2012-07-07

Fluency Features and Elicited Imitation as Oral Proficiency Measurement

Carl V. Christensen
Brigham Young University - Provo

Fluency Features and Elicited Imitation as Oral Proficiency Measurement

Carl Christensen

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

Master of Arts

Deryle Lonsdale, Chair
Dan P. Dewey
William G. Eggington

Department of Linguistics and English Language

Brigham Young University

August 2012

ABSTRACT

Fluency Features and EI as Oral Proficiency Measurement

Carl Christensen
Department of Linguistics and English Language
Master of Arts

The objective and automatic grading of oral language tests has been the subject of significant research in recent years. Several obstacles lie in the way of achieving this goal. Recent work has suggested a testing technique called elicited imitation (EI) can be used to accurately approximate global oral proficiency. This testing methodology, however, does not incorporate some fundamental aspects of language such as fluency. Other work has suggested another testing technique, simulated speech (SS), as a supplement to EI that can provide automated fluency metrics. In this work, I investigate a combination of fluency features extracted for SS testing and EI test scores to more accurately predict oral language proficiency. I also investigate the role of EI as an oral language test, and the optimal method of extracting fluency features from SS sound files. Results demonstrate the ability of EI and SS to more effectively predict hand-scored SS test item scores. I finally discuss implications of this work for future automated oral testing scenarios.

Keywords: Second language oral proficiency, Elicited Imitation, Simulated Speech, Automatic Speech Recognition, language modalities, speech signal processing, computerized oral test

ACKNOWLEDGEMENTS

I am truly grateful for the influence and support of my family. My grandfather, Soren Cox, and parents, Bryce and Mary Christensen, have encouraged, supported, and inspired me greatly. My wife and children have been wonderful in helping me achieve my goals.

My advisor, Deryle Lonsdale, has truly shaped my academic career and inspired, taught, and directed me in my study of linguistics. I am sincerely indebted to him for encouraging my involvement in various research groups which was the genesis for my involvement in the field of study discussed in this work.

I would like to express my appreciation for the help and support of Troy Cox at the English Language Center which was crucial in the completion of this work. Also, I am indebted to my colleagues and former classmates Hitokazu Matsushita and Ross Hendrickson and their work which aided significantly in my progress in this research and was fundamental for this work specifically. I am also appreciative of the students and advisors of the PSST group and ELC that have facilitated the data collection and grading necessary for this work

Table of Contents

List of Figures

List of Tables

Chapter 1 – Introduction

1.1 Background

The need to establish methods for accurately and efficiently determining the oral proficiency of second language learners of a language has received attention from researchers for decades (Henning 1983, Lazaraton 2002). Second-language tests use a variety of methods and theories in order to assess either a particular linguistic skill or global oral proficiency. Most tests rely on test prompts and spontaneous speech in order to evaluate the ability of the learner to produce speech that can be evaluated according to specific criteria. Often, the final scores are holistic in nature. One problem with spontaneous-speech tests that assess global oral proficiency is the difficulty of scoring such tests. Graders are normally asked to assign scores according to test rubrics, which attempt to establish a set of criteria to standardize scoring. But even under the best of scenarios, objectivity in grading is difficult to achieve in oral-language tests.

Recent advances in technology have led many second-language researchers to investigate automated scoring for oral tests. Automated scoring offers the potential benefit of being objective and uniform. The difficulty for automated scoring in oral testing lies in the somewhat generic and abstract factors usually considered when assigning a score. These factors include correct use of grammar and vocabulary, fluency, correct pronunciation, etc. The accurate measurement of many of these linguistic features is currently beyond the current abilities of technology, making test automation either very difficult or completely impossible. As a result, the standard methods of oral testing usually require human testers or scorers.

Involving human testers and scorers makes oral testing expensive and slow. Various testing methodologies have been proposed that require minimal human scoring and make use of automated computer administration and automated scoring via automated speech recognition

(ASR). The difficulty in the use of automated technology, especially in oral testing, comes from the imperfect science of ASR, and the multi-faceted nature of oral testing already discussed. As one might expect, grading a student with an instrument based on an imperfect platform presents problems for automated testing (Mostow and Aist 1000). While ASR quality has improved substantially in the last decade, many of the gains require calibration usually involving reciting a number of predefined phrases, reading a passage of text, and inputting various personal characteristics for the particular speaker – something not feasible in most testing scenarios. Even with calibration, ASR for spontaneous speech for non-native speakers is a challenging task.

One testing methodology that makes use of automated components is elicited imitation (EI). Test items consist of sentences that can be recorded beforehand. The test can be administered on a computer via either a stand-alone program, or a web-based application. The subject hears the stimulus and repeats the sentence. The subject's responses are recorded and either stored locally or sent to a server. Because the desired responses to the test items are known *a priori*, ASR can be used with a high level of accuracy (Graham *et al.* 2008); therefore, the problems mentioned above are less of an issue in this context.

EI also allows for unique test-item construction, which helps target particular grammatical structures or lexical items (Graham *et al*. 2008, Weitze and Lonsdale 2011). Oral test construction usually involves compiling a list of stimuli that will set up a particular scenario and linguistic environment to elicit from the subject the language features being tested. The responses normally follow a particular script or language pattern, but are spontaneous. Specific judgments about the subject's syntactic and lexical abilities are reflected in the holistic score. This process is another factor that makes automation difficult. EI enables more targeted testing of the sort that is more suitable for automation.

In contrast with the spontaneous-speech model used for oral testing, EI requires the response to mirror the stimulus. As already mentioned, this does allow for improved accuracy in ASR-based scoring. The possibility of controlling the responses directly allows for engineering test items to include particular linguistics features. This control over the subject's response also mitigates other issues with spontaneous tests, issues such as data sparsity. For example, eliciting a particular verb form or particular vocabulary item(s) via elicited imitation requires only constructing stimulus sentences that contain the desired features. In spontaneous forms of oral-language testing, the test items can specify a particular way to answer a question (narrative) or can focus on a particular subject, but rely on the response to provide the necessary data for correctly determining whether a linguistic feature has been mastered. These advantages to the EI testing method have led many to further investigate its utility in automatically and accurately assessing oral proficiency. Results from numerous studies have shown that EI is a good indicator of global oral proficiency (Hendrickson *et al*. 2008, Lonsdale *et al*. 2009).

By definition, EI does not incorporate various language phenomena that occur in spontaneous speech that are important indicators or global oral proficiency. Chief among these phenomena is oral fluency. Language fluency is an indispensable—if often difficult to directly define and measure—component of most oral language testing methodologies. Researchers have identified various way of quantifying fluency via features such as hesitation patterns, turn-taking, length of narration, and discourse management (Ellis 1993, Freed *et al*. 2004). Other forms of oral-language testing (such as the OPI) are geared to identifying and testing these features much more accurately. These interview-style oral tests take advantage of normal discourse patterns to evaluate the control of a language learner over various aspects of language. However, as already mentioned, this makes ASR-scoring more difficult. Recent work has focused on using

automatically identified fluency features to serve as a measurement for grading (Koponen and Riggenbach 2000, Segalowitz 2010).

One method of testing that has recently garnered significant attention is referred to as semi-direct or simulated speech (SS). The description of the test as semi-direct and simulated derives from the testing methodology requiring a monologue type response to the stimulus in a simulated environment instead of a dialog- or interview-style test. This method relies on computerized test administration in order to reduce the linguistic resource burden of supplying test administrators for each oral test. It does not, however, typically make use of automated scoring, though many efforts in this area are on-going. These automated methods of scoring usually rely on using a limited vocabulary language model for the ASR engine, phrase or word-spotting, or feature extraction (Zhang *et al*. 2007). The automated method I will explore is that of feature extraction, specifically fluency-feature extraction. Essentially, many of the fluency features that constitute the grading rubric for other oral language testing methodologies can also be found in a computer-administered test—such as an SS test—that elicits spontaneous speech. The SS samples can then be processed via an ASR engine to calculate metrics that can be used for grading.

Matsushita (2011) investigated the utility of the combination of the EI score and a simulated speech test in predicting OPI scores for Japanese. He identified eleven features that he could extract from the test responses using the Julius recognition engine. Table 1 identifies the features that he used in his study. The results were promising and invite validation in English. Therefore, in my study, I will adjust some of the features used in his study to account for available technology and language differences, but will attempt to demonstrate similar advantages in the combination of SS fluency features and EI test scores.

Feature	Description
(1) # Tokens	Number of morpheme tokens in a test item
(2) # Types	Number of morpheme types in a test item
(3) # Pauses	Number of short pauses in speech
(4) Speech Length	Total speech duration in a test item
(5) Silence Length	Total length of silence in a test item
(6) Speech Rate	Number of phonemes per second
(7) # Fillers	Number of filled pauses in a test item
(8) # Runs	Number of fluent speech runs in a test item
(9) Tokens per Run	Number of tokens normalized by fluent speech runs
(10) Speech Time per Run	Speech length normalized by fluent speech runs
(11) Types per Speech Length	Number of types normalized by speech length

Table 1: Fluency features used in Matsushita (2011)

1.2 Thesis Structure

In order to explore the potential for combining EI with fluency features extracted from SS, particularly in automated testing, I will direct my research into four inter-related areas in order to answer the following questions:

1. What information does comparing the results of EI with results of other language tests give in respect to better understanding the role of the EI test in language testing?
2. Which tool is ideal for extracting fluency features from SS test result files?
3. Can machine-learning and statistical techniques utilize the fluency features that are extracted in order to accurately predict holistic SS scores?
4. How do the SS and EI correlate, and does adding automatically extracted fluency features to EI scores better account for a holistic score assigned to an SS test then EI alone?

