter estimates for both persons and items have the quality of *measurement invariance* (Engelhard Jr, 2008). That is, when measuring a unitary construct, person ability estimates are the same regardless of the items that are presented to the examinees, and item ability estimates are the same regardless of the examinees who respond to them. Since the application of the findings of this study are directed for test developers in equating test forms, measurement invariance of the items is highly relevant. Beyond the advantage of measurement invariance, the Rasch analysis can provide information on how well a scale functions and the reliability of the test scores and test items.

Diagnosing rating scales. To evaluate how well a scale functions with Rasch measurement, there are a number of diagnostics available including (a) category frequencies, (b) average logit measures, (c) threshold estimates, (d) category probability curves, and (e) fit statistics (Bond & Fox, 2007). For category frequencies, the ideal is that there should be a minimum of 10 responses in each category that are normally distributed. For average logit measures, the average person ability estimate of each rating category should increase monotonically (Eckes, 2011). The threshold estimates are the logits along the person ability axis at which the probability changes from a person being in one category to another. Those estimates should increase monotonically as well. In order to show distinction between the categories, they should be at least 1.4 logits apart and to avoid large gaps in the variable and the estimate should be closer than five logits (Linacre, 1999). When looking at a graph of the category probability curves, each curve should have its own peak, and the distance between thresholds should be approximately equal. If one category curve falls underneath another category curve or curves, then the categories could be disordered and in need of collapsing. Finally, fit statistics provide one more way to examine a rating scale. If the outfit mean squares of any of the categories are greater than 2.0, then there might be noise that has been introduced into rating scale model (Linacre, 1999). Using these diagnostics through a FACETs analysis, a measurement scale can be analyzed.

Analyzing reliability. Finally, Rasch scaling provides more tools in determining the reliability of test scores, especially when there are multiple facets. Reliability is defined as the ratio of the true variance to the observed variance (Crocker & Algina, 1986). Unlike classical test theory which can only report reliability on the items of a test (e.g. Cronbach's Alpha or Kuder-Richardson 20) or the agreement or consistency of raters (e.g. Cohen's kappa. Pearson's

Correlation coefficient), Rasch reliability reports the relative reproducibility of results by including the error variance of the model in its calculation. Furthermore Rasch reliability provides estimates for every facet (person, rater, item) that is being measured. When the reliability is close to 1.0, it indicates that the observed variance of whatever is being measured (person, rater, item) is close or nearly equivalent to the true (and immeasurable) true variance. Therefore, when person reliability is close to 1, the differences in examinee scores are due to differences in examinee ability. If there are multiple facets such as raters, it might be desirable for a construct irrelevant facet to have a reliability estimate close to 0. If raters were the facet, the indication would be that they were indistinguishable from each other and therefore interchangeable. Any examinee would likely obtain the same rating regardless of which rater were assigned to them. Conversely, if the rater facet had a reliability estimate close to 1.0, then the raters are reliably different and the rating obtained by a given examinee is highly dependent on the rater. When the rater facet is not close to 0, it is necessary that an adjustment be made to the examinee score to compensate for the rater bias.

Summary of Literature

ASR is a complex process that continues to improve but still has limitations. If the text of the speech sample is known beforehand, recognition rates improve. ASR has been successfully used to score speaking assessments where the content is narrowly defined. Spontaneous speech has been more problematic due to the lack of precision in recognizing words. Proxy variables including timing features (e.g. rate of speech, number of pauses, etc.) and proximity to the acoustic system of the target language have been explored for their potential to rate spontaneous speech. No studies have looked specifically at the effect of the speaking prompt difficulty on the proxy variables ASR can successfully recognize. To examine prompt

difficulty, it is important to choose a measurement theory that is robust in its ability to provide diagnostic information on the rating scales used, reliability of the facets being analyzed, and stable difficulty parameters across different testing population. Rasch scaling has been found to meet those criteria as well as provide interval level data that can be used in parametric statistics.

Chapter 3

Methods

The purpose of this study was to establish (a) the ASR timing features that could be used to predict human-rated speaking ability and (b) the extent to which the intended prompt difficulties aligned with ASR scoring and human ratings. This section describes the data collection and analysis procedures used in this study.

Study Participants and Test Administration Procedures

The subjects participating in this study were students enrolled at the *English Language Center* taking their exit exams during winter semester 2012. There were 201 students who spoke 18 different languages (see Table 1).

Table 1

Composition of Subjects by Language and Gender

	Gender			
Native Language	Female	Male	Total	Percent
Arabic	0	3	3	1.5
Armenian	0	1	1	.5
Bambara/French	0	1	1	.5
Chinese	9	4	13	6.5
French	2	0	2	1.0
Haitian Creole	1	3	4	2.0
Italian	0	3	3	1.5
Japanese	0	5	5	2.5
Korean	21	16	37	18.4
Mauritian Creole	1	0	1	.5
Mongolian	3	1	4	2.0
Portuguese	13	15	28	13.9
Russian	2	1	3	1.5
Spanish	54	34	88	43.8
Tajik	1	0	1	.5
Thai	2	0	2	1.0
Ukrainian	3	0	3	1.5
Vietnamese	2	0	2	1.0
Total	114	87	201	
Percent	56.7	43.3		

The students were in the school to improve their English to the point at which they could successfully attend university where the language of instruction was English. They ranged in

speaking ability from novice to superior. Note that three subjects had some audio files that did not record for some of the prompts, therefore there were only 198 complete data sets. The assessment was administered to students as part of their final exams.

Design and Validation of the Data Collection Instrument

The research instrument used in this study was designed to assess speaking ability at proficiency levels 2 through 6 (see Appendix A). It was assumed that after one semester of instruction, all the examinees participating in this study would have some ability to speak English yet none would be considered the functional equivalent of highly educated native speakers. Each level of the speaking rubric was tied to a class level, thus a student with a score of 2 would be ready to study at the Foundations B level. A student with a Level 4 would be ready to study in the Academic A level (see Table 2). The test included 10 prompts, two for each of the targeted levels.

Table 2

Description of Program Level Rubric Scale Scores and OPI Equivalence

Program Level	Rubric Scale/Level	OPI equivalence
Foundations Prep	0	Novice Low
Foundations A	1	Novice Mid
Foundations B	2	Novice High
Foundations C	3	Intermediate Low
Academic A	4	Intermediate Mid
Academic B	5	Intermediate High
Academic C	6	Advanced Low
	7	Advanced Mid

The speaking test was designed with the same framework as an interview-based test. An interview test has four stages. First, the interviewer begins with the easiest items as a warm-up for the examinee. Following the warm-up, the interviewer establishes a baseline at which the examinee can easily function. The interviewer then probes and progresses to more difficult items to see where the examinee experiences breakdown. If the examinee sustains performance,

the interviewer establishes a higher baseline and continues to increase the difficulty to see how much ability the examinee has. If the examinee is unable to sustain performance, the interviewer returns to the baseline. The last part of the interview is a wind-down with question that brings the examinee down to a level at which they can easily respond. Since this test was not adaptive, the items were structured to progress from easier to harder (using prompt 1 at each level) and then back down (using prompt 2 at each level) in a pyramid shape. The prompts were designed in this was so the respondent would be required to demonstrate their ability to speak at the targeted level. The items were designed with varying amounts of preparation time and response time as determine to meet the function of the prompt (see Table 3).

Table 3

Speaking Test Labels, Preparation and Response Times

Label (Level-Item)	Level	Prompt Number	Preparation Time	Response Time
L2-1	2	1	15	45
L2-2	2	2	15	45
L3-1	3	1	45	45
L3-2	3	2	15	45
L4-1	4	1	45	45
L4-2	4	2	30	90
L5-1	5	1	15	45
L5-2	5	2	15	45
L6-1	6	1	45	90
L6-2	6	2	30	90

To determine to what extent the prompts on the instrument aligned with their expected difficulty level, a panel of expert raters was consulted. The rating rubric had been in use for six semesters so the maximum number of semesters a rater could have rated was six. The expert panel consisted of eight raters with an average of 4.75 semesters of rating experience (SD = .88, Range = 3 to 6 semesters).

For each prompt, the raters

- predicted the level of ability on the speaking score rubric that an examinee would need to adequately respond to the prompt,
- identified what objectives were being measured, and
- provided feedback on whether they felt the item would function as intended.

The results of the ratings assigned by the eight raters were to obtain the rater predicted difficulty. The raters were presented 15 prompts, and based on their feedback, 10 prompts were selected for inclusion on the test. The rater predicted difficulties of the 10 selected items rose monotonically in that every Level 2 prompt was easier than every Level 3 prompt and every Level 3 prompt was easier than the Level 4 prompts, etc. (see Table 4).

Table 4

Rater Predicted Difficulty Scores by Test Prompt

Label	Mean Difficulty	SD
L2-1	1.86	0.69
L2-2	2.00	0.76
L3-2	2.71	0.95
L3-1	2.71	0.76
L4-1	3.13	0.83
L4-2	3.38	0.74
L5-2	4.63	0.52
L5-1	4.75	0.89
L6-2	5.00	0.76
L6-1	5.50	0.53

Ratings based on average from 8 different raters

Since the expert rater predicted levels rose monotonically based, it was considered to be evidence that the prompts did reflect the scale descriptors. The rater predicted difficulties were also used to examine the extent to which the estimated difficulty of the speech prompts ordered as expected for both the analytic item level human rating and ASR scoring.

Rating and Scoring Procedures

The scoring rubric for the human rating of the assessment addressed three axes: (a) text type (e.g. word and phrase length, sentence length, paragraph length, etc.), (b) content, and (c) accuracy. Each axis ranged from *no ability* to *high ability* (i.e. the functional equivalent of a well-educated highly articulate native speaker). The scale was intended to be noncompensatory so that a response that is native-like in one area (e.g. pronunciation) could not compensate for a weak performance in another area (e.g. a text type that was only word length).

Since this research examines human rating and ASR scoring, the results obtained from administering the instrument were analyzed in two different ways. Human rating was conducted by two separate groups of raters: one group rated the tests holistically and one group rated the tests analytically (at the item level). The ASR scoring was used on each item of the test and aggregated to create total scores on the test. To ensure the results had the characteristics of interval data and fully justified the use of parametric statistics, both the human ratings (holistic and analytic) and ASR timing measure were converted from raw scores (typically used in classical test theory) to logits and/or the equivalent fair average.

For the human ratings of speaking ability, two different rating schedules were used: one for the raters who rated the tests holistically and one for the raters who rated the tests analytically at the item level (see Appendix B and C). For the ASR scoring, the signal processing software PRAAT was used.

Human rating. The tests were rated on an 8-level scale that roughly corresponded to the ACTFL OPI scale (see Table 2) by raters with ESL training who were working as teachers. This rubric was used for both the holistic ratings and the item level ratings. All of the raters had been trained at various times to use the rubric for the regularly scheduled computer-administrated

speaking tests. The raters had received over 3 hours of training and completed a minimum of 12 calibration practice ratings to ensure sufficient knowledge of the rubric. The existing rater training material was designed to train raters how to use the 8-level scale on a test that was scored holistically. The raters had a packet that contained a copy of the rubric and a printed copy of the exam prompts, the objective of the prompt and the intended difficulty level of the prompt.

Rating designs. In choosing a rating design with human raters, it is important to balance the amount of information gained from a specific design and the cost needed to employ the raters. A *complete* or *fully crossed* design in which all of the raters rate all of the items and examinees provides the most information and is considered the best from a measurement standpoint. This design leads to the most stable parameter estimates as there are no missing data links (Eckes, 2011). It is also the most expensive and is thus not practical to use in most cases (Sykes, Ito, & Wang, 2008). If one is conducting a Many Facet Rasch Measurement analysis, however, a complete design is not a pre-requisite (Linacre, 1994).

Connected designs in which raters rate the same subset of the items or examinees can still provide enough data links to allow all of the facets to be connected (Schumacker, 1999). When there are ***not*** enough data links, a many facet Rasch analysis will result in *disjointed subsets* (Schumacker, 1999), which makes it inappropriate to make comparisons. Connected designs can be engineered by ensuring there is overlap between the facets. The more overlap that occurs, the more stable the parameter estimates are (Eckes, 2011). This connectivity allows various facets to be compared on a shared common metric that contains the rating categories and the facets being analyzed (e.g. examinee, rater, item, etc.). Because it is cost effective and sufficient for Many Facets Rasch Measurement, incomplete connected designs were used for this study.

Human-rated holistic speaking level rating design. All of the tests were rated holistically using the 8-point scale in an *incomplete connected* design. As the existing training materials were designed for holistic ratings, no modification of instructions had to be given to the raters. They simply had to rate in the manner that they had always rated speaking tests. In this design each student was double-rated, all the raters rated a subset of students and then each rater was paired with every other rater. This kind of design has been found to provide sufficient connectivity between raters and examinees (Yu & Brown, 2000). Table 5 provides an example of an incomplete, connected design representing 20 examinees, 5 raters and 10 prompts. This kind of design was necessary to ensure there were enough connections in the data to compute the data points. For the actual study, the design had 201 students, 10 raters and 10 prompts. The complete rating design with the raters and examinees involved can be found in Appendix B. This design provided the data necessary to answer the first question on how well the fluency features could predict the overall speaking score of a test scored by human raters.

Table 5

Incomplete Connected Design for Holistically Rated Speaking Test

	Raters				
Students	1	2	3	4	5
1–10	X	X	X	X	X
11	X	X			
12	X		X		
13	X			X	
14	X				X
15		X	X		
16		X		X	
17		X			X
18			X	X	
19			X		X
20				X	X

Analytic human-rated item speaking level rating design. To answer the second question and get analytic human-rated item level statistics for comparison with the ASR results, each test had to be rated at the item level. There were two possible incomplete connected design

possibilities that could have provided the requisite data. The first was an *incomplete, connected* design in which all the items on a single test were rated by raters who were linked to other raters. While this design is more cost-effective than a complete design, examinee ability estimates can be biased if there is an "unlucky combination of extreme raters and examinees" (Hombo, Donoghue, & Thayer, 2001, p. 20). The second design possibility was an *incomplete, connected spiral* design. This design was differentiated from the prior by assigning individual items to raters and linking raters to other raters through shared item ratings (Eckes, 2011). This design shared the cost-effectiveness of the incomplete, connected designs, but has some distinct advantages. First, when raters listen to the same item from different examinees, they can have a deeper understanding of the response characteristics needed to assign a rating. Second, the spiral design can minimize errors associated with the *halo effect*. Halo effect occurs when performance on one item biases the rating given on subsequent prompts (Myford & Wolfe, 2003). For example, if a rater listens to a prompt and determines the examinee to speak at a Level 4 based on the rubric, then the rater might rate all subsequent prompts at 4 even when the performance might be higher or lower. Finally, spiral rating designs have been found to be robust in providing stable examinee ability estimates in response to rater tendencies (Hombo et al., 2001).

For this design, each rater was assigned to rate a single prompt (e.g. rater 1 scores all of prompt 1, rater 2 scores all of prompt 2, etc.). To avoid having disconnected subsets, a subset of the same students was rated on each item by all the raters. To further ensure raters were familiar with the items, raters rated some additional tests in their entirety. Table 6 is an example of an incomplete, spiral design representing six examinees, four raters and four prompts. For the actual study, the design included 201 students, 10 raters and 10 prompts (see Appendix C).

Table 6

Incomplete Spiral Connected Design for Analytically Rated Speaking Test by Prompt

		Raters			
Students	Prompt	1	2	3	4
1-4	1, 2, 3, 4	X	X	X	X
5	1	X	X		
5	2		X		
5	3		X	X	
5	4		X		X
6	1	X		X	
6	2		X	X	
6	3			X	
6	4			X	X
7	1	X			X
7	2		X		X
7	3			X	X
7	4				X
8	1	X			
8	2		X		
8	3			X	
8	4				X
9	1	X			
9	2		X		
9	3			X	
9	4				X

Since all existing training materials for the rubric were designed in rating tests holistically, these raters had to be given separate instructions. They knew the intended level of the prompt they were scoring, and were told to reference that as they applied the rubric. For example, when rating a prompt that was designed to elicit Level 2 speech samples (ask simple questions at the sentence level), a rater was able to use the entire range of categories in the rubric (0 to 7). Since a rating of 2 would be passing, the only way the higher categories would be used is if the examinee spontaneously used characteristics of those higher categories through the use

of more extended discourse, more academic vocabulary, native-like pronunciation, etc. These instructions were deliberate so to avoid having a restrict of range error (Myford & Wolfe, 2003). This analysis provided the data necessary to answer the second question on how human-rated item difficulties compare to item difficulties computed via ASR.

ASR scoring. For this analysis we used PRAAT ASR software to extract the timing features of the ten different prompts for each of the students being tested. While a few different ASR software packages exist, we chose PRAAT, an open source phonetic software package (Boersma & Weenink, 2005) as it recognized the features more accurately. Table 7 has a detailed list of all the features that were extracted, but they included the (a) total response time, (b) speech time, (c) speech time ratio, (d) number of syllables, (e) speech rate, (f) articulation rate, (g) mean syllables per run, (h) silent pause time, (i) number of silent pauses, (j) mean silent pause time, and (k) silent pause ratio.

Data Analyses to Address Research Questions

To best answer the research questions, a two-step data analysis procedure was followed. In the first step, a Rasch analysis was conducted to verify the functionality of the scale and the reliability of the test scoring method in separating the facets being analyzed.

The programs that were chosen to do the Rasch scaling analyses were FACETS and Winsteps. For the human-ratings, FACETS was used as it can examine parameters beyond person and item including raters and it can compensate for rater bias. This ensured that the items of the examinees were measured as if they had been rated by the average of all the raters. For the ASR ratings, Winsteps was used as it is the recommended default for only two parameters: persons and items (Linacre & Wright, 2009). In the second step, parametric statistics were used to perform either the correlations or regressions, depending on the research question.

Table 7

Speech Timing Fluency Features

Feature	Description
Total Response Time	Speaking + silent pause time (e.g. duration of audio file)
Speech Time	Speaking time, excluding silent pause time.
Speech Time Ratio*	Speech time/total response time.
Number of Syllables	Total number of syllables in a given speech sample was obtained to calculate mean syllables per run, speech rate, and articulation rate.
Speech Rate*	Total number of syllables divided by the total response time in seconds.
Articulation Rate*	Total number of syllables divided by the speech time.
Mean Syllables per Run*	Number of syllables divided by number of runs in a given speech sample. Runs were defined as number of syllables produced between two silent pauses. Silent pauses were considered pauses equal to or longer than 0.25 seconds.
Silent Pause Time	Total time in seconds of all silent pauses in a given speech sample.
Number of Silent Pauses per minute*	Total number of silent pauses per speech sample. Silent pauses were considered pauses of 0.25 seconds or longer.
Mean Silent Pause Time*	Silent pause time / number of silent pauses.
Silent Pause Ratio*	Silent pause time as a decimal percent of total response time.

*Indicates ASR measures that are standardized and can be compared across prompts that are of varying lengths

Research question 1: Use of ASR to predict speaking scores. To answer the first research question and determine the best combination ASR fluency features that predict human-rated speaking proficiency, the following steps were performed. First the human-rated holistic speaking levels were determined using Many Facets Rasch measurement. This produced an estimate of the human-rated holistic speaking level, the dependent variable, based on the fair average of various raters. That was followed by a multiple linear regression to determine which of the ASR timing fluency variables, the independent variables, best predicted student performance as measured by human-rated holistic speaking level.

The facets used for this analysis were examinees and raters. As the first research question only examined the test holistically, prompts were not included in the equation. If the rating scale used across the elements of the facets is constant, the Andrich Rating Scale model is the most appropriate to use (Linacre, 2009). The basic MFRM model to analyze the data can be specified as follows:

$$\ln\left[\frac{p_{njk}}{p_{njk-1}}\right] = \theta_n - \alpha_j - \tau_k \qquad \text{(Equation 1)}$$

where

p_{njk} = probability of examinee *n* receiving a rating of *k* from rater *j* ;

p_{njk-1} = probability of examinee *n* receiving a rating of *k-1* from rater *j*;

θ_n = ability of examinee n;

α_j = severity of rater j;

and

τ_k = difficulty of receiving a rating of *k* relative to *k-1* using a Rasch-Andrich threshold or step calibration scale.

The rubric used for the rating scale (see Table 2) had eight categories and was based on the levels of the program. The examinee fair average score represented the examinees' person ability estimate (θ) and showed the adjusted rating the examinees would have received by controlling for variance in the other facets. A sequential multiple regression was used to determine which combination of ASR timing features best predicts speaking ability. This type of analysis further informs which potential predictors were excluded from the model because they were too highly correlated and thus had multicollinearity. Further, this procedure accounts for the proportion of variability in the human-rated holistic speaking level due to the predictor variables. Multiple regression was used as its "stability, parsimony and algorithmic simplicity" makes it preferable to other methods (Xi et al., 2012).

Research question 2: Impact of prompt difficulty. The second question explored the extent to which the estimated difficulty of each speech prompt was ordered as expected using both the human ratings and ASR scores for each item. The expected difficulty order of the prompts was operationalized by the rater predicted difficulties.

To obtain analytic item level human ratings, the analytic human-rated item speaking level was calculated using a FACETS analysis. The three facets for this analysis included examinees, raters, and prompts. Since the raters used a holistic 8-point scale analytically at the prompt level, the Andrich Rating Scale model was the most appropriate to use (Linacre & Wright, 2009). With the rating scale used across the elements of the facets being held constant, the basic MFRM model to analyze the data for this question was specified as follows:

$$\ln\left[\frac{p_{nijk}}{p_{nijk-1}}\right] = \theta_n - \beta_i - \alpha_j - \tau_k \qquad \text{(Equation 2)}$$

where

p_{nijk} = probability of examinee *n* receiving a rating of *k* from rater *j* on prompt i ;

p_{nijk-1} = probability of examinee *n* receiving a rating of *k-1* from rater *j* on prompt *i*;

θ_n = ability of examinee n;

β_i = difficulty of prompt;

α_j = severity of rater j;

and

τ_k = difficulty of receiving a rating of *k* relative to *k-1* using a Rasch-Andrich threshold or step calibration scale.

The rubric used for the rating scale (see Table 1) was the same as the human-rated holistic speaking level scale but applied to individual prompts (or items) on the exam. The analytic human-rated item speaking level was determined using item fair average scores that represent the item's difficulty parameter (β_i) and show the adjusted rating the item would have received by controlling for variance in the other facets. The categories of the 8-point scale used were also analyzed using a FACETS analysis.

To obtain analytic item level ASR scoring, ASR Regression Predicted Analytic Speaking Levels were calculated. The Regression Predicted Analytic Speaking Levels were determined by applying the regression equation established from the first research question to each prompt of each student to award a score of 0 to 7 mirroring the same 8-point rubric the human raters used. Then, the entire test was analyzed with Rasch scaling. Since the ASR is a single rater and the effect of rater bias does not need to be mitigated using a FACETs analysis, the Winsteps program was deemed to be the best choice in establishing item difficulty parameters.

At this stage, there are three different item statistics that can be compared: rater predicted difficulty, analytic human-rated item speaking level and regression predicted analytic speaking level. The ordering of these item difficulty indices was compared using correlations.

Summary of Methods

To evaluate the potential of using ASR timing fluency features to predict speaking ratings and to examine the effect of prompt difficulty in that process, a speaking test with ten prompts was administered to 201 subjects. The speech samples obtained were then: (a) rated holistically with all ten prompts combined by one set of human raters, (b) rated analytically with all ten prompts separated by a different set of raters, and (c) scored automatically using PRAAT to calculate ten different ASR timing fluency features. The ratings and scores of the speech samples were analyzed with Rasch measurement to evaluate the functionality of the scales and the separation reliability of the examinees, raters, and items. The resulting person and item measures were then used to explore the potential of using ASR timing features to predict human-rated speaking tests and the effect of the prompt difficulty.

Chapter 4

Results

This study examined the potential of using ASR timing fluency features to predict speaking ratings and to examine the effect of prompt difficulty in that process. To accomplish that goal, a preliminary Rasch analysis was conducted to see how well the scales functioned and how reliable the test scoring was for the human-rated holistic speaking level and the analytic human-rated item speaking level. Those results are presented in the preliminary Rasch analysis section and followed by the findings for the first and second research question.

Research Question 1: Use of ASR to Predict Speaking Scores

The first research question asked what combination of ASR timing features best predicts speaking proficiency as measured holistically by human raters? Subquestions were which potential predictors were deleted from the model because of multicollinearity or other reasons and what proportion of the variability in the model is explained by this optimum set of predictors? The Rasch analysis was conducted to diagnose the usefulness of the scale categories and calculate the separation reliability of the facets.

Phase 1: Rasch analysis of human-rated holistic speaking level. The human-rated holistic speaking level represented the scale used by the human raters when they rated tests holistically. As noted earlier, the 8-level scale was derived from the ACTFL proficiency guidelines and was tied to different class levels of the intensive English program. An analysis of the functionality of the scale is followed by a reliability analysis of the test scores from the use of the scale.

Scale diagnosis. While not perfect, the eight-level holistic scale categories (0-7) functioned within acceptable parameters for the study. With the exception of categories 0 (n = 0)

and 1 (n = 2), the relative frequency of each category had a minimum of 10. Since the 0 and 1 categories are typically given to students with little or no English ability, it is not surprising that few students would have such low ability after 14 weeks of intensive English training. The average measures for each category increased monotonically without exception, as did the threshold estimates. The threshold estimates had the minimum recommendation of 1.4 logits between each category indicating that each category showed distinction, however some of the thresholds were over 5 logits apart (e.g. category 2) indicating that some information could have been lost and perhaps the category needed to be split. Furthermore, for the scale to be treated as interval data, it would be more desirable for the spacing of the thresholds to be more regularly spaced (see Figure 1). An examination of the category probability distributions was indicative that each category functioned well (see Table 8), and none of the outfit mean squares exceeded 2.0. Based on this, the conclusion was that the category descriptions of the scale functioned, and there was no need to make adjustments to the categories.

Reliability analysis. One advantage of a Many Facets Rasch Measurement analysis is that the facets can be compared on a vertical scale that shows the link between the measurement scale and the facets. Figure 2 shows the logit in the first column, the examinee ability level in the second column, the rater severity in the third column and the scale equivalency in the fourth column. The 0 in the middle of the vertical scale is tied to the mean of the examinee ability estimates or logits. An examinee with an ability logit of 0 (the second column) would have a 50% chance of being rated in category 4 (the fourth column), by raters R12, R13 or R15 (the third column).