1.3 EI and Other Language Modalities

The distinct characteristics of the EI test have already been discussed in detail, but these fundamental characteristics require more probing investigation in establishing the relationship between EI and other modalities of language testing. Most of the oral-testing methods discussed above are graded largely on the basis of considerations completely inapplicable to the EI test. This investigation of the combination of SS and EI tests as a more accurate method of determining global oral proficiency proceeds on the assumption that EI provides a reliably accurate measurement of oral production. And while researchers have demonstrated a good correlation between EI and other oral tests (Graham 2006), the test remains a fundamental outlier in oral-language testing.

In this vein, beyond oral-language testing, I will study the correlation of EI test scores with scores in other language-testing areas, such as grammar, reading, and listening. The unique aspects of the EI test have led many to question which aspect of language is really being tested – the listening or comprehension, or the speaking (Hood and Lightbrown 1978, Vinther 2002). Understanding this distinction in modality is fundamental to correct usage of the EI test. As previously mentioned, the possibility of designing test *responses* allows research into language-acquisition testing not directly possible in most other oral-language tests. Because of the quick and direct access to particular language features, EI could serve as a supplement to various testing methods to reduce uncertainty about acquisition of particular linguistic features, and provide useful insight into the cross-over of various language modalities into the oral realm.

1.4 Fluency-Feature Extraction Tool

The definition of oral fluency is often elusive. Regardless of the exact interpretation of the concept of fluency, oral-language testing proceeds on the basic assumption that particular characteristics or features of oral speech are both desirable and indicative of global proficiency. As previously mentioned, Matsushita (2011) investigated the utility of an ASR engine for extracting fluency features. The main supposition underlying his study was that although transcription of open vocabulary, non-native sound files via ASR is flawed, it is predictably flawed. This is, however, not always the case. Modern language models and acoustic models are dependent largely on a window of prior context. With varying acoustic or language context, results could conceivably vary considerably (Mostow and Aist 1999).

I will attempt to contrast this ASR-based feature-extraction methodology with a methodology relying solely on signal processing or lower-level analysis of the speech signal's properties. Because signal processing does not rely on context-dependent models as does ASR, results can be expected to be more accurate, if less complex and detailed. The number and type of features available in these differing systems are not the same; therefore, I will identify candidate features for each independently. More features are available for ASR extraction simply because of the additional resources available in the system. However, one could expect the accuracy of the extracted features from the signal-processing method to be significantly higher without the increased complexity. This comparison will provide useful insight into the methodology that should be used to extract fluency features.

1.5 Fluency Features and Machine Learning

Researchers use various ways to map fluency features to test scores (Higgins *et al*. 2011). For this study, I will analyze the features using machine learning (ML). Machine learning has found growing application in the field of linguistics and language learning in the last decade. By passing the fluency features extracted via the tools into a ML component, I will be able to model the importance of the features in correctly predicting scores on an SS test and in demonstrating the correlation of these scores and the holistic outcome of the test. The ML will also be useful in determining which extraction method is superior by making it possible to compare the accuracy of the ML model produced by the set of features extracted via the ASR to the accuracy of the model created by relying on the signal-processing method. Statistical modeling is also relevant and applicable for this type of research. Inputting the fluency features into a regression model makes it possible to calculate the amount of variance accounted for by these features. The regression model can also be called upon to predict or extrapolate SS scores. The scores predicted via the models will serve to complement the EI ASR-based scores and provide better correlation with a holistic score.

1.6 EI and SS Testing Battery

While many other language-testing batteries are more standard, I will explore the advantages of a battery of tests composed of SS and EI exams. Besides allowing for automatic scoring, as I have already indicated, both of these tests can be graded objectively. While a variety of scoring methods do exist for both fluency-style features and EI-test items (Tomita *et al*. 2009; Matsushita 2011), for the purpose of this study I will choose only the most standard or logical choices for scoring. Because all scoring is automatic and based solely on features or

characteristics static across test examinations and test subjects, other methods of scoring should map naturally into this method. Thus other methods will remain peripheral to this study.

I will investigate the utility of the combination of these tests for more accurately predicting other oral-language measurements than can be done with either of the tests separately. By combining the strengths of both the EI test and a spontaneous-speech test such as the sLAT, I will investigate the correlation of automated scoring with human scoring. If the automated results can be obtained quickly and predictably for the EI and SS tests, while maintaining a high correlation for other oral global-language measurements, it will demonstrate potential efficiency gains for second-language testing. These gains could indicate an quicker, easier and more efficient way in which to provide a global measure for a second language learner

Chapter 2 – Literature Review

In this chapter I will discuss the body of research conducted by scholars in the area of elicited imitation (EI) and simulated speech (SS). I will examine studies on differing L2 language-test modality correlations and their application to fundamental questions about EI. Finally, I will also survey the work done in the area of fluency-feature extraction for L2 language analysis and discuss the importance of choosing correct fluency features in order to reach the goal of this study. Familiarity with the history of the EI and SS testing methodologies allows for greatly improved understanding of the role of each test in current language-testing research and in real-world applications. The historical development of SS testing will also frame the issues surrounding methods of extracting fluency features from the test results. Similarly, the historical use of the EI test also provides insight into the research of various types of language modality overlap and how that overlap can affect correct use of EI tests today.

2.1 Elicited Imitation Background

The basic concept of using EI as a language assessment tool has been analyzed in various studies. The procedure for EI usage is quite uniform: a subject hears an utterance and repeats it back verbatim (Chaudron 2003). Tests usually consist of multiple-stimulus sentences which elicit a target grammatical construction, morpheme, lexical item, etc. The actual linguistic processes involved are, however, still under debate (Hood and Lightbrown 1978, Jessop *et al*. 2007).

The EI test relies on the assumption that a subject's ability to repeat the stimulus is a reflection of his ability to process the input. According to the theory, the stimulus is processed by chunking the sentence into existing language structures in the brain and storing the

representation in working memory (Bley-Vroman and Chaudron 1994, Vinther 2002). The subject must then reconstruct the utterance from the memory, once again using his available language resources to repeat the utterance. Figure 1 below illustrates the round-trip between stimulus and response according to EI theory as depicted by Matsushita (2011), following the stimulus through the subject's language comprehension and production faculties. In the limit, native speakers of a language will not be able to repeat a phrase of arbitrary length because of working memory constraints. However, for sentences under a particular length threshold, the level of proficiency of a language—both for non-native speakers and children—seems to directly influence the ability of a subject to correctly repeat the utterance, (Hendrickson *et al*. 2008, Natalica 1976).

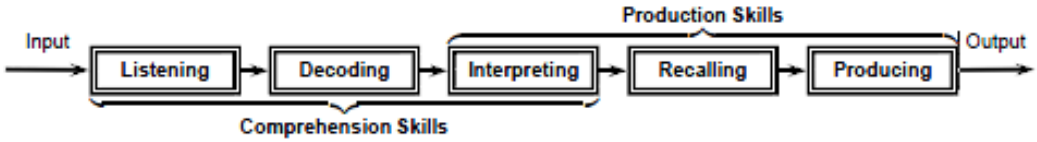


Figure 1: EI theory schematic from Matsushita (2011)

The main variable that contributes to the difficulty of an EI test item is the length, measured most commonly by the number of syllables the utterance contains (Chapman and Miller 1975). For extremely short sentences, otherwise difficult components can be stored in working memory without being processed or completely understood, which results in the subject essentially parroting the stimulus. For arbitrarily long items, the constraints of working memory are exceeded regardless of the proficiency in the language. However, inside a band of item

lengths, particular linguistic features also seem to have significant impact on subjects' ability to chunk and repeat the stimulus correctly (Hendrickson *et al.* 2008).

A substantial amount of legacy work in the EI field has been in the area of first-language acquisition (e.g. Menyuk 1964, Hood and Lightbrown 1978, McDade *et al*. 1982). In particular, many studies focused largely on the capacity of the EI methodology to measure grammatical acquisition (Carrow 1974, Ambridge and Pine 2006). A recent study conducted by Valian *et al*. (2006) explored the effect of direct-object predictability on EI test items for first-language acquisition among children. They showed that the predictability of the direct object reliably correlates with item scores. Essentially, as the naturalness of the verb-object combination increased, so did test scores. An item with an improbable or odd combination seemed to require the children to memorize more and chunk less. The study has direct relevance in the second-language acquisition research into EI items as test item design has shown to affect L2 learners as well (Christensen *et al.* 2010). In essence, this study demonstrates that either the production or comprehension components of a subject are affected by the syntactic complexity of a test item. This finding aligns with more recent work indicating that the EI test is a measure of linguistic knowledge, not merely memory capacity (Okura and Lonsdale 2012).

Research into the usage of EI as a language-assessment tool for second-language learners has increased markedly in the last few decades. Just as in studies of L1 acquisition, EI research for L2 learners has largely focused on grammatical structures or lexical items. Various studies have investigated which features affect the difficulty of an EI test item. Graham *et al*. (2010) conducted a study in which they constructed items from particular lexical frequency bands. The lexical density of an item showed a distinct correlation with the subject's ability to repeat the stimulus correctly. Hendrickson *et al*. (2008) investigated the role of features in a

particular syllable band. Their study identified various syntactic, morphological, and lexical features that are statistically significant. These studies underscore the ability of EI to test various aspects of a language learner's linguistic knowledge.

The differences between EI and natural scenarios of repetition (such as classroom interaction) are highlighted by Jessop *et al.* (2007). These researchers cite the laboratory settings and the content of target grammatical structures that make EI unique. They also point out that this form of testing allows researchers to elicit structures otherwise difficult to observe in natural repetition or other oral-testing scenarios. This potential research application, however, also creates issues with naturalness and predictability, as pointed out above. Jessop suggests that corpus linguistics can resolve some issues of sparsity, which would also alleviate naturalness concerns for EI repetition and the use of contrived EI items for global oral-proficiency investigation.

Previous investigation into using corpus resources in order to identify target features inside of naturally occurring sentences has shown that extracting items from real-world utterances significantly impacts the EI test results and their correlation with other oral-proficiency testing methods (Christensen *et al*. 2010). Beyond enhancing direct-object predictability, extracting sentences from existing corpora provides for items that are less stilted and contrived. As demonstrated in the Valian *et al*. (2006) study for L1 learners, the less natural and predictable the item, the more the subject must memorize the stimulus. This effect apparently applies similarly for L2 EI test items.

The tools discussed in our earlier work (Christensen *et al*. 2010) also make it much easier to generate test items. As already discussed, much of the focus on EI has been regarding particular grammatical features, lexical entries, morphology, etc. In order to elicit the desired

linguistic features, linguists would be required to arbitrarily generate sentences which, besides the side-effect of producing often being strange and improbable items, took a nontrivial amount of time to create. The item-generation tool thus allows for fast and targeted test-item generation from a large test bank of EI items. This serves to move toward easier test automation.

2.2 Language-Test Modality Correlations

A long-standing complaint about the EI test from linguists is that the test is simultaneously a listening or comprehension test as well as a speaking or production test, as depicted in Figure 1. This indictment initially slowed the interest in the use of the test substantially. This overlap of modalities thus makes the results too ambiguous as to whether the subject's comprehension or production is tested (Hood and Lightbrown 1978, Vinther 2002, Jessop *et al*. 2007). The counterargument holds that the need to disambiguate is not pressing since the test can still be used to accurately approximate global oral proficiency. A more detailed and nuanced understanding of the production/comprehension overlap in EI is, however, desirable.

There is significant prior work investigating the acquisition of varying language modalities in both L1 and L2. Payne and Whitney (2002) showed that development of chatroom-writing skills produced a substantial gain in oral skills. Feyten (1991) analyzed both the correlation of listening skills and language achievement in general and oral proficiency in particular, as well as the ability to predict language achievement based solely from listening skills. Feyten points out that there are different levels of listening and that they serve different purposes. His discrimination among listening skills is easily understood by contrasting the type

of listening required to do a task such as EI versus the listening skills required to answer comprehension questions.

Despite the recent surge in interest regarding EI, little work exists disambiguating the nature of the test as either a production- or comprehension-focused test. The majority of work has focused on the production aspect of EI (Hakansson and Hansson 2000, Fujiki and Brinton 1987), despite the original objects of researchers as to the dual nature of the test. Because the written modality of language can also be classified under production or comprehension (reading vs. writing), the correlation of EI with non-oral modalities could also provide helpful evidence as to just how much the correlation between oral tests can be relied on to accurately profile the comprehension and/or production measure in the EI test.

Additionally, although EI has no written component, EI is also unique because of its strong focus on grammatical structures and lexical items, otherwise difficult to elicit during an oral test (Naiman 1974). This gives it potential overlap with components from writing and reading modalities as well, modalities in which grammar and vocabulary are more often the focus.

2.3 Elicited Imitation Scoring and Automation

Because of the unique nature of the EI test among oral language tests, scoring has evolved considerably through continued research. Initially items or tests were scored holistically, in the same fashion that scoring was done on more traditional oral tests (Keller-Cohen 1981). Subsequent research has identified more objective and standardized methods of scoring.

Graham (2006) proposed a syllable-based scoring procedure in which each syllable in the test item is given a binary score of correct or incorrect. The cumulative test-item score is then

calculated by summing all the correct syllables in the utterance. Subsequent research has strongly validated this method of scoring as highly internally consistent among human graders (see Lonsdale *et al*. 2009). The development of various tools to automate the scoring of EI using this syllable-scoring method, including the development of an aid to human graders and a process for fully automated scoring via automatic speech recognition (ASR), has yielded promising results.

The first attempt at fully automatic scoring of EI test items was documented by Graham *et al*. (2008). This study demonstrated the ability of an ASR engine to return scores for EI test items that are highly consistent with the scores returned by a human scorer. Using the SPHINX ASR engine (Lee 1989) and custom recognition grammars designed for use in EI scoring, the researchers reported a correlation of 88% between ASR and human-scored items. These results highlighted the potential to achieve a fully automatic EI test. A more detailed discussion of the grammar and process will be given in Chapter 5, as this process was also implemented to score the EI tests used for this study.

More recent work has moved researchers closer to full EI test automation and real-world application. A few of these studies are as follows:

1. Cook *et al*. (2011) use EI test results to automatically predict OPI scores for English.
2. Lonsdale and Christensen (2011) propose a system of machine learning that will identify the most ideal next item for a student given past responses, for implementation in a predictive test environment.
3. Matsushita (2011) and Millard and Lonsdale (2011) discuss the implementation of ASR for EI for Japanese and French respectively.

Despite these advances, much of this research into the potential for using elicited imitation tests for assessing global oral proficiency has focused largely on methods of scoring

and demonstrating that there is a high correlation between EI test results and more standard proficiency measures, such as the oral proficiency interview (OPI), provided by Language Testing International (LTI), which has standardized guidelines that target particular language features for assessing oral proficiency (ACTFL 1999) . The goal of the studies cited above was to propose the EI test as a viable alternative to more expensive, slower forms of oral language testing. Until quite recently, however, these studies have not addressed a fundamental issue at stake in an EI oral test: namely, that oral fluency is a fundamental component of almost all other oral-language testing methods and receives significant attention when grading these tests (Housen and Kuiken 2009) and is completely absent in the scoring of EI tests, and arguably absent from the test as well. Some work in the area of researching fluency metrics extracted from EI test responses has proved unfruitful (Matsushita and LeGare 2010) because the speech samples were too short and lacked the characteristics of spontaneous speech.

Another study that examined pseudo-fluency features from EI items provided slightly better results, but not promising enough to justify hope that automatically generated scores could effectively be substituted for human grading (Müller *et al.* 2009). Müller, in advancing an alternative to the automatic scoring method proposed by Graham (2006), tried to incorporate both accuracy and fluency metrics in the calculation of a score for a test item. For their human-scored metrics, Müller and his colleagues attempted a dual-scoring holistic-style rubric, which is demonstrated in Figure 2. While this combination of scoring methods is attracting substantially more interest in current research and is the direction of this study, it is applied very differently. Müller's attempt to extract pseudo fluency features convolutes the nature of the EI scoring to some degree, and attempts to quantify fluency in test items that have been demonstrated to not contain sufficient fluency information (Matsushita and LeGare 2010). In practice, the differences

between the scales depicted in Figure 2 are insufficient to base both fluency and accuracy measurements on.

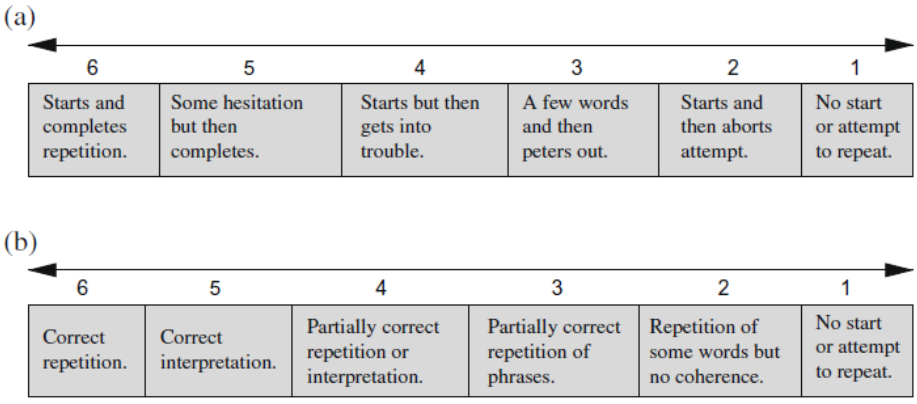


Figure 2: Grading scales from Müller *et al*. (2009) – (a) fluency of EI item, (b) accuracy of the item

In the last year, studies by Matsushita and others (Matsushita and Lonsdale 2012) have identified the need to incorporate other automated testing methods in order to increase the effectiveness of the EI test, and to more accurately mimic the global nature of interview-style exams.

2.4 Simulated Speech

Another language-testing method that has widespread application and has received significant attention within the last few decades in respect to automatic scoring is referred to as a semi-direct oral test (O'Loughlin 2001), or simulated speech (SS). This method of testing historically incorporates technology in various ways. Originally, SS tests could be administered via tapes (Malabonga *et al*. 2005). In most modern applications that implement the SS test, the test stimuli are delivered via computer (Malone 2007). The application then records the subject's response (as well as other peripheral information) to a sound file to be scored later. The stimuli

will give the examinee situations in which he or she will respond appropriately with a description, response, or other speech act (Bernstein *et al*. 2010).

Findings have indicated that SS tests also correlate strongly with OPI-style tests. One SS-style test—called the Simulated Oral Proficiency Interview (SOPI)—reports correlation of between 0.89 and 0.93 (Clark and Li 1986, Shohamy *et al*. 1989). Despite these promising correlation statistics, other researchers have highlighted the differences—such as the lack of turn-taking, and discourse strategies (Koike 1998)—between the normal interview-style test and SS-style tests.

The SS test shares some of the advantages of the EI test while maintaining desirable characteristics of face-to-face tests, most notably that of requiring the subject to produce spontaneous speech. Applications of this testing method are currently much more widespread than those of EI. Currently, the internet-based TOEFL test (iBT), the SOPI, and the Computer Assisted Screening Tool (CAST) all make use of the SS testing methodology (Clark and Li 1986, Malone 2007).

The advantage of SS testing is its utility in providing a simplified and more standardized administration procedure. Fewer graders are required because tests can be graded remotely. As already mentioned, substantial work has already been done to increase the possibility of incorporating automatic-grading capabilities in these types of tests. While some aspects of oral language are typically omitted in this type of test (such as turn-taking, discourse management, etc.), this type of omission must be expected in tests where interaction with another speaker is not available. The decision as to whether the cost/benefit ratio of losing some linguistic knowledge about a subject while gaining the possibility of increased test automation must be made on a case-by-case basis.