Table 8

Human-rated Holistic Speaking Level Rating Scale Category Statistics

Category	*Absolute Frequency*	*Relative Frequency*	*Average Measure*	*Outfit*	*Threshold*	*SE*
0	0	0%	NA	NA	NA	NA
1	2	<1%	-11.50	1.6		
2	55	11%	-8.18	0.8	-13.85	0.77
3	130	27%	-1.95	1.0	-5.52	0.27
4	156	32%	1.39	0.9	-0.30	0.16
5	89	18%	3.95	1.0	3.25	0.17
6	49	10%	6.69	1.0	5.86	0.23
7	7	1%	9.37	1.5	10.56	0.49

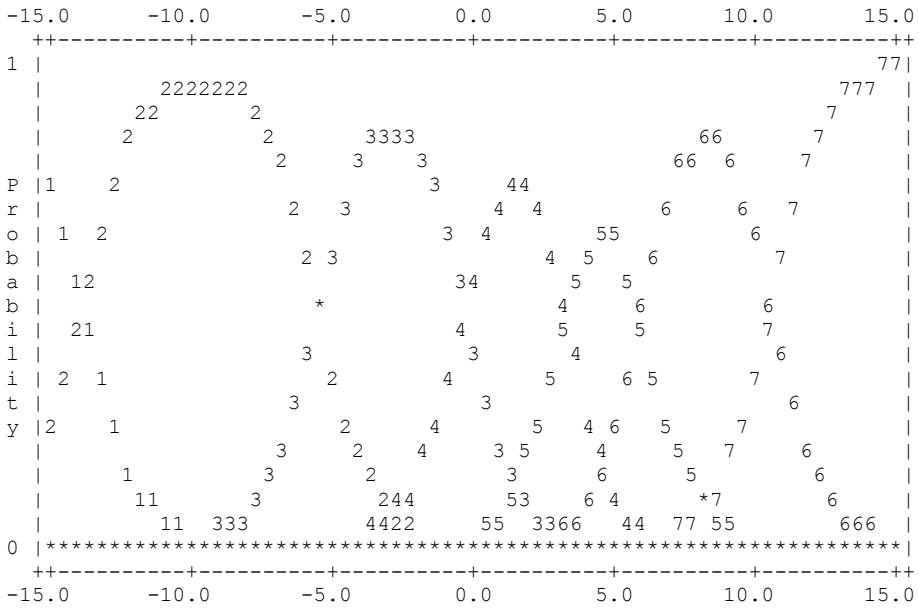

Figure 1. Human-rated Holistic Speaking Level Rating Category Distribution

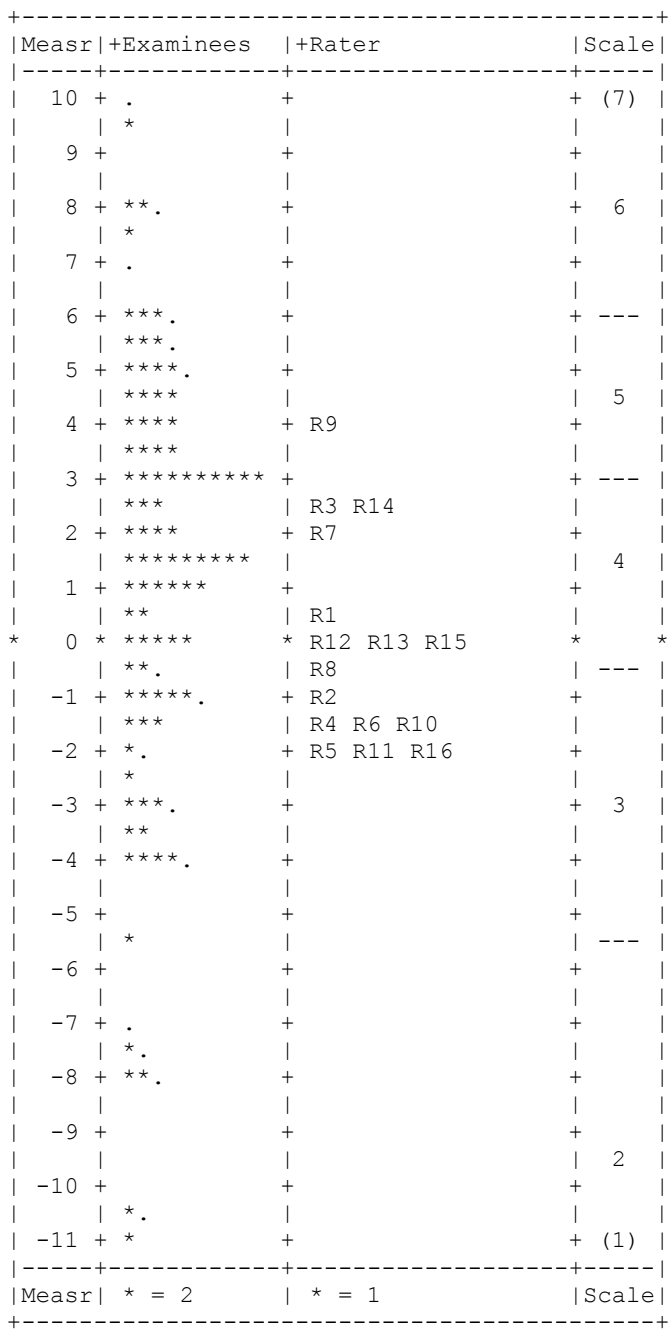

Figure 2. Human-rated Holistic Speaking Level Vertical Scale

From Figure 2 we can see that the examinee abilities ranged from category 1 to 7 on the scale. The analysis found that the separation reliability between the examinees was .86, indicating that we can be confident that estimated person ability parameters are indicative of reliable differences in the examinees' abilities.

The analysis of the raters produced a separation reliability of .95 that is indicative that the raters judged the examinees with enough varying degrees of severity that an examinee's unadjusted rating would be biased depending who rated him or her. From Figure 2, we can see that rater R9 was the most generous and raters R5, R11, and R16 were the most severe. The fit statistics do however suggest that even though the raters had differences in their rating, they were internally consistent (the fit statistics were close to 1.0 with a range between .5 and 2.0 which is considered acceptable in most cases). The average of the outfit mean squares was 1.09 and the average of the infit mean squares was .98. Thus, despite the varying degrees of severity, the ratings can be used if the fair average values are used to compensate for any differences.

The Rasch analysis found that the scale categories functioned within recommended guidelines and that the separation reliability of the facets were indicative of different levels. The person fair averages of human-rated holistic speaking level were deemed adequate measures for the purpose of answering the first research question.

Phase 2: Statistical analysis of human-rated and ASR scored speaking tests. Prior to performing the regression, Pearson Product Moment correlation coefficients were computed to describe the relationship between the human-rated holistic speaking level and each of the ASR-scored fluency features of (a) total response time, (b) speech time, (c) speech time ratio, (d) number of syllables, (e) speech rate, (f) articulation rate, (g) mean syllables per run, (h) silent pause time, (i) number of silent pauses, (j) mean silent pause time, and (k) silent pause ratio (see

Table 9). The correlations between the human-rated holistic speaking level and each of the ASR fluency features were moderate with almost all of them being statistically significant ($p < .05$), the exception being the correlation between the human-rated holistic speaking level and the number of silent pauses ($p = .31$). It should be noted that even though some of the relationships were statistically significant, for the correlations lower than ±.30, the effect size was small (Larson-Hall, 2010).

While the study did not explicitly examine the relationship of the ASR fluency variables with each other, there are a few things of interest. First, the total response time had very few significant correlations. On further investigation, it was discovered that the computer programming of the test did not let examinees end the recording time early so that each total response between the examinees was constant. Since total response time was used to calculate speech time ratio, speech rate and silent pause ratio, it had near perfect correlations with the other variables. Those derived variables did vary between examinees as they are ratios and were able to be used in the analysis. However, total response time was excluded from the analysis and was not used in the regression equation.

There were two other perfect correlations but they had inverse relationships: speech time and silent pause time and speech time and silent pause ratio. The speech time was the total time in which speaking occurred while the silent pause time indicated the total time it was silent and the silent pause ratio was the decimal percent of silence. By definition, then the relationship between Speech Time and these two variables would produce a perfectly inverse relationship.

Table 9

Correlations between Human-rated Holistic Speaking Level and ASR Timing Features

	Speech Time	Speech Time Ratio	Number of Syllables	Speech Rate	Articulation Rate	Mean Syllables per Run	Silent Pause Time	Number of Silent Pauses	Mean Silent Pause Time	Silent Pause Ratio
Human-rated Holistic Speaking Level	.35	.51	.51	.52	.43	.17	-.36	-.07	-.43	-.35
Speech Time		.79	.79	.79	.16	.69	-1.00	-.76	-.66	-1.00
Speech Time Ratio			.99	.99	.72	.58	-.79	-.49	-.65	-.79
Number of Syllables				1.00	.72	.58	-.79	-.48	-.67	-.79
Speech Rate					.72	.58	-.79	-.48	-.67	-.79
Articulation Rate						.14	-.16	.09	-.36	-.16
Mean Syllables per Run							-.69	-.75	-.28	-.69
Silent Pause Time								.76	.66	1.00
Number of Silent Pauses									.07	.76
Mean Silent Pause Time										.66

Note: correlations greater than ± .14 are significant at $p < .05$ (2-tailed) and correlations ± .28 are significant at $p < .01$ (2-tailed)

To determine which combination of features best predicted human-rated holistic speaking level, a sequential linear regression was run. The sequence of the ASR variables was based on previous work (Ginther et al., 2010). The variables that were excluded from the model based on statistical significance were: silent pause time and number of syllables, but the remaining variables showed high levels of multicollinearity. The variance inflation factor (VIF) should be less than 5, but in the initial regression, the only variable that met that standard was mean syllables per run (VIF = 2.33) with the remaining variables having VIFs that were higher ((a) VIF = 25.65 for articulation rate, (b) VIF = 188.45 for speech rate, (c) VIF = 183.63 for speech time ratio, (d) VIF = 6,099.15 for speech time, and (e) VIF = 6,050.36 for silent pause ratio). Through an iterative process, the variables of speech time, speech time ratio, and articulation rate were subsequently omitted to obtain a model that had VIFs lower than 5. For the final analysis, the regression was again with the following three variables: speech rate, mean syllables per run, and number of silent pauses (See Table 10).

Table 10

Multiple Regression Table Predicting Human-rated Holistic Speaking Level

Model	R	Total R^2	ΔR^2	Speech Rate B	Mean Syllables per Run B	Number of Silent Pauses B	Silent Pause Ratio B
1	0.52*	0.27*		0.017 (.013, .021)			
2	0.54*	0.29*	0.025	0.021 (.016, .026)	-0.03 (-.053, -.007)		
3	0.55*	0.31*	0.016	0.021 (.016, .026)	-0.009 (-.039, .022)	0.003 (.000, .006)	

*p < .05, Intercept for Model 3 = .38 (-.67, 1.4) with the three predictors Speech Rate, Mean Syllables per Run and Number of Silent Pauses, the regression equation is statistically significant, F(3,194) =28.7, p<.001.

Tests for multicollinearity indicated a very low level multicollinearity with these for variables ((a) VIF = 3.19 for speech rate, (b) VIF = 2.70 for mean syllables per run, and (c) VIF = 3.74 for number of silent pauses). The speech rate accounted for 27% of the variance. When mean syllables per run was added, an additional 2% of the variance was accounted for. The number of silent pauses added another 2%, but silent pause ratio did not significantly add more information. Table 10 reports the obtained values of R, R^2, change in R^2, the unstandardized regression coefficients (B), and the 95% CIs. The resulting prediction formula included the three variables, speech rate, mean syllables per run and number of silent pauses, which accounted for 31% of the variance.

In summary, to answer the first research questions, the combination of the ASR timing features that best predicts human-rated holistic speaking scores are speech rate, mean syllables per run and number of silent pauses. The regression equation that can be used to predict the speaking score is y = .38 + .021 (speech rate) + -.009 (mean syllables per run) +.003 (number of silent pauses). The variables that were eliminated due to multicollinearity were speech time ratio, articulation rate, silent pause time, mean silent pause time and silent pause ratio. Finally the proportion of variability explained by the model was 31%.

Research Question 2: Impact of Prompt Difficulty

The second research question asked to what extent did the rater predicted difficulty of the speech prompts align as expected with (a) the analytic item level human ratings and (b) the empirical ASR scoring for each prompt? Two different Rasch analyses were conducted to answer this research question. The first examined the analytic item level speaking ratings obtained from the human raters. The second examined the ASR scoring at the prompt level obtained from the regression prediction equation derived from question 1 of this study.

Phase 1: Rasch analysis of analytic human-rated item speaking level. The analytic human-rated item speaking level was based on the same 8-level rubric that the holistic raters employed in the holistic rating to calculate the human-rated holistic speaking level. The only difference in its use was that it was applied to each prompt of the assessment rather than the overall score.

Scale diagnosis. The eight categories of the analytic human-rated item speaking level functioned within acceptable parameters for the study (Table 11). With the exception of the category 0 (n = 3), the relative frequency of each category had a minimum of 10 in each category. The average measure increased monotonically from 1 to 7 without exception, as did the threshold estimates. The threshold estimates had the minimum recommendation of 1.4 logits between each category indicating that each category showed distinction, and none of the thresholds were over 5 logits apart, and the spacing of the categories was more evenly spaced than the human-rated holistic speaking level. An examination of the category probability distributions was indicative that each category functioned well (see Figure 3).

For the fit statistics, the outfit mean squares of the categories did not exceed 2.0 with the exception of the 0 category which only had 3 responses. The only category that did not fit the guidelines of a good scale was 0. Since the 0 category is typically reserved for little or no production and since the students had one semester of instruction, this category could be combined with category 1 if it were only used as an end of instruction scale. However, since the scale is used for placement testing as well, all eight categories were retained. The analytic human-rated item speaking level scale functioned within acceptable parameters to be used in the analysis.

Table 11

Analytic Human-rated Item Speaking Level Rating Scale Category Statistics

Category	*Absolute Frequency*	*Relative Frequency*	*Average Measure*	*Outfit*	*Threshold*	*SE*
0	3	0%	-3.04	2.0		
1	54	2%	-3.72*	1.1	-7.13	0.59
2	345	12%	-2.39	1.0	-5.06	0.77
3	933	33%	-.23	1.0	-2.27	0.27
4	916	32%	1.62	.9	.77	0.16
5	449	16%	2.91	1.0	3.00	0.17
6	142	5%	3.88	1.1	4.55	0.23
7	24	1%	4.93	1.0	6.14	0.49