2.5 Simulated Speech and Elicited Imitation as Test Battery

As documented in 2.1 sections and 2.2, both the EI and SS tests offer numerous advantages. However, both come with significant drawbacks for linguists attempting to create a fully automated language-testing method. While significant advancements have been made in using both testing methodologies independently, a strong argument has been made in favor of combining these tests to provide better global assessment of oral proficiency.

The strength of the EI test includes its utility in eliciting particular linguistic structures and lexical items otherwise difficult to obtain. It also allows for fully automatic grading based on current ASR technology which focuses totally on the accuracy of the utterance without requiring any judgment of quality or fluency (Graham *et al*. 2008). Because of the unique nature of the EI test, various linguistic features can be investigated more in depth in the EI test by requiring the subject to try to mimic the stimulus regardless of whether he or she has acquired the lexical item or grammatical structure in question (Naiman 1974).

On the other hand, the SS test permits the subject to produce spontaneous speech and produce linguistic cues not evident or available in the EI test. The SS test has been widely researched with respect to fluency features. Numerous features have been identified as effective indicators of oral proficiency that can automatically be extracted from the SS results (Ellis 1993, Laver 1994, Chambers 1997). By using the extracted fluency features to assess overall fluency in the language and using EI as an oral accuracy measurement, linguists can create a more complete and sophisticated type of automated language assessment. This testing battery is particularly attractive because administration and scoring can be done objectively and without any manual processes.

This combination was proposed by Matsushita (2011). Though both tests are used independently in the testing community, using them in concert to predict global oral language proficiency is an innovative approach. Matsushita investigated this testing combination in Japanese and found that it provided a significantly better correlation with the OPI than did either test independently. With these encouraging results, the same technique merits investigation in English in order to compare and validate and conclusions drawn from the Matsushita study.

2.6 Extracting Fluency Features

Significant work has gone into identifying what features of oral language are indicative of global oral proficiency and fluency, or, for our purposes, fluency features. Recently, researchers have been most interested in which fluency features can be extracted via automated techniques. Two main methods of extracting features have emerged: feature extraction via ASR, and feature extraction via a signal-processing tool.

The nature of the features extracted through these two methods differs. ASR-based features vary widely; however, they all reflect the capacity of the ASR acoustic model and language model either to provide insight into the subject's fluency as indicated in acoustic scores and/or language scores for words, or to provide high-level understanding of the subject's abilities via the recognized output produced by the whole of the ASR engine (Müller *et al*. 2009). This use of ASR output always relies on data beyond the textual output of the system, such as time-stamps and then uses various post-processing techniques to calculate metrics for features that apply to fluency (Cucchiarini *et al*. 2000, Neumeyer *et al*. 2000, Xi *et al*. 2008). Most systems rely on a combination of these in order to give the most complete understanding of the subject's overall fluency.

The difficulty with the ASR approach lies in the fact that ASR is imperfect technology. Any system that relies on the output of the models is assuming that the engine gives a more-or-less accurate representation of the original utterance or—at a minimum—that the results are at least predictably inaccurate. However, the underlying complexity of ASR technology introduces significant variability. From the mapping from acoustic signal to candidate phonemes (via acoustic models) to mapping phonemes to dictionary entries (via specialized ASR dictionaries) to mapping candidate dictionary entries to strings of words (via language models), the output is too often unreliable in unpredictable ways.

ASR research has made significant strides in open-vocabulary language recognition in the past decade; however, many of these advances require custom adjustments for the acoustic model for the speaker or a customized language model for each speaker. These adjustments are difficult but not completely unfeasible, but the added complexity of non-native speakers makes high recognition quality nearly impossible. Despite these difficulties, the implementations of fluency-feature extraction systems with ASR technology have proved successful in approximating either global or target areas of oral proficiency (Ginther *et al*. 2010).

Signal-processing tools have also been used to extract fluency features with favorable results (De Jong and Wempe 2009). The PRAAT tool in particular has found wide-spread use in feature extraction from sound files (Préfontaine 2010). These systems differ in that they rely on no underlying models to correctly, successively map output to input. The approach is simple and straightforward in that the features are calculated by analyzing the acoustic signal for silence, voicing, syllable nuclei, etc. as determined by heuristics.

Both methods of feature extraction have provided excellent results in measuring fluency and estimating characteristics of oral proficiency. However, the fundamental differences in

approach highlight the need to identify which extraction technique provides the optimal results for use in automated scoring.

Chapter 3 – Elicited Imitation vs. Other Language Tests

One of the most fundamental questions facing linguists using the elicited imitation (EI) testing methodology in the past has been the fundamentally dual nature of the test. Because the test involves listening to and presumably comprehending the test item and then repeating or (re)producing it back, it is unclear whether the test scores more fully represent the comprehension ability or the production ability of the subject (Hood and Lightbrown 1978, Vinther 2002, Jessop *et al*. 2007). Most studies have assumed that production is the more important ability tested in EI. Despite the ubiquity of the discussion about the production and comprehension duality of EI in the literature, opinions differ about the significance of this distinction. As expressed in Chapter 1, this work aims to address the question of the role of EI among other language tests by comparing the scores of other language tests with those of EI.

In recent decades, pragmatists have demonstrated that, regardless of which capacity is most represented in test scores, EI scores correlate well with scores from other global oral-proficiency measures. Vinther (2002) argued, however, that it would be possible to attribute improved EI scores to listening comprehension if a subject had received listening training. Vinther also points out that a subject with good listening-comprehension skills but bad production skills and a subject with poor listening comprehension skills and good production skills could end up with the same results from an EI test. Naiman (1974: 1) makes the clearest statement on the distinction by stating that EI is a "conservative estimate of second language comprehension skills and a non-conservative estimate of second language production skills."

A clear elucidation of the relationship of EI with a listening exam, and the relationship of EI with another speaking exam is necessary to better clarify the role of EI in comprehension or

production testing. Various experimental methods could be employed to better ascertain the focus of EI. However, for this study, I will examine EI by comparing test results from various testing modalities with results from EI in an attempt to gain greater understanding from a global perspective. While it is safe to assume that the production vs. comprehension modality question can be investigated by contrasting results from aural and oral exams with results from EI, it is also possible to gain insight and additional evidence of the focus of the EI test by investigating the correlation—or lack thereof—between the scores from EI and the scores from parallel textual tests, such as a grammar (writing) test, focusing on production, or a reading test, focusing on comprehension. One of the strengths of EI that is often emphasized is that it gives researchers a way to access grammatical knowledge (Vinther 2002, Jessop *et al*. 2007). Thus one would expect the correlation between EI and grammar tests to reflect a greater overlap than would be accounted for by the similarities in their focus on language production.

3.1 Methodology and the Data

In order to better understand the role and nature of the EI test, I compared scores from EI with the scores from four other language tests including speaking, listening, grammar, and reading. By comparing the scores of EI against those of these other tests for all the subjects, I obtained correlation statistics that help to better understand what the EI test measures and how similar it is with other oral or aural tests.

The testing data used to examine the relationship between EI and other test modalities were acquired from the English Language Center (ELC) at Brigham Young University, which is an English for Academic Purposes (EAP) institution. The ELC administers a battery of tests for placement at the beginning of semesters and a series of final exams at the ends of semesters. At

the end of the semester, the students complete a series of Language Achievement Tests (LAT), which include tests of grammar, reading, listening, and speaking (sLAT). An EI test is administered simultaneously. The reading, listening, and grammar tests are traditional fill-in-the-blank and multiple choice style tests, while the sLAT is a simulated speech (SS) style test where the students hear a stimulus and then respond in monologue fashion. These tests are given as semester-ending achievement tests designed to measure improvement and achievement. Tests are designed to adequately test all levels of language learners – from basic to academic level – with scores placing the subject somewhere on the scale between those extremes. The scores used for this test were well distributed between the beginning level learners and academic speakers, with scores for comprehension test slightly higher on average than for the production-focused test.

For this study, data from three semesters were used, with the number of students having completed the tests for each semester running between 169 and 190, for a total of over 500 student tests. Each student took all five tests, though not all students successfully completed all tests. Additionally, some grading data were not available; therefore, the total number of student tests for each test analyzed here is 492.

Both the reading and listening tests are designed to target comprehension skills. The grammar and speaking tests are more geared to measure production. Each test is scored automatically. The descriptive statistics for the data are shown in Table 2.

Table 2: Details of the data used for comparison of EI and other test modalities

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
sLAT	492	1	7	3.99	1.038
Listening	492	233	853	671.74	93.300
Grammar	492	100	1100	629.41	151.489
Reading	492	100	806	615.19	96.498
EI	492	0	523	376.63	54.753
Valid N (listwise)	492				

The EI scores used for comparison in this study are those procured via ASR scoring. As mentioned in Chapter 2, significant previous work has focused on the ability of ASR to score EI accurately (Graham *et al*. 2008). In order to validate this method for the scenario under consideration, I undertook a small validation experiment from a subset of the data.

The work reported on in Graham *et al.* (2008) found that the correlation of ASR-scored EI items and hand-scored items was r = 0.88. Using the Julius recognition engine[1] after investigating various techniques to build language models specifically for the test in question, Matsushita (2011) reported correlation of r = 0.91 at the item level for Japanese EI. These studies provided the impetus for significant further work in automatic grading of EI.

In order to obtain ASR-scored data for the tests included in this dataset, I implemented the same framework that was established for the prior work discussed previously in Chapter 2. In order to verify the prior work in the context of the ASR-scoring results for the data used in this work, several scorers supplied by the Pedagogical Speech and Software Technology research

[1] Available at http://julius.sourceforge.jp/. See also Lee and Kawahara (2009) for more technical details.

group hand-scored the data for one of the semesters. I then compared the hand-scored results and the ASR-scored results from the corresponding semester. A correlation coefficient of r = 0.84 was obtained for subject-level scores. Although this coefficient indicates that the level of correlation in this study is not quite as high as has been reported in other seminal studies in this area, it is sufficiently high to deem the findings of the previous work pertaining to the automatic scoring of EI as a legitimate scoring methodology as relevant in comparison with hand-generated scores in this analysis. As the improvement of EI scoring techniques is not the focus of this study, I conducted no further investigation. Figure 3 shows the scatterplot for the hand-scored and ASR-scored subject scores for the winter 2011 semester. Points lying below the best-fit line denote subject-level scores that are lower for the ASR-grading method relative to the hand-scoring method, and conversely, the points above the line denote subject-level scores higher for ASR-grading relative to hand-scoring results.

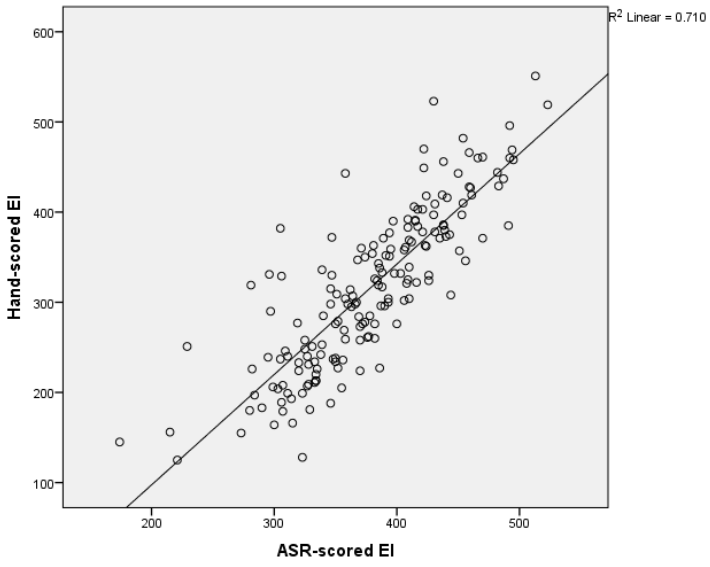


Figure 3: Scatter-plot showing correlation of ASR and hand-scored subject-level EI scores

3.2 Production vs. Comprehension

Modality distinctions in language are made in various ways for different fields of linguistic research—written and spoken, production and comprehension distinctions being the most common (Feyten 1991, Hakansson and Hansson 2000). EI is unquestionably categorized under an oral modality. But because both comprehension and production are used in repeating the stimulus sentence, it remains largely unclear as to whether to classify EI in the production modality or in the comprehension modality. I investigated this distinction in two ways. First, I compared the correlation of scores from EI with scores from sLAT, on the one hand, with the correlation of scores from EI with scores from the listening LAT, on the other. Second, I used a regression model to compare the scores from both production tests (the sLAT and grammar) and from both production tests (listening and reading) with the scores from EI.

For the oral-aural tests, scores from the EI tests correlated slightly better with the scores for listening than with those from the sLAT. The correlation coefficient for listening scores and EI scores was $r = 0.534$ ($N = 492$, $p < 0.01$). The sLAT scores and EI scores returned a correlation of $r = 0.463$. Both of these correlations easily reach the level of statistical significance ($p < 0.05$ is used as the level of statistical significance in this work), indicating that a relationship exists between EI and both of these tests. Using the Fisher r-to-z transformation, I compared the correlation statistics and found that the difference between the correlation statistics does not, however, reach statistical significance ($p = 0.139$). This seems to signify that EI does not clearly focus solely on production or on comprehension. In Figure 4 the correlation of the EI and listening tests is depicted via a scatterplot. The scatterplot shows a definite relationship, but also demonstrates that the relationship is not tightly linear.

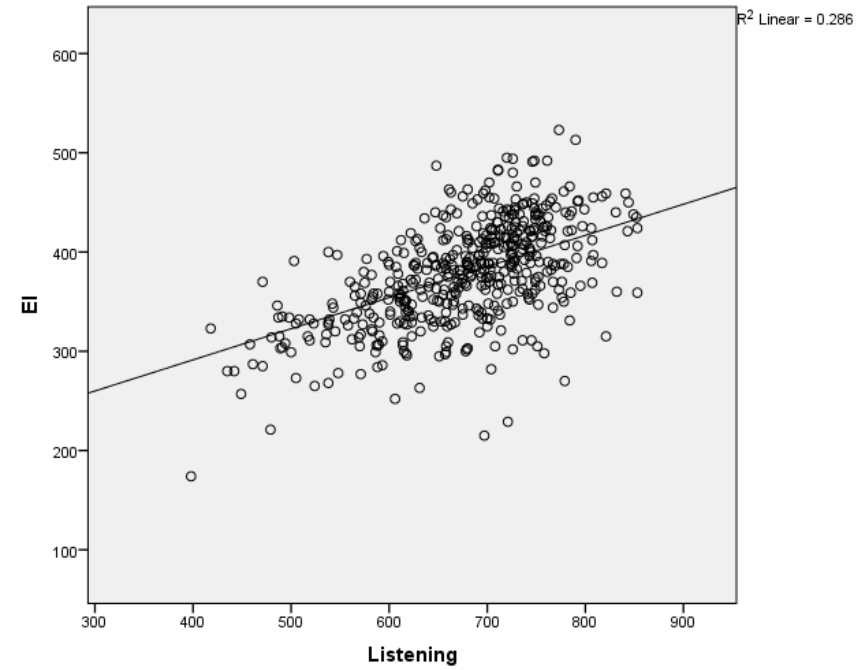


Figure 4: Correlation results of LAT listening for all semesters and EI test score

In order to further investigate the production/comprehension distinction beyond the spoken realm, I used multiple regression to investigate the relationship between EI and production tests (grammar/writing and speaking), and EI and comprehension tests (reading and listening). The models produced results as shown in Table 3 and Table 4 which demonstrates that while comprehension tests provide a slightly better model of the data, the results, once again, do not give conclusive evidence of a particular distinction in EI between comprehension and production. Therefore, no conclusion about the focus on the test can be made. Both regression models demonstrate a similar R value (both R values are statistically significant ($p < 0.01$), indicating that both production and comprehension elements of language play similar roles as factors of the EI test. In the context of using EI as an oral proficiency measure, this could be seen as a less-than-desirable overlap as other oral language exams do not have this dual focus.

Table 3: Regression model statistics for EI and comprehension tests ($p < 0.01$)

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.535[a]	.286	.283	46.362

a. Predictors: (Constant), Reading, Listening

Table 4: Regression model statistics for EI and production tests ($p < 0.01$)

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.519[a]	.270	.267	46.885

a. Predictors: (Constant), Grammar, sLAT

3.3 Grammar Test

Researchers have long used EI tests to assess the acquisition of grammatical features (Jessop *et al.* 2007). Because EI requires subjects to produce and/or comprehend grammatical structures regardless of whether they have learned the structure yet, the EI test, in theory, is somewhat analogous to a grammar test. Traditional grammar tests are textual, and therefore should still be starkly different from an oral-aural test, regardless of the grammatical focus. However, if the focus of EI is truly testing a subject's grammatical ability, one would expect a stronger relationship to exist between EI and grammar tests than between EI and textual tests focused on other aspects of language. Thus, the correlation between scores from the grammar LAT and the scores from EI should be significantly different from the correlation between the scores from the reading LAT and those from EI.

In fact, the scores from the reading test returned a correlation with the scores from EI of $r = 0.296$, which is well above the level of statistical significance ($N = 492$, $p < 0.01$). The scores from the grammar test, however, correlated even more strongly with the scores from EI, with an r value of 0.380 ($N = 493$, $p < 0.01$). Although the difference in correlations did not reach statistical significance ($p = 0.139$), the difference in correlations is greater than the difference in correlations for production and comprehension tests. Because this difference in correlations is below the level of statistical significance, no conclusions can be drawn from this about the ability of EI to test grammar acquisition in comparison to other linguistic modalities. However, more data could show this slightly larger correlation to be significant, therefore more research is necessary.

3.4 Discussion

This study demonstrated that EI cannot be satisfactorily classified as either a production test or as a comprehension test. Although much has been made of the need to disambiguate the dual nature of the EI test, it remains unclear how much this distinction matters in practice. It seems clear that the EI test is simultaneously a representation of some aspects of listening skills and some aspects of oral production. This duality has historically been the issue that has caused some to regard EI skeptically (Hood and Lightbrown 1978, Vinther 2002, Jessop *et al*. 2007). However, despite the conceptual ambiguity inherent in this test, it seems abundantly clear from the body of research that EI can be used as a good approximation of global oral proficiency.

The use of EI as a grammar-testing tool seems to be trending toward validation by this research but it is certainly not the case that EI is a strong indicator of grammar test scores. Although the EI test that was administered in these exam iterations was not specifically designed

to focus on grammatical acquisition, a moderate correlation (usually defined between 0.3 and 0.7) still exists between results from a traditional grammar test and those from EI.

Chapter 4 – Extracting Fluency Features from SS Test

4.1 The Data

The simulated speech (SS) data used to examine the optimal method of extracting fluency features were also included in the LAT testing described in Chapter 3. The sLAT is a computer-administered test that consists of a series of questions that prompt spontaneous speech responses – thus an SS test. The responses are recorded by a testing application at the ELC and then given a holistic score by a human grader on a scale of 1 to 7. Each test is double-scored by raters at the ELC according to a grading rubric provided at the ELC. The test-administrator then runs the scores though Facets (Many-Facet Rasch Measurement) for an in-depth analysis of rater bias. Finally, a weighted average score is assigned each test. The test files in this study ranged between 20 seconds to just under 2 minutes, and the average length of file for each student was between 50 seconds and a minute.