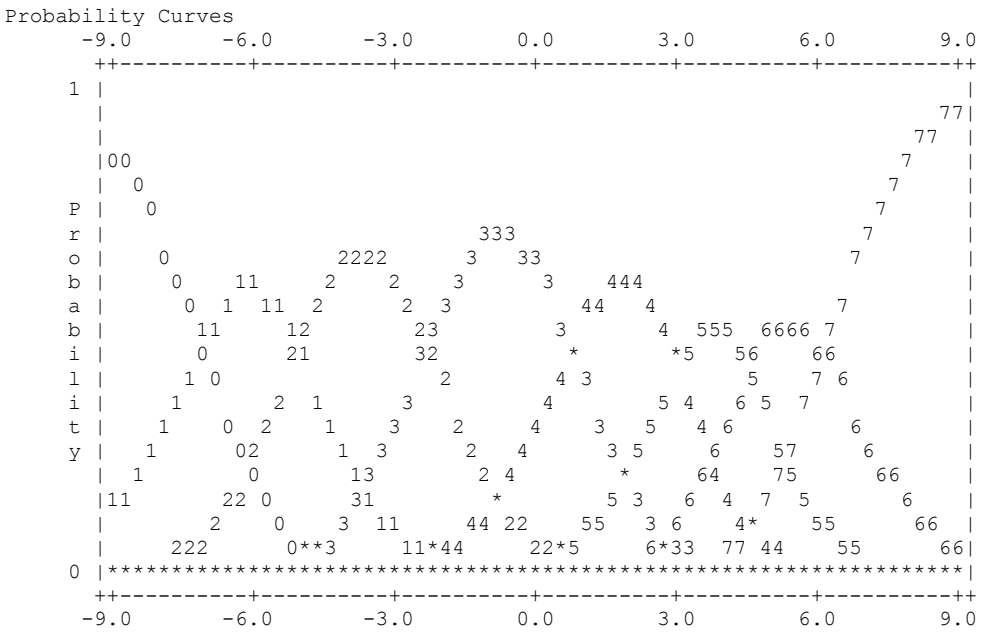

Figure 3. Human-rated Item Speaking Level Rating Category Distribution

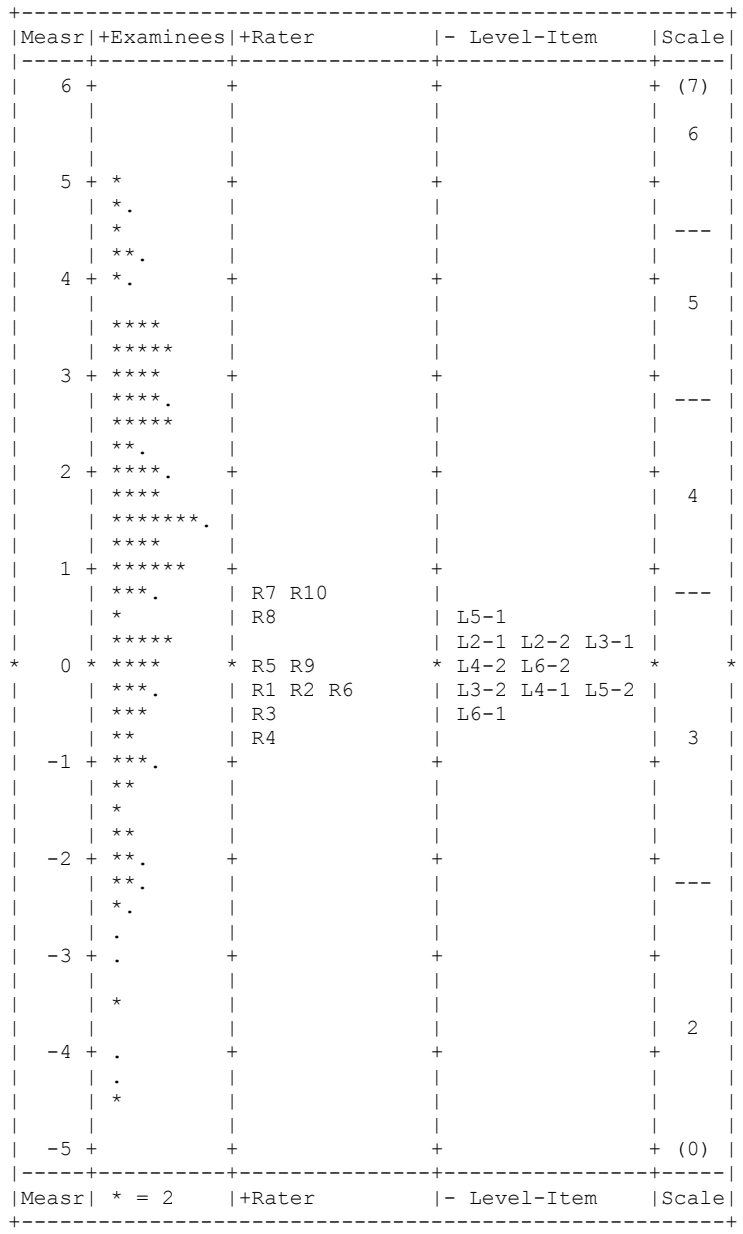

Figure 4. Analytic Human-rated Item Speaking Level Vertical Scale

Reliability analysis. The reliability statistics on item level scoring found that all three facets were reliably separated. Figure 4 is a vertical scale map similar to the map in Figure 2 with the exception that the third column shows the item difficulties with prompt being labeled with the intended level and the item. So L3-1 represents the prompt that was intended to elicit Level 3 speech and it was Item 1. Please note that the mean of the items and raters is centered at 0 logits. Figure 4 showed that the examinee ability ranged from category 2 to 6. This had a more restricted range than the human-rated holistic levels, however with the examinees separation reliability of .94, we can be even more confident of the different ability levels of the examinees.

The analytic human-rated item speaking level raters judged more homogenously than the human-rated holistic speaking level raters with the analytic human-rated item speaking level standard deviation = .48 as opposed to the human-rated holistic speaking level standard deviation = 1.82. Even with the smaller standard deviation, the analytic human-rated item speaking level had a reliability coefficient of .96 indicating that the raters could not be used interchangeably. In Figure 4, we can see that the raters R7 and R10 were the most generous and rater 4 was the most severe. As with human-rated holistic ratings, the fit statistics were indicative of high internal consistency with an average mean outfit square of 1.0 and an average mean infit square of 1.0.

The item facet had a reliability of .89 indicating that items could be not used interchangeably without compensating for their difficulty level. In Figure 4, the third column represents the intended level and item number. The easiest item was L6-1 (i.e. Intended Level 6, Item 1) and the most difficult item was L5-1. While it was expected that the prompts would have varying item difficulties and that some kind of item equating would need occur to create equivalent test forms, it was unexpected that the item difficulty means did not order in their

intended levels. It was also notable that the prompts clustered around category 3 (SD = .27) and had a narrower range than the raters (SD = .48). The prompt fit statistics were indicative of high internal consistency with an average mean outfit square of 1.0 and an average mean infit square of 1.0. While the prompt alignment was not as expected, the separation reliability was deemed high enough to use in the analysis to see if the human-rated empirical difficulties would align with the ASR item difficulties. The analytic human-rated item speaking level item MFRM fair averages were therefore used to answer the second research question.

Phase 2: Rasch analysis of ASR regression predicted analytic speaking level. To compute prompt difficulty, each prompt of each test of every examinee was processed using the PRAAT ASR software. The ASR regression predicted analytic speaking level was calculated by taking the three ASR features found to account for the most variance (speech rate, mean syllables per run, and number of silent pauses) and applying the regression equation determined in Research Question 1 to score each prompt as if a human had rated it. Since the ASR was able to rate all prompts of all students, there were only two parameters: items and persons.

Scale diagnosis. The regression predicted analytic speaking level eight-level holistic scale (0-7) did not function as expected (see Table 12). Only five of the eight levels had a minimum of 10 in each category. There were no instances of any of categories 6 or 7. Category 0 only had nine responses and from the average measures, should have been combined with Category 1. The average measure of categories 1 to 5 increased monotonically as did the threshold estimates.

The threshold estimates satisfied the minimum recommendation having at least 1.4 logits between each category indicating that each category showed distinction, but category 4 had a range greater than 5 logits indicating that that category could lack distinction between examinees

of different ability levels. This could be problematic because examinees with a wide range of ability levels would all be awarded the same category rating. Unfortunately there is little that can be done to remedy a category that is too broad. An examination of the category probability distributions revealed that the categories 2 through 5 were distinct (see Figure 5) and the outfit mean squares did not exceed 2.0.

To improve the functionality of a scale, categories are collapsed. It was not possible in this case. Furthermore the failure of the scale to use all of the categories made it questionable to use it as a future rating scale for ASR scoring. This was likely because the regression equation upon which the scale was based only accounted for 31% of the variance. However, the scale was used for this study since these are likely to be the categories that would be used in ASR scoring of spontaneous speaking tests.

Reliability analysis. The reliability separation analysis found that both examinees and ASR rated items were reliably separate. Figure 6 is a vertical scale similar to the other ones presented with the exception that it shows only the examinees and the items and since it was generated from Winsteps does not have the category scale column. Since the ASR rated all of the items, there is no rater column.

In the first column of Figure 6, we see the examinees have a range of logit values from just under -4 to above 6. The separation reliability between the examinees was .90, indicating that this scoring method could reliably separate test takers into different ability groups. At first glance this would appear promising as it means that some aspect of the timing fluency construct reliably separates examinees.

Table 12

Regression Predicted Analytic Speaking Level Rating Scale Category Statistics

Category	*Absolute Frequency*	*Relative Frequency*	*Average Measure*	*Outfit*	*Threshold*	*SE*
0	9	0%	-2.35	2.98		
1	31	2%	-2.52	1.07	-4.26	.38
2	190	10%	-1.08	1.13	-3.61	.19
3	758	38%	.84	.86	-1.55	.09
4	929	47%	3.62	.93	2.01	.06
5	79	4%	6.23	.91	7.41	.13
6	0	0%	NA	NA	NA	NA
7	0	0%	NA	NA	NA	NA

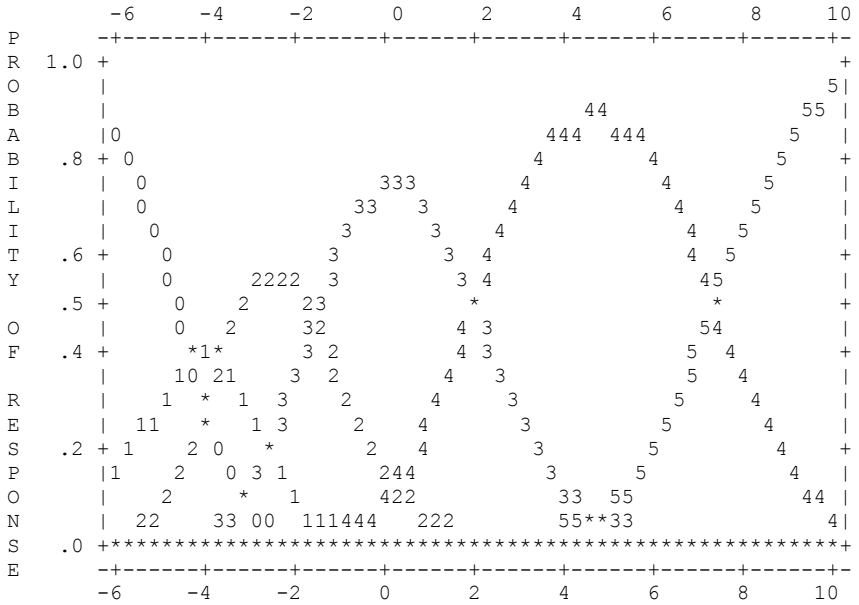

Figure 5. Regression Predicted Analytic Speaking Level Rating Category Distribution

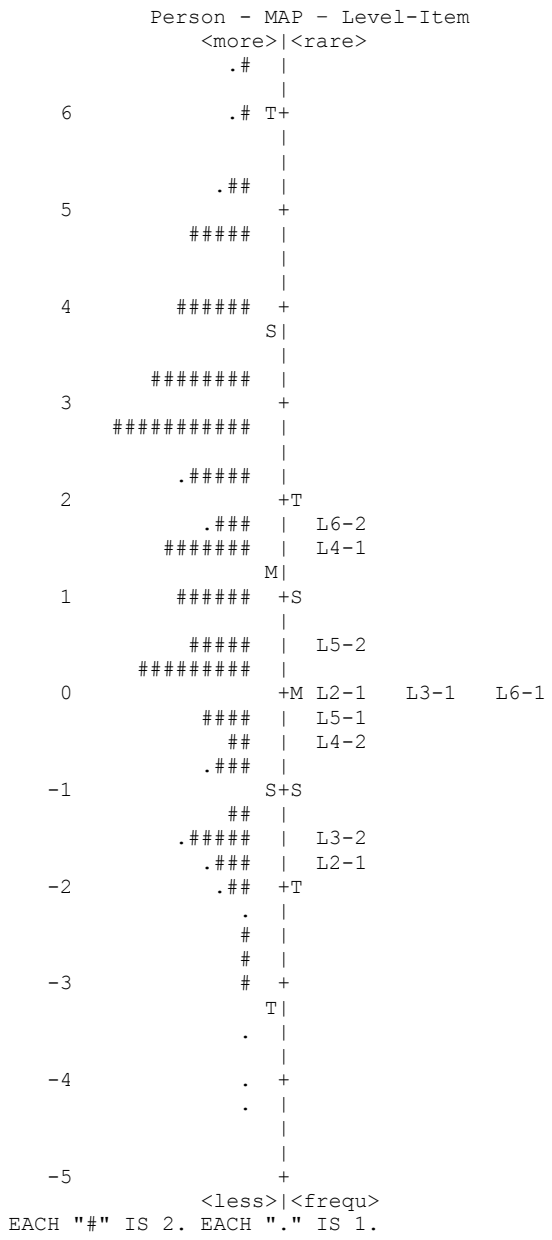

Figure 6. ASR Regression Predicted Analytic Speaking Level Vertical Scale

However, simply dividing examinees reliably into different groups is insufficient unless those groups would have those same divisions if humans rated them. Since the focus of this study is on the prompts and their item difficulties, no further discussion of this will be presented though it is an interesting area of research.

The item reliability was .98 indicating that the items had distinctly different difficulty levels and could not be used interchangeably (see Figure 6). The most difficult item was L6-2 and the easiest item was L2-1, and while that initially appears to align with the rater predicted difficulties, many of the other items do not align.

This misalignment once again confirms the study rationale that for equated tests to be created, some kind of prompt statistics needed to be established in order to make equivalent forms of the assessment. The regression predicted analytic speaking level item logits were used to answer the second research question. However, in the initial comparison with the expert rater predicted difficulties and analytic human-rated item speaking level, the fair average equivalents were calculated.

Rasch analyses were conducted on the analytic human-rated item speaking levels and the ASR regression predicted analytic speaking levels. The scale categories of the human ratings were found to function within recommended guidelines; however, the scale categories of the ASR regression equation were problematic. Despite the problems, the scale was kept, as it would most likely be used if an ASR were to score at the item level. Both analyses found that the prompts had reliably different item difficulty parameters and could be used in comparison with the expert rater predicted difficulties.

Phase 3: Statistical analysis of human-rated and ASR-scored prompt difficulty. To answer the second question and determine the extent to which the rater predicted difficulty were

aligned with the analytic human-rated item speaking level and regression predicted analytic speaking level difficulty levels, the item difficulty statistics for the analytic human-rated item speaking level and regression predicted analytic speaking level were calculated. The scale ranged from 0 to 7, and the averages of all three measures were in the middle of the scale with means ranging from 3.49 to 3.66 (see Table 13).