For this study, I examined data from all three of the 2011 sLAT test administrations. For each semester included in the data, the test was administered to between 179 and 196 students. The test results for the sLAT constitute one holistic score for the entire 10-item test. For both the ASR and Praat feature-extraction systems, I calculated sLAT fluency features for both the ASR and signal-processing methodologies for approximately 500 student tests (5000 sound files).

4.2 Feature and Tool Selection

In order to set up the dichotomy between fluency features extracted via an ASR engine and those extracted via low-level signal processing objectively, I first tried to identify the most prevalent and successful features used in similar studies. A list of some of the identified features from similar studies is shown in Table 5. A more comprehensive list of common fluency features

extracted from an ASR-based system is outlined in Matsushita (2011). Likewise, for signal-processing-feature extraction, a variety of features are listed in Table 5 as well, though the possibilities of features extracted via signal-processing methods are much more limited.

Many of the features extracted, either via ASR or via signal-processing, can be more accurately quantified by human graders. However, some features presumably cannot be assigned by a human grader, such as articulation rate or acoustic model. In either case, for the hand-scoring of SS tests, one can assume that the grader does not consciously quantify these metrics but rather takes into account an abstract representation of the subject's fluency comprising some combination of these metrics and other factors in the subjective assignment of a grade. The use of these computationally extracted fluency features can therefore be viewed more as an attempt to quantify the perception of a speaker's fluency. The accurate identification of the most influential and discriminative features is consequently of utmost importance.

Although a variety of features exist for both signal processing and ASR, these features by and large map semi-directly from tool to tool. For example, the token count of words that can be calculated via ASR maps straightforwardly to a syllable count (Graham *et al*. 2008) which can be extracted via signal processing. This similarity in features further highlights the need to identify which of the tools is better at extracting features of comparable types. This work seeks to directly answer the question posed in Chapter 1: Which tool is ideal for extracting fluency features from SS test result files?

Common ASR-based Fluency Features	Common Signal-processing-based Fluency features
Acoustic model score	Pausing or silence information
Language model score	Speech rate
Pausing or silence information	Articulation rate
Articulation rate	Total file duration
Length of speech runs in words or time	
Unique number of words (types)	
Total number of words (tokens)	

Table 5: Brief compilation of common features based on numerous ASR studies

I chose two studies in particular to model – Matsushita (2011) and De Jong and Wempe (2009). This modeling ensures that the results from the comparison should be an accurate representation of the utility of ASR versus signal-processing in real-world applications. The features selected for use in this study were those used by Matsushita (2011), with the exception of the number of fillers, or disfluencies consisting of the subject's uttering a meaningless word or sound. Although the filler feature proved significant in the results as the fifth most influential variable of the eleven used in of Matsushita study, some doubts about the usefulness and relevance of fillers have been raised in other studies. In particular, Ginther *et al*. (2010) enumerate the uses of fillers in ways that are not detractors from fluency but instead serve pragmatic functions in language. Also, the ability to extract fillers is complicated in this study as a result of the varying L1s of the subjects. In light of such prior research and other concerns, I have omitted this feature from this study.

I selected features for the signal-processing-based method based on the study by De Jong and Wempe (2009), including the additional features added from more recent work based on their findings. It should be noted that the De Jong and Wempe study was focusing on the

potential for using these fluency features to quantify speech rate in particular and to compare the automatically assigned speech rate score with scores given by human graders. While speech rate has been shown to be a strong indicator of oral proficiency in many studies, that was not the focus of these researchers. Table 6 below shows all the features used for both the ASR and signal-processing systems.

Table 6: Fluency features used for Sphinx and Praat systems

SPHINX Features	Praat features
Speech Time Per Run	Number of Syllables
Speech Rate (number of phonemes per second)	Number of Pauses
Word Types/Speech Length	Duration of File
Tokens Per Run	Phonation Time
Speech Length	Speech Rate (number of syllables/duration of file)
Silence Length	Articulation Rate (number of syllables / phonation time)
Number of Word Types	Average Syllable Duration (speaking time/Number of Syllables)
Number of Word Tokens	
Number of Runs	
Number of Pauses	

I used the Sphinx ASR engine built at CMU for the extraction of fluency features. Although this was not the same ASR engine used by Matsushita (2011) (which was the Julius ASR engine for Japanese), it was preferable to maintain one ASR engine for use in this work. As already documented, the SPHINX system has been tuned extensively for work with the EI system; therefore, it was deemed easier to adjust the SPHINX system to extract the necessary SS

features than it would have been to adjust the Julius engine to provide EI-item syllable scores in English. Differences between the Julius and SPHINX engines should not yield results substantially different from those in the Matsushita (2011) study. SPHINX is recognized as one of the premier open-source ASR systems and is used in many similar applications and thus is a natural choice for this study. I contrast the feature results obtained from SPHINX against those extracted via Praat tool, which served as my signal-processing component, and was also used in the De Jong and Wempe (2009) study. Praat is an open-source signal-processing and acoustic-analysis tool developed at the University of Amsterdam (Boersma and Weenink 2005).

Once the features are extracted from both of these systems, I used machine-learning and statistical modeling in order to determine which set of features better explains the scoring data. This analysis addressed the third research question posed in Chapter 1: Can machine-learning and statistical techniques utilize the fluency features that are extracted in order to accurately predict holistic SS scores? For the machine learning (ML) component, I used both the ASR and Praat features listed in Table 6 to predict sLAT scores. The role of ML in computational linguistics has grown steady in the last decade, and ML here serves as an effective means of obtaining item-score predictions based on a feature set. In this case it serves the dual purpose of predicting the sLAT scores and providing a common framework in which to compare the results produced by each system. For the ML component in this study, I used the Tilburg Memory Based Learner (TiMBL) to analyze the features and to identify the features that lead to a more accurate prediction of the human-assigned sLAT score. The TiMBL program is commonly used in the ML field, though usually in the context of language analysis and related issues (Daelemans *et al.* 2010).

For the statistical modeling of the features, I calculated correlations and regression models that demonstrated the relationship of fluency features with the sLAT score. By investigating the features via both ML and statistical modeling, I was able to determine which features are the most representative of oral proficiency.

As previous mentioned, the SPHINX system is currently used in many language-testing applications (Mostow and Aist 1999). Because the system allows for custom-built language models and acoustic models, the ASR engine can be tuned to optimize performance for any given task. As previously discussed in the explanation of the choice of the SPHINX engine for this study, substantial work has gone into creating the optimum scoring procedure for EI test items using a custom grammar, a procedure that has produced excellent results. However, for SS only generic components were employed for two reasons: first, the data are not known *a priori* and therefore represent an open-vocabulary type recognition task that makes custom development of components time-consuming, and second, generic components are used in other similar studies.

4.3 ASR Fluency-Feature Extraction

In order to extract the fluency features from the sLAT test items, the files were first converted from the .aiff audio format to .wav format – the type of audio format that Sphinx recognizes natively. The input audio also had to be normalized to 16-bit 16000 Hz mono. For the language model (LM), there were a few publicly available options that that I tried, including the Hub 4, GigaWord, and Wall Street Journal (WSJ) models. There were also various acoustic models to experiment with, such as the Communicator, Hub 4, and WSJ models. The best configuration was determined qualitatively by running several files through the engine and

comparing the results to the sound file manually. While this process was not ideally automated or empirically verifiable, it was necessary in the absence of the requisite body of sLAT transcription data and needed only to be completed once in order to obtain the results for all the data. The ideal configuration of the system was a combination of the Hub 4 language model and the WSJ acoustic model.

I had to resolve several questions also, such as, what length of silence constitutes the end of a speech run? What length of silence duration as marked by ASR stamps should be included in the total silence duration? Previous work (Freed *et al*. 2004, Matsushita 2011) has used a 400 millisecond boundary for the minimum silence duration to separate continuous speech runs. This boundary was also initially used in this study. But when I investigated a portion of the data empirically, a shorter minimum seemed to reflect human-perceived pauses better and also improved results; accordingly, I shortened the silence duration employed. I omitted from silent-feature calculations any long silences at the beginning or end of the sound files.

The output of the SPHINX ASR engine can be given either at the word level with timestamps, or at the phoneme level with timestamps. In order to compare this work most closely with that of Matsushita, I used the word-level timestamp output. The recognized results of each file produced text files with time-aligned text results, as shown in Figure 5. As evidenced from this figure, the ASR results have an incredibly high word error rate (WER), which one could assume would affect fluency features based on ASR type count, and potentially token count. The high WER is indicative of the difficulty of transcribing non-native speakers in an open-vocabulary scenario. As will be demonstrated later, the effects of the high WER do not seem, however, to adversely affect the ability to get relatively accurate fluency features and promising results.

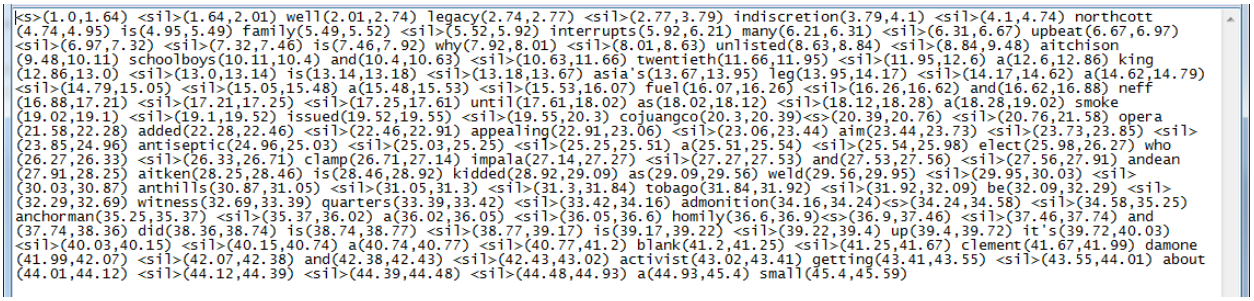


Figure 5: Sphinx ASR time-aligned word output

With the output for all the data, I ran a post-processing script that I created over the text files in order to calculate the metrics for each test item. I then aggregated the item totals on a per-student basis, which allowed me to create a feature vector for each student that could be used in the ML components. Each item vector consisted of the 10 aggregate ASR fluency features listed in Table 6 for the student along with his overall sLAT score.