Table 13

Comparison of Speaking Level Item Statistics

Item	Rater Predicted Difficulty	Analytic Human-rated Item Speaking Level Fair Average	Analytic Human-rated Item Speaking Level Logit	Regression Predicted Analytic Speaking Level Fair Average	Regression Predicted Analytic Speaking Level Logit
L2-1	1.86	3.54	0.29	3.21	1.07
L2-2	2.00	3.58	0.20	3.27	0.84
L3-1	2.71	3.52	0.36	3.92	-1.94
L3-2	2.71	3.79	-0.30	3.82	-1.31
L4-1	3.13	3.74	-0.19	3.71	-0.77
L4-2	3.38	3.67	-0.03	3.48	0.1
L5-1	4.63	3.5	0.39	3.37	0.51
L5-2	4.75	3.75	-0.22	3.53	-0.06
L6-1	5.00	3.82	-0.38	3.38	0.46
L6-2	5.50	3.71	-0.11	3.2	1.12
Mean	3.57	3.66	0.00	3.49	0.002
SD	0.71	0.11	0.27	.24	1.04

The spread of the three measures was quite different. The rater predicted difficulty was the most disparate (SD = 0.71) while the other two measures were much more homogenous. As shown in Figure 7, the analytic human-rated item speaking level had the smallest standard deviation (SD = 0.11) and regression predicted analytic speaking level slightly larger (SD = .24).

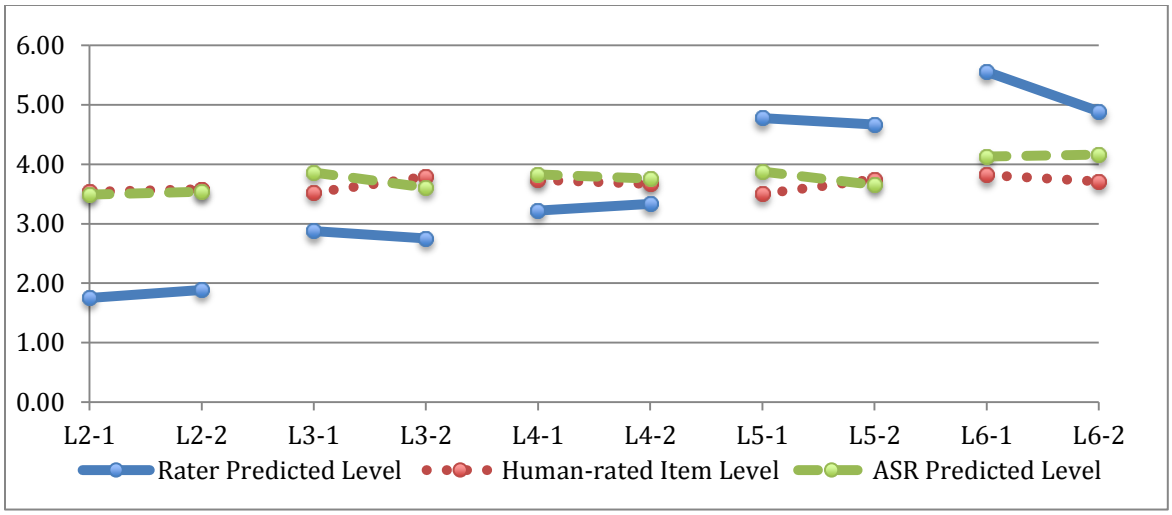

Figure 7. Means of Item Difficulty Measures

A Pearson Product moment correlation was computed between all of the possible pairs of item measures and none of them were found to be significant (see Table 14). What was most surprising was that human rated analytic human-rated item speaking level and the human predicted difficulty levels had an inverse relationship.

Table 14

Correlations between Item Difficulty Measures

	Analytic Human-rated Item Speaking Level	Regression Predicted Analytic Speaking Level
Rater Predicted Difficulty	-.43	.25
Analytic Human-rated Item Speaking Level		.09

To answer the second research question, therefore, neither of the analytic scoring methods aligned as expected with the rater predicted difficulty levels. None of the relationships between the variables were significant and the human-rated measures even had an inverse relationship.

Post-Study Question: Use of an At-Level Scale to Rate Items

These results raise an interesting question that was not initially part of the study. Why was the alignment of the rater predicted difficulties so different from the human-rated analytic human-rated item speaking level? There are a number of possible explanations as to why the rater predicted difficulties did not align with the analytic human-rated item speaking level. The first could be that the descriptors in the scale upon which the items were written were flawed. While possible, the scale was based on the well-established ACTFL scoring rubric that has been in use for over 30 years. Second, there is the possibility that the items did not adequately reflect the scales' descriptors. The raters that evaluated the prompts to determine their intended difficulty levels had a minimum of 3 semesters of rating experience with the average number of semesters being 4.75 semesters. These raters felt that the items did align with the rubric they used for rating.

Another possibility is likely the existence of a pervasive restricted range error in using a holistic rubric to rate an item. When an item is targeted at Level 6, and the rater knows it is targeted at Level 6, that rater might be hesitant to give scores on the lower end of the scale (0, 1, and 2) even if the respondent language is characteristic of those levels. Similarly an item targeted at Level 2 might result in ratings that are not in the higher part of the range (5, 6, and 7) because the prompt did not elicit language in that upper range. This range restriction in scoring could have resulted in fair item averages that clustered close to the mean of all the items. One

piece of evidence of this possibility is the fact that analytic human-rated item speaking level had the smallest fair average standard deviation (see Table 13) of all the scoring methods.

There are a number of reasons that could cause this phenomenon. First, there could be the rater bias classified as central tendency error. In this situation, raters fail to use the extremes of a the scale either due to the inability to discriminate between good and bad performance or a desire hedge their rating by being moderate (Myford & Wolfe, 2003). Another possibility is that there is a halo effect in which one part of the rubric (e.g. pronunciation) is causing raters to compensate for weak performance in another part of the rubric. So when a prompt that is targeted at Level 2 gets a rating that is substantially higher than a 2 (e.g. 4 or higher), it could be an indication that the examinee did an outstanding response at Level 2 in terms of accuracy (e.g. pronunciation and fluency), but there could be no evidence that the text type or content was at that higher level. The converse could also be true—a lower rating than the level that the prompt was targeting could indicate failure of the prompt but not evidence that the examinee could perform language based on the description of the rubric.

Phase 1: At-level scale Rasch analysis. To compensate for the possibility that a restricted range bias impacted the scores, the human rated raw data were recoded to a five-point scale that will be referred to as the at-level scale. Since the raters knew the intended level of the prompt they were rating, the scale reflected whether the student response was below the targeted prompt level, at level or above level. Table 15 shows how each intended level's 8-point rubric was converted to the 5-point at-level scale.

Table 15

Speaking Rubric to At-Level Scale Conversion Matrix

	At-level Rating	Intended Item Difficulty Level				
		2	3	4	5	6
Below by 2 or more levels	1	0	0 1	0 1 2	0 1 2 3	0 1 2 3 4
Below by 1 level	2	1	2	3	4	5
At level	3	2	3	4	5	6
Above by 1 level	4	3	4	5	6	7
Above by 2 or more levels	5	4 5 6 7	5 6 7	6 7	7	

Scale diagnosis. An additional FACETs analysis was conducted with the recoded data to evaluate whether the item measures would more closely align with the rater predicted difficulty. The at-level scale functioned within the parameters needed for a reliable scale (see Table 16). The relative frequency of each category had a minimum of 10 in each category. The average measure increased monotonically without exception, as did the threshold estimates. The threshold estimates had the minimum recommendation of 1.4 logits between each category indicating that each category showed distinction, and none of the thresholds were over 5 logits apart. Furthermore, the spacing between the thresholds was the most evenly-spaced of all the scales used in the study (see Figure 8). The outfit mean squares of the other categories did not exceed 2.0.

Table 16

At-Level Scale Rating Scale Category Statistics

Category	*Absolute Frequency*	*Relative Frequency*	*Average Measure*	*Outfit*	*Threshold*	*SE*
1	770	27%	-6.14	1.0		
2	553	19%	-2.52	.6	-3.86	.08
3	589	21%	-.06	1.0	-1.39	.07
4	552	19%	2.38	1.2	1.19	.07
5	402	14%	5.22	1.1	4.07	.09

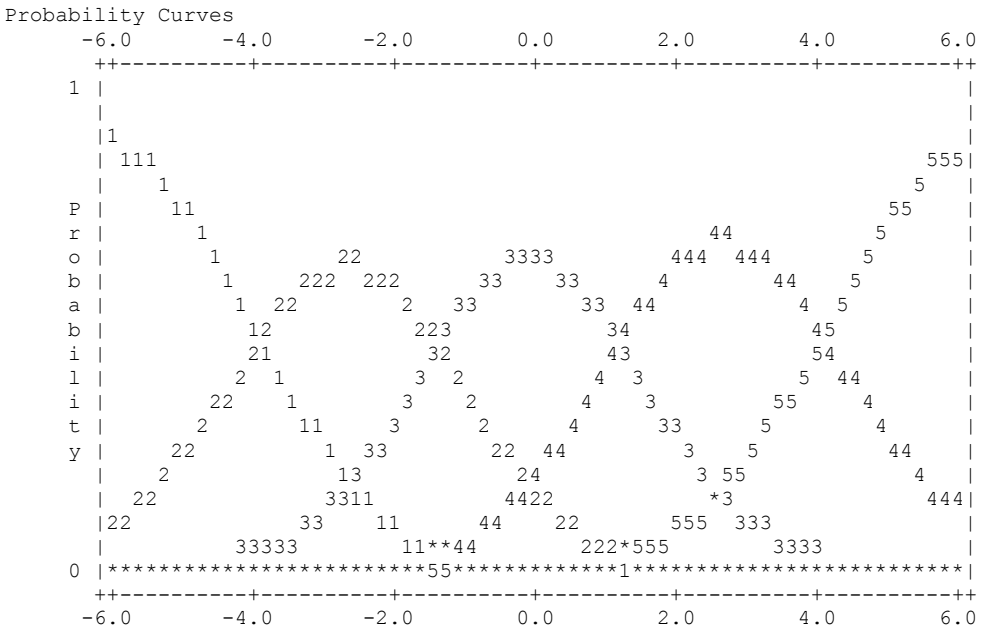

Figure 8. At-Level Scale Rating Category Distribution

Reliability analysis. The reliability statistics on the at-level item scoring found that all three facets (examinees, raters, and items) were reliably separated. In Figure 9, we can see that the examinees have a range from categories 1 to 5. Note that the significance of these categories did ***not*** signify the levels of the speaking rubric facets of examinee, but rather whether how well they performed the task at its intended level. This analysis found that the separation reliability between the examinees was .93 and that the examinees could be separated reliably into different groups.

The raters had a reliability of .96 the same as the analytic human-rated item speaking level analysis with the standard deviation being slightly larger than the at-level scale analysis (analytic human-rated item speaking level SD = 0.55 compared to at-level scale SD =0.49). The raters still exhibited different levels of severity with R7 being the most generous and raters R3 and R4 being the most severe. Comparing the raters in Figure 4 and Figure 9 we see that the ordering of the severity and generosity of the raters is very similar with a high correlation ($r = .78$, $p < .05$) between the at-level rater and analytic human-raters. The fit statistics were indicative of high internal consistency with an average mean outfit square of 1.0 and an average mean infit square of 1.0.

Most noteworthy though was the fact that the at-level scale item facet jumped to a reliability of 1.00 indicating that it would be virtually impossible to have the same score with the different items. Furthermore, the differences in difficulty aligned closely with the raters predicted level (see Figure 9). The prompts that were intended to elicit Level 6 language (L6-1 and L6-2) were the most difficult while the prompts intended to elicit Level 2 language were the easiest (L2-1 and L2-2). Table 17 illustrates the manner in which the item difficulty parameters increase monotonically as the intended difficulty increased.

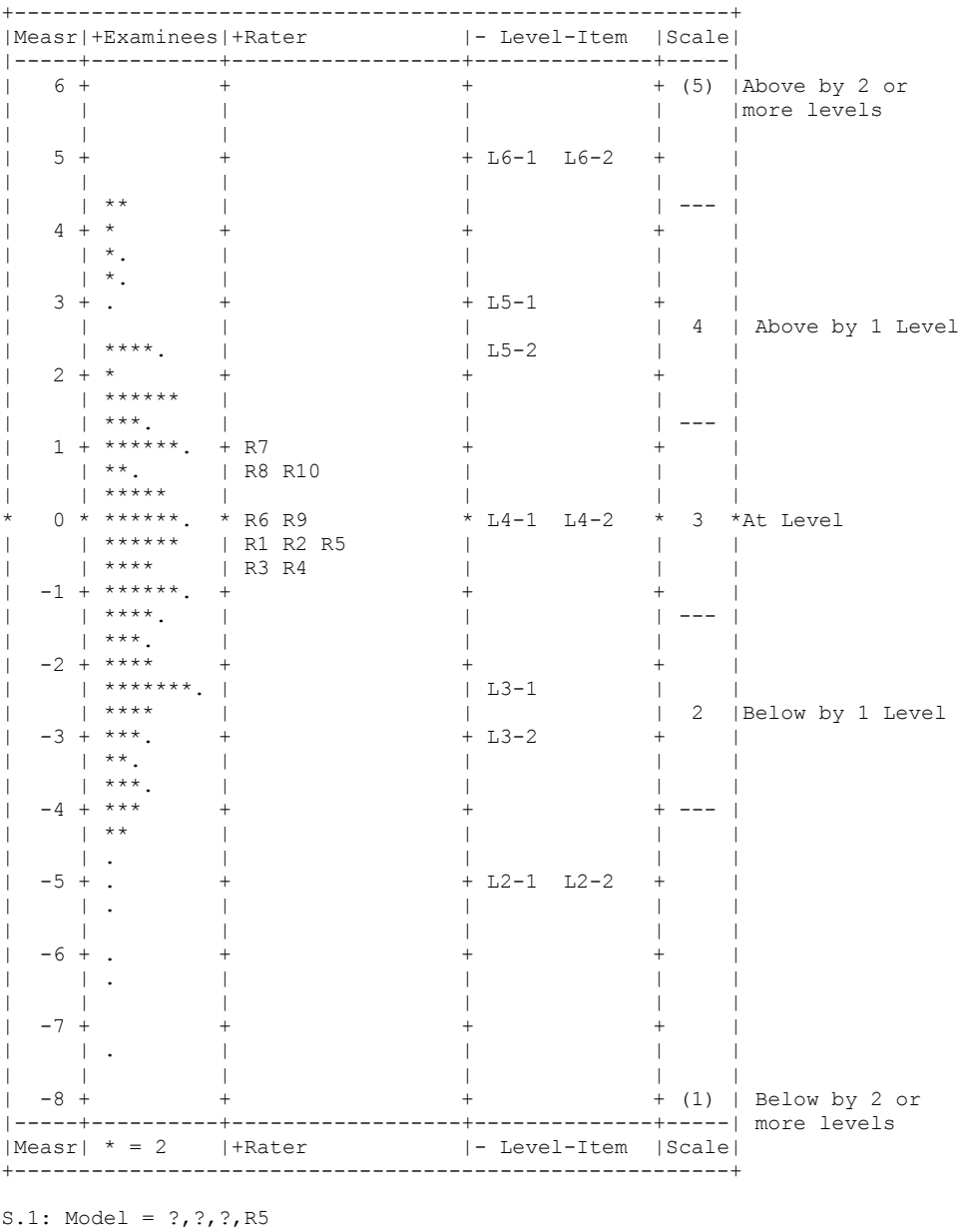

Figure 9. At-Level Vertical Scale of Examinees, Raters, and Items

Another benefit of this scale was that it gave information on prompts that were intended to elicit language at the same level. For example, the prompts at Levels 2, 4 and 6 could be used interchangeably with the other prompts at those intended levels because their item difficulty parameters had comparable values.