4.4 Praat Feature Extraction

The construction of the Praat component required only a little adjustment to the updated script made available[2] from the De Jong and Wempe study (2009). The adjustments consisted solely of minor changes to the file processing in order to navigate the directory tree and file structure of the sLAT repository easily. The main disadvantage of this script is that it currently cannot be run in console mode, and therefore requires the Praat program to be open and running. This disadvantage hinders the current automation potential. However, the time required to perform the feature extraction is considerably less than that required to employ the ASR component. This time differential is a serious consideration when discussing a broadly

[2] https://sites.google.com/site/speechrate/

implemented testing system with potential for calculating real-time scores. The Praat program also requires fewer external resources, such as the language model and the acoustic model.

The configuration of the Praat script requires manual calibration of settings, such as the minimum threshold length for silence, a decibel threshold tuning parameter (defining silence within a speaker's utterance), as well as the minimum decibel dip (defining the distinction between syllable peaks). These settings were calibrated as follows: three-tenths of a second as the minimum length of silence, -25 decibels as the tuning parameter for silence, and 2 decibels as the minimum dip between syllables.

Once configured, the script processed all of the sound files and printed the fluency feature results in the Praat script output window. Figure 6 shows the Praat output text for a few of the sound files. Once all the files were processed, I ran a post-processing script over the results, a script similar to the script necessary to process the SPHINX output.

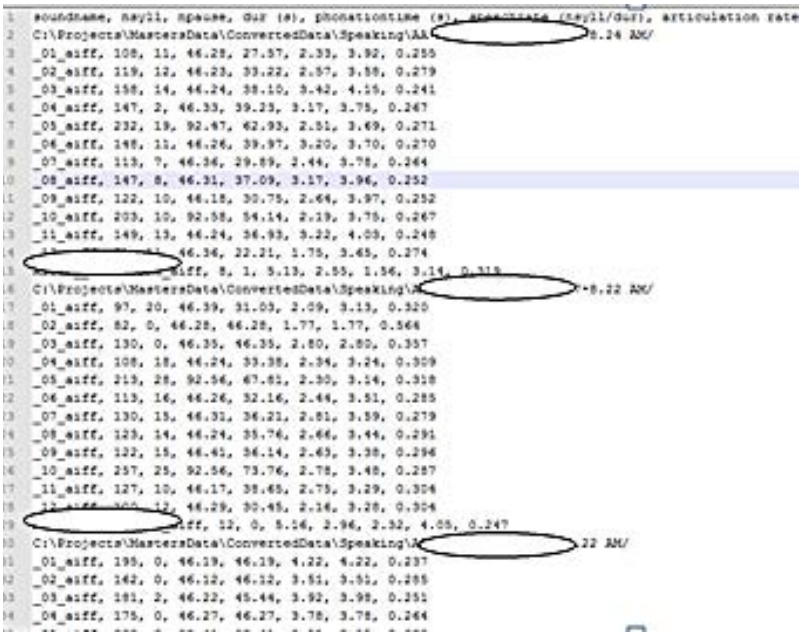


Figure 6: Praat script fluency feature output (student names removed)

4.4 ML Results

I ran both sets of feature vectors through TiMBL and obtained test prediction accuracy scores via the leave-one-out method of prediction. The ASR training file consisted of 484 vectors consisting of the ten fluency features extracted from the ASR transcription results, while the Praat training file had 536 vectors consisting of the seven fluency feature extracted from the sound files, including a few sounds files that were not successfully recognized by the ASR system. The TiMBL system calibrates a model in order to predict future results. The results returned the predicted score given the model calibrated with the fluency features. The accuracy of the model is then scored by comparing the actual outcome versus the predicted outcome of the SLAT. TiMBL also includes ranked features in the results, showing the relative information gain provided by the features incorporated in the model, or the quality of the feature in assisting in accurately predicting the correct outcome. The outcome for both the ASR and Praat features is depicted in Table 7 below. Besides determining the exact-match score showing the number of predictions that were entirely accurate, I also calculated the within-one (or adjacent-score) accuracy. The scores used to train the system in this case were weighted averages, which were used because human scores often differ by a point or more. This additional margin for error is consistent with human-rating scores. While no system's exact accuracy was as high as the results reported in Matsushita (2011), some degradation in prediction accuracy must be expected as the sLAT grading score is a 7 point scale (7 outcomes) as compared to the 3 or 4 level scale (3 or 4 outcomes) being employed in that study.

Table 7: TiMBL accuracy predication rates for fluency features

	ASR	**Praat**
Exact Accuracy	0.3908	0.3645
Within-one Accuracy	0.8376	0.8299

The results demonstrate that both ASR and Praat achieve exact accuracy above 30%. That increases meaningfully for within-one scores to an accuracy of between 83% and 84% for both systems. Based on a statistical analysis of the correlations of the predicting scores from the two models, these differences are statistically insignificant ($p = 0.69$) showing that both models are equally good at predicting sLAT scores based on their respective features. The results also demonstrate that for within-one prediction, ASR and Praat prediction of sLAT scores gives reasonably good results and either should be considered a prime candidate for the automation of fluency-feature extraction techniques.

Interestingly, the ranking of the variables and the information-gain supplied to the model (that is, the gain in the statistical power of the ML model to predict outcomes accurately) for the ASR features was quite similar to that reported by Matsushita. Despite the use of the different ASR engine and the differing target language and L1 backgrounds of the subjects, the same variables proved to be the most discriminative among test takers as shown in Table 8. Although the order of the most influential variables was not the same, of the five most influential variables from that study, four appeared among the most influential variables for this study. The top five discriminative features in this study were: (1) Number of Runs, (2) Number of Pauses, (3) Number of Word Types, (4) Number of Word Tokens, and (5) Tokens Per Run.

Among the Praat features, speech rate emerged as the top discriminative feature. Other discriminative features included number of syllables, articulation rate, and average syllable duration. Another feature that overlapped with ASR features—number of pauses—proved significant in both models by appearing in the top three discriminative features in both ASR and Praat results. Not surprisingly, the simplest score extracted, total duration of the file, had the least discriminative effect on the predicted scores. Many of the features that have equivalents in the ASR feature set (e.g. number of syllables, number of pauses) proved influential in both the ASR and Praat features. One obvious exception was speech rate, which as calculated by Praat proved to be the most discriminative feature, but in SPHINX came up next to last in importance. This variation in feature importance could be a reflection of the quality of the speech rate as calculated by ASR versus by Praat, or it could just be a reflection of the inherent inaccuracy of the pseudo-phonemes used by counting letters in orthography.

Table 8: TiMBL results for fluency features

Variables by order of significance	
SPHINX Features	**Praat features**
1. Number of Runs 2. Number of Pauses 3. Number of Word Types 4. Number of Word Tokens 5. Tokens Per Run 6. Silence Length 7. Speech Length 8. Speech Time Per 9. Speech Rate (number of phonemes per second) 10. Run Word Types/Speech Length	1. Speech Rate (number of syllables/duration of file) 2. Number of Syllables 3. Number of Pauses 4. Articulation Rate (number of syllables / phonation time) 5. Average Syllable Duration (speaking time/Number of Syllables) 6. Phonation Time 7. Duration of File

4.5 Statistical Results

The results from ML contrast slightly with those obtained via a regression model. In Table 9, the model summaries show that the Praat features yield a slightly better regression model than the ASR features. Both of the models are statistically significant (ASR model: $F = 25.459$, $p < 0.01$; Praat model: $F = 41.536$, $p < 0.01$).

(a) ASR Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.586[a]	.343	.329	.912

(b) Praat Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.596[a]	.356	.347	.896

Table 9: The regression model statistics for the features from the two feature extraction systems.

The significance of the individual features and their respective impact on the model was also analyzed by t-tests for each feature. The results of the analysis are displayed in Table 10. For the ASR model, the features that reached the level of statistical significance ($p < 0.05$) are (1) number of word types, (2) silence length, (3) speech length, and (4) number of runs. The features that reached statistical significance for the Praat model were (1) number of syllables, (2) file duration, (3) articulation rate, and (4) phonation rate. These results overlap only partially with the ML results.

(a) ASR Feature Coefficients[a]

Model	Unstandardized Coefficients		Standardized Coefficients	t	Sig.
	B	Std. Error	Beta		
(Constant)	3.518	.312		11.271	.000
Speech time per run	-.022	.072	-.033	-.303	.762
speechRate	.000	.009	.001	.029	.977
typesDivSpeechLength	.051	.082	.076	.615	.539
tokensPerRun	-.045	.039	-.068	-1.172	.242
speechLength	-.107	.043	-.548	-2.490	.013
silenceLength	-.146	.035	-.582	-4.136	.000
numTypes	.117	.029	1.761	4.045	.000
numTokens	.030	.031	.510	.951	.342
numRuns	-.142	.060	-.648	-2.383	.018
numPauses	-.023	.016	-.250	-1.465	.144

a. Dependent Variable: sLAT FAIR

(b) Praat Feature Coefficients[a]

Model	Unstandardized Coefficients		Standardized Coefficients	t	Sig.
	B	Std. Error	Beta		
(Constant)	4.043	.715		5.653	.000
numSyl	.016	.004	.450	3.948	.000
npause	.020	.013	.119	1.631	.103
dur	-.089	.016	-.328	-5.445	.000
phonationTime	.038	.016	.268	2.309	.021
speechRate	-.099	.130	-.055	-.763	.446
artRate	.321	.117	.180	2.735	.006
ASD	-.100	.141	-.028	-.713	.476

a. Dependent Variable: SLAT FAIR

Table 10: Individual feature analysis in the regression models of (a) ASR and (b) Praat features

4.6 Discussion and Implications

A fundamental difference between the Matsushita study and this study lies in the values being predicted. Whereas this study had direct access to human-scored SS test items, Matsushita used the class level as designated by the placement procedure as the outcome variable in the feature vector through which he identified the most salient fluency features.