The prompts at Levels 3 and 5 however were not comparable. Item L5-1 was more difficult than Item L5-2 and similarly item L3-1 was more difficult than Item L3-2. If there were more prompts, then test developers could chose those that would create equivalent test. The prompt fit statistics were indicative of high internal consistency with an average mean outfit square of 1.0 and an average mean infit square of 1.0.

Table 17

At-Level Scale Item Statistics in Order of Measure

Item	Fair Average	Logit
L2-1	4.49	-4.86
L2-2	4.54	-5.04
L3-1	3.61	-2.22
L3-2	3.86	-2.93
L4-1	2.83	-.15
L4-2	2.76	.02
L5-1	1.65	2.92
L5-2	1.88	2.27
L6-1	1.16	4.88
L6-2	1.13	5.10
Mean	2.79	0.00
SD	1.30	3.74

The Rasch analysis found that the at-level scale functioned most closely to the recommended guidelines and that the separation reliability of the facets were indicative of different levels. Most notable was that the item facet had the highest separation reliability with this method. The at-level item difficulty fair averages were then compared to the other item difficulty parameters calculated in the study.

Phase 2: Statistical analysis of human-rated and ASR-scored prompts. A Pearson Product moment correlation was run between all of the item measures (see Table 18): the rater predicted difficulty, the analytic human-rated item speaking level, the regression predicted analytic speaking level, and the at-level scale. The highest correlation was between the at-level scale and the rater predicted difficulty ($r = .98$, $p < .001$). The at-level scale still did not have a significant relationship with the ASR regression predicted analytic speaking level so the original findings with research question 2 remain the same.

Table 18

Post-Study Correlations between Item Difficulty Measures

	At-Level Scale	Analytic Human-rated Item Speaking Level	Regression Predicted Analytic Speaking Level
Rater Predicted Difficulty	.98	-0.43	0.25
At-Level Scale		-0.42	0.23
Analytic Human-rated Item Speaking Level			0.09

Note: correlations greater ±.77 are significant at $p<.01$ (2-tailed)

Thus to answer the supplementary research question that evolved from the research study, the holistic scale does not drill down well to the item level. A modified scale that has a smaller range that is related to the original rubric yet operationalized on the item level had great potential as a human scoring method on the item level.

Summary of Results

The ASR timing features of speech rate, mean syllables per run, and number of silent pauses were the best predictors human-rated speaking tests, but these features only accounted for 31% of the variance. The correlations between (a) ASR-scored item difficulties, (b) rater

predicted item difficulties, and (c) analytic human-rated item speaking levels were not significant. A post-study analysis found that changing the analytic human-rated item speaking levels from a holistic scale to an at-level scale resulted in a better and more accurate correlation with the rater predicted difficulties.

Chapter 5

Discussion and Conclusions

This study evaluated the ability of ASR-scored speaking tests to predict what ratings human judges would have assigned the tests. Furthermore, the impact of prompt difficulty on the two scoring processes (human and ASR) was evaluated. While not initially part of the study, the impact of having human raters use a holistic rating rubric to rate responses at the item level was analyzed.

Review of Findings

The first question explored the extent to which the ASR scoring of timing features could predict human speaking ratings. While the relationships between the human-rated holistic speaking level and the ASR timing features were statistically significant (see Table 9), they were moderate. The regression equation found that only 31% of the variance in the speaking scores was accounted for by the speech rate, mean syllables per run and number of silent pauses (see Table 10). So, while many of the timing fluency features were related to overall speaking ability, they would likely be insufficient to predict overall speaking ability consistently and accurately. This finding is similar to what other researchers have found (Zechner et al., 2009)—that ASR-scored tests may be suitable for low stakes situations but are not likely a sufficient replacement for the human rating of speaking in high stakes situations involving important decisions.

The second research question this study addressed explored how the intended prompt difficulties aligned with the ASR scoring and the human rating. Neither the human rating method of using a holistic scale at the prompt level nor using the ASR regression equation method of scoring aligned with the intended difficulty levels (see Table 14). The failure of the ASR functioning might have been expected because the equation upon which it was based only

accounted for 31% of the variance. Or to restate it, 69% of the variance is due to other factors besides the ASR timing features.

The greater concern was the relationship between the intended difficulty levels and the human-rated item level statistics. First, the item difficulty statistics had very little variance (SD = .27). In fact, Figure 4 illustrated that the difference in the raters was greater than that of the items (SD = .48). Furthermore, the correlation between the rater predicted prompt difficulties and the human-rated item statistics were not statistically significant and inversely correlated (r = -.43). The incongruence of trained raters (a) being able to predict differences in prompt difficulty yet (b) being unable to find performance differences from the prompts led to an investigation of the rating approach.

The most surprising finding was that the human-item-level ratings did not align with the rater predicted difficulty. Using the holistic 8-level scale on each item proved to be problematic. The holistic rating scale included a full range of language possibilities from simple sentences on familiar topics (Level 2) to extensive, complex speech on abstract academic topics (Level 6), but each of the individual prompts was aimed at only one of those levels. In writing a speaking proficiency test, an item writer would try to elicit a specific proficiency level of speech in the construction of the prompt. Consider the following prompts: (1) Describe your house and the neighborhood you live in, and (2) What is the impact of government subsidized housing on the quality of life in a neighborhood?

In the first prompt, the intent is to elicit speech at the ACTFL Intermediate level, whereas in the second prompt, the intent is to elicit speech at the Superior level. If those intended prompt difficulties do not align with the ASR scoring or the human rating, there are important implications for item writers attempting to create parallel test forms. The determination of item

equivalence from one test to the next needs to be justified by demonstrating that item writers can reliably write prompts to an intended difficulty level.

This misalignment between the expert rater intended difficulties and the empirical human-item-level ratings created challenges and perhaps even cognitive dissonance for the raters. The use of an 8 level holistic scale that represented the whole range of speaking proficiency at the item level could have introduced a restricted range error. This could be an artifact of telling the raters the intended level of the prompt they were rating, but it could also be failure to use the rubric properly. For instance, it would be difficult for a prompt targeted at a Level 2 task to elicit a speech sample much higher than a Level 3 or 4, even if the examinee did respond with more extensive speech. The rater might be reticent in awarding a rating that was more than 2 levels higher than the prompt's intended difficulty level (e.g., elaborating in such a way as to demonstrate speaking ability beyond what the prompt was designed to elicit). Conversely, when a rater was scoring a failed attempt at a prompt targeted at Level 6 task, it might be difficult to know why an examinee was failing to perform at that level and there might be little evidence about what level the examinee could accomplish. The failure to offer an academic opinion on complex topics could mean the examinee was a beginning speaker with almost no speaking ability or it could be an intermediate speaker suffering linguistic breakdown because of the increased cognitive load. Raters might not know how low to rate such breakdown and may be reticent to assign a rating more than 2 or 3 levels below the prompt's intended difficulty level. Thus the ratings for all of the prompts judged with the holistic rubric clustered around the mean (SD = .11).

Using an at-level scale for each item (through the conversion of the holistic rubric ratings) functioned much better from a measurement perspective. First, there was a wider

dispersion of the prompt difficulty means (SD = 1.30). Second, the Rasch analysis showed the categories had the most uniform distribution so the categorical differences in ratings examinees received were the most equidistant (see Figure 9). Finally, there was a much stronger relationship ($r = .98$, $p < .01$) between the rater predicted difficulties (see Table 18) than there had been with the holistic scale ratings. Therefore, the low relationship established through using the holistic scale at the item level could be more indicative of scale misuse than the inability of the raters to differentiate performance when judging the different prompts. The results seem to indicate that either responses obtained from lower level prompts did provide some evidence of the examinee's ability to speak at a higher level or that raters tended to rate the respondents overall quality of the response on an 8 level scale in a compensatory manner (e.g. native-like fluency compensated for a failure to use academic language). Either way, an analysis of the at-level scale data verifies that the intended prompt difficulty did affect the overall assessment of speaking ability and the instruments ability to obtain a valid speech sample (i.e., one that represents the individual's true speaking ability).

Implications

ASR-scored open response speaking tests seem to fit two of the criteria test developers in higher education are seeking: the scoring is faster and cheaper. ASR scoring, however, is not better than, and as yet, is not even as good as human scoring. This reality leads to some important implications for test developers who want to employ current ASR scoring technology with their assessments.

Use of ASR to predict speaking scores. The current state of ASR-scoring allows for speaking tests with mixed task types and with mixed scoring methods. For this type of speaking test, some sections of the test could have prompts that elicit restricted speech while other

prompts elicit spontaneous speech. The restricted speech sections could use audio transcription to score the accuracy of the speech that was elicited, and the spontaneous sections could use extracted timing features to score the fluency. The scoring of this mixed task method might then provide a conjoint model in which different rating levels could be determined by requiring a predetermined level of accuracy (as determined through word recognition of the restricted response sections) and a predetermined level of fluency (as determined by the "timing" scores assigned in the spontaneous speech section).

Impact of ASR-scoring and human-rating on prompt difficulty. As developers create prompts for parallel test forms, they would need to consider two prompt characteristics. First, they need to ensure the prompts on a test range in difficulty in order to maintain construct validity of the test. Then, if equivalent forms of a test are desired, the prompts would need to be selected in such a way that the prompts that have the same intended difficulty in terms of the item difficulty statistics are similar from one form of the test to another. In creating a parallel form, there would need to be sufficient prompts at each of the targeted levels to assure the test composition had construct validity and to assure that the scoring of the parallel form would be equivalent.

For example, with this study, the test was comprised of ten items with two items targeted at five different levels of the rubric. Were we to create two parallel forms, we could build two, five-item tests, each test having one item at each intended level. If the ASR had scored those items in the same way the humans rated them, we could have parallel test forms. If, as study found, the ASR did not score the items in the same way as the humans, the forms would not be equivalent. Had there been more items, it would have been possible to select prompts that aligned with the human rater predicted difficulties. Furthermore, prompts that did not align with

ASR scoring and human ratings could be excluded. This could serve as a model for parallel form development with current ASR technology.

It is important to note that the rate of speech, mean length of run and number of pauses did vary across prompts. Since in its current state, ASR cannot make judgments on quality of the content the respondents produce and must score solely on proxy variables, it is important to ensure the scoring of the proxy variables aligns with human ratings of prompt difficulty. This requirement would be essential if the desire is to create adaptive tests in which the scoring of one question leads to the selection of the next question.

Use of an at-level scale in the human rating of items. Using holistic rubrics to rate individual items that are targeted at specific levels is problematic and should be done only with caution and verification that the ratings will be free from the rater error (central tendency, range restriction or logical errors). In this study a 5-point at-level scale linked to a holistic rubric but targeted at the intended level of the prompt yielded much better results. By using this scale, prompts that were targeted to elicit speech at the same level were more likely to represent their intended empirical difficulty levels. There was a clear separation in the scoring based on the intended prompt difficulty levels. A logical interpretation of this result suggest that the item difficulty is important and scoring, both human and ASR, is not invariant to the prompts used.

From the result of this analysis it was also noted that those prompts intended to elicit evidence of speaking ability at Levels 2 (L2-1, L2-2), 4 (L4-1, L4-2), and 6 (L6-1, L6-2) were of equal difficulty within level, but the prompts at Levels 3 (L3-1, L3-2) and 5 (L5-1, L5-2) were not of equal difficulty. Analyzing prompts in this way can provide evidence of test equivalence when attempting to create parallel forms of an assessment. It also provides a item level statistic of difficulty that could be used when creating item banks.

Limitations and Future Research

With any research, there are weaknesses and limitations, and the self-reflective researcher must evaluate how the study may have been improved. Often, these reflections lead to additional questions to be addressed in future research.

Use of ASR to predict speaking scores. The study might have yielded different results had a few different choices been made in the methodology of the study. First, the software that was used was designed for signal processing, more specifically for identifying timing features of speech that did not incorporate audio transcription. Had other signal features been analyzed such as rhythm or prosody been measured, the results may have been different. Second, though word recognition rates with audio transcription software are not very accurate, the ASR engines with that capability do provide information on pronunciation by comparing how close the sounds of the test file match those in the ASR language and acoustic dictionaries. Perhaps a pronunciation component would have improved the ASR scoring. Furthermore, if examinee responses from a previous trial administration of the prompts been available, those responses could have been transcribed and specific dictionaries built. This step would have increased the probability of later successfully transcribing the audio. If the words could be successfully transcribed, analyses on the frequency and diversity of the vocabulary used could have increased the ability of the ASR to predict human speaking ratings.

Impact of ASR-scoring and human-rating on prompt difficulty. This study examined a proficiency test with prompts that had a wide range of predicted difficulties. This test type is appropriate when placing students into a language-learning program or evaluating their general proficiency, but there are times when a screening test would be more appropriate. For instance, in assessing readiness to speak in an English medium university, having all the prompts targeting

the same difficulty level threshold (e.g. English needed to succeed at the university) might have yielded different results. If there had been more items at the same intended level, there might have been different findings. The items in the study consisted of two that were targeted at each level, but within the same level, there are a wide range of topics and tasks. For example, in this study, the Level 3 items included one prompt in which the examinee had to do simple future narration based on an itinerary and one in which the examinee had to give simple travel advice. Had both of the Level 3 items tested future narration, the ASR may have rated the item difficulties more similarly. Further analysis with more prompts that share more characteristics might yield more information on how the types of prompts (instead of intended levels) affect the fluency features of the responses and therefore how the ASR processes the responses.

Use of an at-level scale in the human rating of items. In this study, the raters who judged the item responses had been trained to rate overall performances with a holistic scale. They were not given any instruction or exemplars on how to apply the scale at the item level, and the task may have been untenable, as the rubric was not designed for use at the micro level of item. This design weakness might have been overcome if there had been more rater training that focused on the item level. Through the training, the challenge of implementing a holistic scale at the item level could have emerged and a change to the design could have been implemented at that time. Fortunately, the existing holistic scale could be converted after the fact so a more accurate analysis could still be made. Were the research to be done again, it would be better to (a) initially design the scale at the item level and (b) trial the scale to ensure it functions as intended. This would have avoided the step of needing to conduct a post hoc analysis.

Treating a rubric with three distinct axes (text type, content and accuracy) as a unidimensional construct could have affected the rating as well. Raters making expert judgments

of performance have a cognitive load placed upon them that could be simplified by letting them focus only on one aspect at a time. Then, if a multi-dimensional IRT model had been applied, the findings might have been different as well.

Conclusion

To summarize study findings, the combination of ASR timed fluency features that best predicted the human-rated holistic speaking ratings were speech rate, mean syllables per run, and number of silent pauses. However only 31% of the score variance was due to these features, so that relationship was not strong enough for the fluency features alone to predict speaking ability. Also, the item difficulties calculated by the ASR, did not align with the intended prompt difficulty predicted by the experts.

The use of automatic scoring with speaking may be more a matter of time than anything else. The Pearson Test of English (PTE) uses ASR exclusively to score the speech samples that are on its tests, so as technology improves in recognizing unrestricted speech, the inclusion unrestricted prompts to be rated by ASR scoring will likely increase. The ASR could then score the tests exclusively as in the case of the PTE or it could be paired with a human rater. The ETS TOEFL Writing test uses this combined scoring model, and any disagreement between the automatic scoring and human rating is resolved by a second human rater. The pressure to use ASR engines for scoring will come from the financial incentive of cost saving in paying human raters and in the potential of receiving test scores immediately.

Before embarking down the path of automated scoring, the unintended consequences of that approach should be explored. With spontaneous speech, the timing features alone are insufficient to predict human rating. As the technology improves and other features including pronunciation and word recognition are added to the ASR model, it will be imperative to conduct

additional research to ensure the assessments used also consider the validity, reliability, and interpretation of results. As the technology improves and improved regression equations are developed, ASR item level scoring might align more closely with the intended item difficulty of the experts. Whether improved item-level scoring would improve the accuracy of the ASR generated proficiency ratings would still need to be verified.

After noting that technological change can be either life threatening or life enhancing, Postman (1999) noted, "Only a fool would blithely welcome any technology without having given serious thought to the question"(p. 44). Test developers need to be vigilant in giving serious thought to the impact technology might have in the assessment process to ensure that the tests that are creating follow the psychometrician's oath to do no harm.

References

Adams, R. J., Griffin, P. E., & Martin, L. (1987). A latent trait method for measuring a dimension in second language proficiency. *Language Testing*, *4*(1), 9-27.

Alexander, A., Dessimoz, D., Botti, F., & Drygajlo, A. (2005). Aural and automatic forensic speaker recognition in mismatched conditions. *The International Journal of Speech, Language and the Law*, *12*(2), 214-234.

Anusuya, M. A., & Katti, S. K. (2011). Front end analysis of speech recognition: A review. *International Journal of Speech Technology*, *14*(2), 99-145.

Aron, J. (2011). How innovative is apple's new voice assistant, Siri? *The New Scientist*, *212*(2836), 24.

Bachman, L. F. (2004). *Statistical analyses for language assessment.* Cambridge, UK: Cambridge University Press.

Baylor, C., Hula, W., Donovan, N. J., Doyle, P. J., Kendall, D., & Yorkston, K. (2011). An introduction to item response theory and Rasch models for speech-language pathologists. *American Journal of Speech-language Pathology / American Speech-Language-Hearing Association*, *20*(3), 243-59.

Beigi, H. (2008). *Whether computer analyses can predict human ratings of speaking proficiency* (Technical Report: RTI-20081205-01). Retrieved from http://www.recognitiontechnologies.com/~beigi/homayoon/publications.html

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.

Boersma, P., & Weenink, D. (2005). Praat: Doing phonetics by computer [computer program]. *Version*, *5*, 21.

Bond, T. G., & Fox, C. M. (2007). *Applying the Rasch model : Fundamental measurement in the human sciences.* Mahwah, NJ: Lawrence Erlbaum Associates.

Brown, J. D. (1996). *Testing in language programs.* Upper Saddle River, NJ: Prentice Hall Regents.

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

Carr, N. T. (2011). *Designing and analyzing language tests.* Oxford, UK: Oxford University Press.

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

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.

Christensen, C. V. (2012). *Fluency features and elicited imitation as oral proficiency measurement.* Thesis. Retrieved from http://hdl.lib.byu.edu/1877/etd5468

Chun, M. (2010). Faster, better, cheaper: The iron triangle of higher education assessment. *Improving Writing and Thinking Through Assessment*, 87.

Cincarek, T., Gruhn, R., Hacker, C., Noth, E., & Nakamura, S. (2009). Automatic pronunciation scoring of words and sentences independent from the non-native's first language. *Computer Speech and Language*, *23*(1), 65-88.

Cox, T., & Davies, R. S. (2012). Using automatic speech recognition technology with elicited oral response testing. *CALICO Journal*, *29*(4), 601-618.

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory.* Fort Worth, TX: Harcourt Brace.

Cucchiarini, C., Neri, A., & Strik, H. (2009). *Oral Proficiency Training in Dutch L2: The Contribution of ASR-based Corrective Feedback.*

Cucchiarini, C., Strik, H., & Boves, L. (2000). Different aspects of expert pronunciation quality ratings and their relation to scores produced by speech recognition algorithms. *Speech Communication*, *30*, 109-119.

David, L. (Writer), & Cherones, T. (Director). (1993). The puffy shirt. [Television Series Episode]. In G. Shapiro, H. West, J. Seinfeld, & L. David (Executive Producers), *Seinfeld.* New York, NY: Sony Pictures.

Delattre, P. (1964). Comparing the vocalic features of English, German, Spanish and French. *International Review of Applied Linguistics*, *2*(2), 71-97.

Eckes, T. (2011). *Introduction to many-facet Rasch measurement: Analyzing and evaluating rater-mediated assessments.* Frankfurt, Germany: Peter Lang.

Engelhard Jr, G. (2008). Historical perspectives on invariant measurement: Guttman, Rasch, and Mokken. *Measurement*, *6*(3), 155-189.

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

Esling, J. H., & Wong, R. F. (1983). Voice quality settings and the teaching of pronunciation. *TESOL Quarterly*, *17*(1), 89-95.

Franco, H., Bratt, H., Rossier, R., Rao Gadde, V., Shriberg, E., Abrash, V., & Precoda, K. (2010). EduSpeak®: A speech recognition and pronunciation scoring toolkit for computer-aided language learning applications. *Language Testing*, *27*(3), 401-418.

Fulcher, G. (2003). *Testing second language speaking.* Pearson Education.

Fulcher, G., & Reiter, R. M. (2003). Task difficulty in speaking tests. *Language Testing*, *20*(3), 321-344.

Gammill, T. & Pross, M. (Writers), & Ackerman, A. (Director). (1994). The pledge drive. [Television Series Episode]. In G. Shapiro, H. West, J. Seinfeld, & L. David (Executive Producers), *Seinfeld.* New York, NY: Sony Pictures.

Ghosh, P. K., & Narayanan, S. (2011). Automatic speech recognition using articulatory features from subject-independent acoustic-to-articulatory inversion. *The Journal of the Acoustical Society of America*, *130*(4), 251-257.

Ginther, A., Dimova, S., & Park, S. (2012). *Interaction between fluency measures and task characteristics*. In *Workshop Fluent Speech. Utrecht, Netherlands.* Retrieved from http://nivjadj.wix.com/workshopfluentspeech

Ginther, A., Dimova, S., & Yang, R. (2010). Conceptual and empirical relationships between temporal measures of fluency and oral English proficiency with implications for automated scoring. *Language Testing*, *27*(3), 379.

Goto, H. (1971). Auditory perception by normal Japanese adults of the sounds "L" and "R". *Neuropsychologia*, *9*(3), 317-23.

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)

Griffin, P. E. (1985). The use of latent trait models in the calibration of tests of spoken language in large-scale selection-placement programs. In Y. P. Lee, A. C. Y. Fok, R. Lord, & G. Low (Eds.), *New directions in language testing* (pp. 149–161). Oxford: Pergamon.

He, X., & Zhao, Y. (2007). Prior knowledge guided maximum expected likelihood based model selection and adaptation for nonnative speech recognition. *Computer Speech and Language*, *21*(2), 247-265.

Henning, G. (1992). Dimensionality and construct validity of language tests. *Language Testing*, *9*(1), 1-11.

Hombo, C. M., Donoghue, J. R., & Thayer, D. T. (2001). *A simulation study of the effect of rater designs on ability estimation* [Research Report (RR-01-05)]. Educational Testing Service. Retrieved from Google Scholar

Huang, X. D., Ariki, Y., & Jack, M. A. (1990). *Hidden Markov Models for speech recognition.* Edinburgh: Edinburgh University Press.

Ip, E. H. (2010). Empirically indistinguishable multidimensional IRT and locally dependent unidimensional item response models. *The British Journal of Mathematical and Statistical Psychology*, *63*(Pt 2), 395-416. doi:10.1348/000711009X466835

Iwashita, N., McNamara, T., & Elder, C. (2002). Can we predict task difficulty in an oral proficiency test? Exploring the potential of an information-processing approach to task design. *Language Learning*, *51*(3), 401-436.

Jeon, J. H., & Liu, Y. (2012). Automatic prosodic event detection using a novel labeling and selection method in co-training. *Speech Communication*, *54*(3), 445-458.

de Jong, N. H., & Wempe, T. (2009). Praat script to detect syllable nuclei and measure speech rate automatically. *Behavior Research Methods*, *41*(2), 385-90. doi:10.3758/BRM.41.2.385

Kawahara, T., & Lee, A. (2005). Open-source speech recognition software Julius. *Journal of the Japanese Society for Artificial Intelligence*, *20*(1), 41-49.

Kim, H. J. (2006). Providing validity evidence for a speaking test using FACETS. *Teachers College, Columbia University Working Papers in TESOL & Applied Linguistics*, 6(1).

Kinnunen, T., & Li, H. (2010). An overview of text-independent speaker recognition: From features to supervectors. *Speech Communication*, *52*(1), 12-40.

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

Kolar, J., Liu, Y., & Shriberg, E. (2010). Speaker adaptation of language and prosodic models for automatic dialog act segmentation of speech. *Speech Communication*, 52(3), 236-245.

Koolagudi, S. G., & Krothapalli, R. S. (2011). Two stage emotion recognition based on speaking rate. *International Journal of Speech Technology*, *14*(1), 35-48.

Larson-Hall, J. (2010). *A guide to doing statistics in second language research using SPSS.* New York; London: Routledge.

Lee, K. -F. (1989). *Automatic speech recognition: The development of the Sphinx recognition system.* Boston, MA: Kluwer Academic Publishers.

Linacre, J. M. (1991). Log-odds in Sherwood Forest. *Rasch Measurement Transactions*, *5*, 162-63. Retrieved from http://www.rasch.org/rmt/rmt53d.htm.

Linacre, J. M. (1994). *Many-Facet Rasch measurement.* Chicago: MESA Press.

Linacre, J. M. (1999). *A user's guide to FACETS*. Chicago: MESA press.

Linacre, J. M., & Wright, B. D. (2009). *A user's guide to WINSTEPS*. Chicago: MESA press.

Livingston, S. A. (2004). *Equating test scores (without IRT).* Princeton, NJ: Educational Testing Services.

Loukina, A., Kochanski, G., Rosner, B., Keane, E., & Shih, C. (2011). Rhythm measures and dimensions of durational variation in speech. *The Journal of the Acoustical Society of America*, *129*(5), 3258-3270.

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

MacSpeech Dictate. (2010). [Computer Software]. Burlington, MA: Nuance Communications.

Manning, C. D., & Schütze, H. (1999). *Foundations of statistical natural language processing.* Cambridge, MA: MIT press.

Matsushita, H. (2011). *Computerized oral proficiency test for Japanese: Measuring L2 speaking ability with ASR technology.* Thesis. Retrieved from http://contentdm.lib.byu.edu/cdm/ref/collection/ETD/id/27

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

McNamara, T., & Knoch, U. (2012). The Rasch wars: The emergence of Rasch measurement in language testing. *Language Testing.*

McNamara, T. F. (1996). *Measuring second language performance.* London, UK: Longman.

Mehlman, P. (Writer), & Ackerman, A. (Director). (1994). The Chinese woman. [Television Series Episode]. In G. Shapiro, H. West, J. Seinfeld, & L. David (Executive Producers), *Seinfeld.* New York, NY: Sony Pictures.

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

Moustroufas, N., & Digalakis, V. (2007). Automatic pronunciation evaluation of foreign speakers using unknown text. *Computer Speech & Language*, *21*(1), 219-230.

Myford, C. M., & Wolfe, E. W. (2003). Detecting and measuring rater effects using many-facet Rasch measurement: Part I. *Journal of Applied Measurement*, *4*(4), 386-422.

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. doi:10.1007/s10459-010-9222-y

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).

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

Owren, M. J. (2008). GSU Praat tools: Scripts for modifying and analyzing sounds using praat acoustics software. *Behavior Research Methods*, *40*(3), 822-9.

Petersen, N. (2007). Equating: Best practices and challenges to best practices. In N. J. Dorans, M. Pommerich, & P. W. Holland (Eds.), *Linking and aligning scores and scales* (pp. 59-72). New York, NY: Springer.

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.

Postman, N. (1990). Informing ourselves to death. *Speech at the German Informatics Society.* Retrieved from http://w2.eff.org/Net_culture/Criticisms/informing_ ourselves_to_death.paper

Postman, N. (1999). *Building a bridge to the 18th century: How the past can improve our future.* New York: Alfred A. Knopf : Distributed by Random House.

Raab, M., Gruhn, R., & Noth, E. (2011). A scalable architecture for multilingual speech recognition on embedded devices. *Speech Communication*, *53*(1), 62-74.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests.* Copenhagen: Danmarks Paedogogiske Institut.

Raykov, T., & Marcoulides, G. A. (2010). *Introduction to psychometric theory.* New York, NY: Routledge.

Reckase, M. D. (2009). Historical background for multidimensional item response theory (MIRT). *Multidimensional Item Response Theory*, 57-77.

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

Sangwan, A., & Hansen, J. H. L. (2012). Automatic analysis of Mandarin accented English using phonological features. *Speech Communication*, *54*(1), 40-54.

Schumacker, R. E. (1999). Many-facet Rasch analysis with crossed, nested, and mixed designs. *Journal of Outcome Measurement*, *3*(4), 323-38.

Shavelson, R. J., Baxter, G. P., & Gao, X. (1993). Sampling variability of performance assessments. *Journal of Educational Measurement*, *30*(3), 215-232.

Stansfield, C., & Kenyon, D. (1996). Comparing the scaling of speaking tasks by language teachers and by the ACTFL guidelines. In A. Cumming & R. Berwick (Eds.), *The concept of validation in language testing* (pp. 124-153). Clevedon, Avon: Multilingual Matters.

Stevens, S. S. (1946). On the theory of scales of measurement. *Science*, 677-680.

Sykes, R. C., Ito, K., & Wang, Z. (2008). Effects of assigning raters to items. *Educational Measurement: Issues and Practice*, *27*(1), 47-55.

Tajima, K., Zawaydeh, B. A., & Kitahara, M. (1999). A comparative study of speech rhythm in arabic, english, and japanese. *Proceedings of the XIV ICPhS, San Francisco, USA.*

Taylor, L. B. (2011). *Examining speaking: Research and practice in assessing second language speaking.* Cambridge; New York: Cambridge University Press.

van der Walt, C., de Wet, F., & Niesler, T. (2008). Oral proficiency assessment: The use of automatic speech recognition systems. *Southern African Linguistics and Applied Language Studies*, *26*(1), 135-146.

van Doremalen, J., Strik, H., & Cucchiarini, C. (2009). Optimizing non-native speech recognition for CALL applications. *Optimizing Non-native Speech Recognition for CALL Applications.*

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.

Ward, N. G., Vega, A., & Baumann, T. (2012). Prosodic and temporal features for language modeling for dialog. *Speech Communication*, *54*(2), 161-174.

Wright, B. D., & Linacre, J. M. (1989). Observations are always ordinal; measurements, however, must be interval. *Archives of Physical Medicine and Rehabilitation*, *70*(12), 857.

Wu, S., Falk, T. H., & Chan, W. -Y. (2011). Automatic speech emotion recognition using modulation spectral features. *Speech Communication*, *53*(5), 768-785.

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

Xi, X., Higgins, D., Zechner, K., & Williamson, D. (2012). A comparison of two scoring methods for an automated speech scoring system. *Language Testing*, *29*(3), 371-394.

Xue, S. A., & Deliyski, D. (2001). Effects of aging on selected acoustic voice parameters: Preliminary normative data and educational implications. *Educational Gerontology*, *27*(2), 159-168.

Yu, Y., & Brown, W. L. (2000). Raters and single prompt-to-prompt equating using the facets model in a writing performance. *Objective Measurement: Theory Into Practice*, *5*, 97-111.

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.

Zelinka, P., Sigmund, M., & Schimmel, J. (2012). Impact of vocal effort variability on automatic speech recognition. *Speech Communication*, *54*(6), 732-742.

Appendix A

English Language Center Speaking Rubric

Level	Text Type	Accuracy	Content
7—leaving Academic C	Exemplified speaking on a paragraph level rather than isolated phrases or strings of sentences. Highly organized argument (transitions, conclusion, etc.).Speaker explains the outline of topic and follows it through.	• Grammar errors are extremely rare, if they occur at all; wide range of structures in all time frames; • Able to compensate for deficiencies by use of communicative strategies—paraphrasing, circumlocution, illustration—such that deficiencies are unnoticeable; • Pausing and redundancy resemble native speakers; • Intonation resembles native-speaker patterns; pronunciation rarely if ever causes comprehension problems; • Readily understood by native speakers unaccustomed to non-native speakers;	• Discuss some topics abstractly (areas of interest or specific field of study); • Better with a variety of concrete topics; • Appropriate use of formal and informal language; • Appropriate use of a variety in academic and non-academic vocabulary;
6—starting Academic C	Fairly organized paragraph-like speech with appropriate discourse markers (transitions, conclusion, etc.) Will not be as organized as level 7, but meaning is clear.	• Grammar errors are infrequent and do not affect comprehension; no apparent sign of grammatical avoidance; • Able to speak in all major time frames, but lacks complete control of aspect; • Pausing resembles native patterns, rather than awkward hesitations; • Often able to successfully use compensation strategies to convey meaning;	• Uses appropriate register according to prompt (formal or informal) • Can speak comfortably with concrete topics, and discuss a few topics abstractly; • Academic vocabulary often used appropriately in speech;
5—starting Academic B	Simple paragraph length discourse.	• Uses a variety of time frames and structures; however, speaker may avoid more complex structures; • Exhibits break-down with more advanced tasks—i.e. failure to use circumlocution, significant hesitation, etc. • Error patterns may be evident, but errors do not distort meaning; • Pronunciation problems occur, but meaning is still conveyed • Understood by native speakers unaccustomed to dealing with non-natives, but 1st language is evident;	• Able to comfortably handle all uncomplicated tasks relating to routine or daily events and personal interests and experiences; • Some hesitation may occur when dealing with more complicated tasks; • Uses a moderate amount of academic vocabulary;
4—starting Academic A	Uses moderate-length sentences with simple transitions to connect ideas. Sentences may be strung together, but may not work together as cohesive paragraphs.	• Strong command of basic structures; error patterns with complex grammar; • Pronunciation has significant errors that hinder comprehension of details, but not necessarily main idea; • Frequent pauses, reformulations and self-corrections; • Successful use of compensation strategies is rare; • Generally understood by sympathetic speakers accustomed to speaking with non-natives;	• Able to handle a variety of uncomplicated tasks with concrete meaning; • Expresses meaning by creating and/or combining concrete and predictable elements of the language; • Uses sparse academic vocabulary appropriately;

3—starting Foundations C	Able to express personal meaning by using simple, but complete, sentences they know or hear from native speakers.	• Errors are not uncommon and often obscure meaning; • Limited range of sentence structure; • Intonation, stress and word pronunciation are problematic and may obscure meaning; • Characterized by pauses, ineffective reformulations; and self-corrections; • Generally be understood by speakers used to dealing with non-natives, but requires more effort;	• Able to successfully handle a limited number of uncomplicated tasks; • Concrete exchanges and predictable topics necessary for survival; • Highly varied non-academic vocabulary;
2—starting Foundations B	Short and sometimes incomplete sentences.	• Attempt to create simple sentences, but errors predominate and distort meaning; • Avoids using complex/difficult words, phrases or sentences; • Speaker's 1st language strongly influences pronunciation, vocabulary and syntax; • Generally understood by sympathetic speakers used to non-natives with repetition and rephrasing;	• Restricted to a few of the predictable topics necessary for survival (basic personal information, basic objects, preferences, and immediate needs) • Relies heavily on learned phrases or recombination of phrases and what they hear from interlocutor; • Limited non-academic vocabulary
1—starting Foundations A	Isolated words and memorized phrases.	• Communicate minimally and with difficulty; • Frequent pausing, recycling their own or interlocutor's words; • Resort to repetition, words from their native language, or silence if task is too difficult; • Understood with great difficulty even by those used to dealing with non-natives	• Rely almost solely on formulaic/memorized language; • Very limited context for vocabulary; • Two or three word answers in responding to questions;
0—starting foundations prep.	Isolated words.	• May be unintelligible because of pronunciation; • Cannot participate in true conversational exchange; • Length of speaking sample may be insufficient to assess accuracy;	• No real functional ability; • Given enough time and familiar cues, may be able to exchange greetings, give their identity and name a number of familiar objects from their immediate environment;

Appendix B

Rating Design for Human-rated Holistic Speaking Level

Below is the rating design used for the Human-rated Holistic Speaking Level. It is an incomplete connected design with 16 raters and 222 examinees. This design allowed for connected subsets in performing the FACETS analysis.

Students	Tasks	Rater 1	Rater 2	Rater 3	Rater 4	Rater 5	Rater 6	Rater 7	Rater 8	Rater 9	Rater 10	Rater 11	Rater 12	Rater 13	Rater 14	Rater 15	Rater 16	Total
1 to 5	1 to 10	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	16
6	1 to 10	X	X	X	X					X								5
7	1 to 10	X	X	X	X													4
8 to 40	1 to 10	X			X													2
41	1 to 10	X			X		**X**			X								4
42 to 60	1 to 10	X					**X**											2
61 to 76	1 to 10		X				**X**											2
77	1 to 10		X				**X**			X		X						4
78 to 93	1 to 10		X									X						2
94	1 to 10		X							X		X				X		4
95 to 97	1 to 10		X									X				X		3
98 to 110	1 to 10			X												X		2
111 to 112	1 to 10			X												X	X	3
113 to 127	1 to 10			X													X	2
128 to 129	1 to 10			X						X					X		X	4
130 to 134	1 to 10			X											X			2
135 to 144	1 to 10					X				X					X			3
145 to 146	1 to 10					X								X	X			3
147 to 161	1 to 10					X								X				2
162	1 to 10					X				X			X	X				4
163	1 to 10					X							X	X				3
164 to 171	1 to 10					X							X					2
172 to 178	1 to 10							X					X					2
179	1 to 10							X			X		X					3
180	1 to 10					X		X		X	X		X					5
181 to 187	1 to 10							X			X							2
188	1 to 10							X	X		X							3
189 to 196	1 to 10								X		X							2
197	1 to 10								X	X	X							3
198 to 201	1 to 10								X	X								2

Appendix C

Rating Design for Analytic Human-rated Item Speaking Level

Below is the rating design used for the Human-rated Holistic Speaking Level. It is an incomplete spiral connected design with 10 raters and 201 examinees. This design allowed for connected subsets in performing the FACETS analysis.

Student	Prompt	Rater 1	Rater 2	Rater 3	Rater 4	Rater 5	Rater 6	Rater 7	Rater 8	Rater 9	Rater 10	Total
1 to 3	1 to 10	X	X	X	X	X	X	X	X	X	X	10
4 to 10	1	X	X									2
4 to 10	2	X	X									2
4 to 10	3	X		X								2
4 to 10	4	X			X							2
4 to 10	5	X				X						2
4 to 10	6	X					X					2
4 to 10	7	X						X				2
4 to 10	8	X							X			2
4 to 10	9	X								X		2
4 to 10	10	X									X	2
11 to 20	1	X	X									2
11 to 20	2		X									2
11 to 20	3		X	X								2
11 to 20	4		X		X							2
11 to 20	5		X			X						2
11 to 20	6		X				X					2
11 to 20	7		X					X				2
11 to 20	8		X						X			2
11 to 20	9		X							X		2
11 to 20	10		X								X	2
21 to 30	1	X		X								2
21 to 30	2		X	X								2
21 to 30	3			X								2
21 to 30	4			X	X							2
21 to 30	5			X		X						2
21 to 30	6			X			X					2
21 to 30	7			X				X				2
21 to 30	8			X					X			2
21 to 30	9			X						X		2
21 to 30	10			X							X	2
31 to 40	1	X			X							2
31 to 40	2		X		X							2
31 to 40	3			X	X							2
31 to 40	4				X							2
31 to 40	5				X	X						2
31 to 40	6				X		X					2
31 to 40	7				X			X				2
31 to 40	8				X				X			2
31 to 40	9				X					X		2
31 to 40	10				X						X	2
41 to 50	1	X				X						2
41 to 50	2		X			X						2
41 to 50	3			X		X						2
41 to 50	4				X	X						2
41 to 50	5					X						2
41 to 50	6					X	X					2
41 to 50	7					X		X				2

41 to 50	8						X				X			2
41 to 50	9						X					X		2
41 to 50	10						X						X	2
51 to 60	1	X						X						2
51 to 60	2		X					X						2
51 to 60	3			X				X						2
51 to 60	4				X			X						2
51 to 60	5						X	X						2
51 to 60	6							X						2
51 to 60	7							X	X					2
51 to 60	8							X		X				2
51 to 60	9							X				X		2
51 to 60	10							X					X	2
61 to 70	1	X							X					2
61 to 70	2		X						X					2
61 to 70	3			X					X					2
61 to 70	4				X				X					2
61 to 70	5						X		X					2
61 to 70	6							X	X					2
61 to 70	7								X					2
61 to 70	8								X	X				2
61 to 70	9								X			X		2
61 to 70	10								X				X	2
71 to 80	1	X								X				2
71 to 80	2		X							X				2
71 to 80	3			X						X				2
71 to 80	4				X					X				2
71 to 80	5						X			X				2
71 to 80	6							X		X				2
71 to 80	7								X	X				2
71 to 80	8									X				2
71 to 80	9									X	X			2
71 to 80	10									X			X	2
81 to 90	1	X									X			2
81 to 90	2		X								X			2
81 to 90	3			X							X			2
81 to 90	4				X						X			2
81 to 90	5						X				X			2
81 to 90	6							X			X			2
81 to 90	7								X		X			2
81 to 90	8									X	X			2
81 to 90	9										X			2
81 to 90	10										X	X		2
91 to 100	1	X										X		2
91 to 100	2		X									X		2
91 to 100	3			X								X		2
91 to 100	4				X							X		2
91 to 100	5						X					X		2
91 to 100	6							X				X		2
91 to 100	7								X			X		2
91 to 100	8									X		X		2
91 to 100	9										X	X		2
91 to 100	10											X		2
101 to 201	1	X												1
101 to 201	2		X											1
101 to 201	3			X										1
101 to 201	4				X									1
101 to 201	5						X							1
101 to 201	6							X						1
101 to 201	7								X					1
101 to 201	8									X				1
101 to 201	9										X			1
101 to 201	10												X	1