Because the dependent variable in this study (the sLAT score) is the same as the score that we want to more accurately generate, the outcome of the machine learning from this study directly yields a model that can be used to predict these scores for future SS tests. On the other hand, in the Matsushita study, the ML was used to identify which features should be used in another algorithm to directly score the SS items. For this study, the hand-graded sLAT scores provide more accurate assessment of oral proficiency as determined by this test than the class-level variables Matsushita was forced to use. Additionally, because the purpose of this work is to demonstrate the utility of EI and SS fluency features in better predicting the human-graded SS score, EI and fluency features can be used directly in ML and statistical modeling to demonstrate improved results attained by combining EI and fluency features which will be discussed in chapter 5.

Also, the granularity of these studies differed. Whereas Matsushita used fluency features at both the test-item level and at the subject level as feature vectors, I aggregated the scores from the full SS test and averaged the features to obtain one fluency-feature vector per student. This was done in order to maintain consistency in scoring granularity of the SS test. Each sLAT test at the ELC is assigned a single holistic score, and since fluency features can vary between sound files, I determined that it was advisable to aggregate the features in order to use the holistic score more appropriately.

The Praat results, while not quite as good as the ASR results, did not perform appreciably worse. This is important to note because, despite the numerous fundamental differences already discussed in Chapter 2 and above in 4.2, the model yielded results substantially above the baseline of chance (~14%) and quite well for within-one predictions. The additional features available to the ASR system did not, in this study, increase the utility of the ML model in correctly predicting the score. Advantages of an automated SS system that implemented a Praat feature extraction would include increased speed of extraction and simpler processing without the need of additional models.

The comparable nature of the available features in both systems is evident in these results. However, it is quite likely that a combination of the features from both systems would provide even better results. Although the majority of the Praat features are available in some manner via the ASR outputs already available, the diminished complexity and quicker access to the features present compelling reasons to use lower-level processing where available for feature extraction. Other fluency features available via the ASR system but not analyzed in this study may also provide improved results. The additional complexity of the ASR features appears to have been of no additional help in the correct prediction of SS scores.

Chapter 5 – EI and Fluency Features

So far, I have investigated the utility of the elicited imitation (EI) test as an oral production measurement and have weighed options for extracting fluency features from a simulated speech (SS) test for the purpose of combining EI and SS fluency features. The purpose of the EI test is to measure the linguistic accuracy of a non-native speaker. Because of the automatic grading available for the EI test, much of the burden of manual test grading can be alleviated. However, as already discussed, more spontaneous speech tests require more nuanced and complicated measures. Much of this complexity originates from the multifaceted nature of oral speech, complexity that makes it challenging to measure characteristics such as fluency and accuracy. By assessing the accuracy of the subject via EI and the fluency of the speaker via fluency features extracted from an SS test, researchers can rapidly and effectively assess two of the major speech characteristics used for rating global oral efficiency. As demonstrated in Chapters 3 and 4 respectively, EI and fluency features independently correlate moderately with the sLAT results. Because of the overlap between the results that these two automated tests yield, using them in concert should augment the correlation coefficient and prediction accuracy. My final research question, how do the SS and EI correlate, and does adding automatically extracted fluency features to EI better account for a holistic score assigned to an SS test then EI alone, is investigated in this chapter.

As indicated in Chapter 4, the two available extraction tools for assessing fluency features appear roughly equivalent in their utility; consequently, either one should yield similar results when combined with EI. Therefore, I will base my work in this chapter on the ASR results in order to more closely parallel the methodology employed by Matsushita (2011). Using

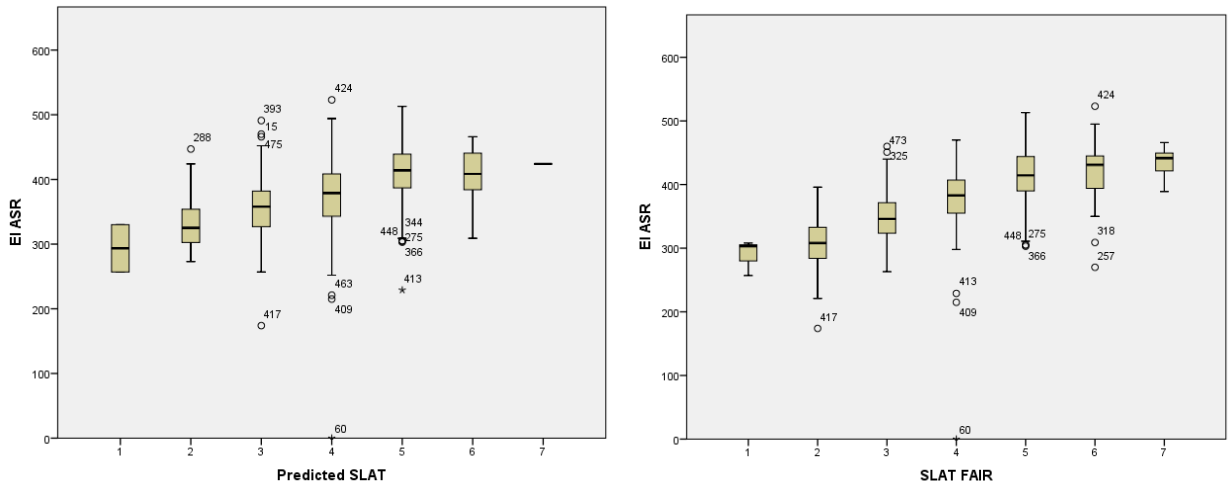
the ASR features, I will explore two ways to validate the combination of EI and SS fluency features: First, I will combine the EI scores with the fluency features and run them through TiMBL to demonstrate the superior predictive power of this procedure. Second, I will use regression models to show the significance of the EI scores alongside the fluency features in creating a stronger relationship with the sLAT scores.

These statistical and ML techniques are relevant here, just as they were in Chapter 4, because the EI score can be seen as another feature used in predicting the sLAT score. Because hand-scored results for the SS test are available, the regression model is applicable, because the combination of EI and SS fluency features will better reflect the FAIR average test score of the sLAT.

5.2 ML Results

Placing the ASR-generated EI scores in the feature vector of fluency features for each student had the hoped-for effect of dramatically improving the utility of the ML results in predicting SS scores. Not surprisingly, the EI score was the single most discriminative feature. The information-gain metrics were affected for the other features by the addition of EI, but the order of feature importance did not vary. The difference in the prediction accuracy was significant, with exact accuracy jumping to 49%, a 10% increase from the ML results reported in Chapter 4. The within-one accuracy reached more than 86%, a 3% increase. Figure 7 shows the hand-scored and predicted values of the sLAT test scores' relationship with EI. The boxplots reveal the similar predicted trends of the sLAT predictions vs. EI but also underscore the difficulty of predicting values particularly at the upper-end of the range.

The composite of the fluency features and the EI test yields a noticably improved reflection of the global oral profiency measure assigned manually. This validates the assumption that the information overlap between EI test scores and fluency features extracted from an SS test does not reach the level where no additional information about oral proficiency can be gleaned when the results from one are added to the other. Although the nearly 50% accuracy is still not at the accuracy level reported in Matsushita (2011), it does approach the human-agreement metrics for the scoring of the sLAT files.



(a) EI and predicted sLAT boxplot (b) EI and hand-scored sLAT boxplot

Figure 7: Boxplots showing the relationship between EI ASR versus predicted sLAT scores and hand-scored sLAT scores

5.3 Regression Results

Similar improvement of results is evident in the regression-model improvement. Table 11 gives additional model statistics. The overall improvement in model R^2 was 0.124. The difference in the R values is significantly higher with EI results included ($p < 0.03$) as

determined by the Fisher r-to-z transform. The R^2 value of this new model approaches 0.5, indicating the about half of all the variance in the sLAT test scores can be accounted for by EI and fluency features. As demonstrated with the ML results as well as this regression model, EI scores give significant additional information to the fluency features and improve the ability of the models to predict sLAT scores.

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.684[a]	.467	.455	.821

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	279.000	11	25.364	37.647	.000[b]
	Residual	317.998	472	.674		
	Total	596.998	483			

Table 11: Regression model statistics and ANOVA for ASR fluency features and EI scores

The analysis of the individual features of the model provides results similar to those for the model outlined in Chapter 4, a.467 model that did not include EI. As expected and demonstrated in section 5.1 in ML results, EI scores produced the most significant t value. Importantly, none of the significant features in the fluency-feature-only model were made obsolete by the addition of the EI scores. The significant features were reordered in their level of significance, however. As identified by this regression model, the order of the most significant fluency features (excluding EI) is (1) silence length, (2) number of runs, (3) number of word

types, (4) speech length. Table 12 shows the additional statistical-feature information for the regression model (compare with table 10a).

Coefficients[a]

Model	Unstandardized Coefficients		Standardized Coefficients	t	Sig.
	B	Std. Error	Beta		
(Constant)	.965	.402		2.401	.017
Speech time per run	.000	.066	.000	-.005	.996
speechRate	-.006	.009	-.035	-.750	.453
typesDivSpeechLength	.039	.075	.059	.513	.608
tokensPerRun	-.065	.038	-.090	-1.706	.089
speechLength	-.109	.039	-.561	-2.805	.005
silenceLength	-.106	.033	-.424	-3.249	.001
numTypes	.076	.027	1.144	2.851	.005
numTokens	.053	.029	.898	1.827	.068
numRuns	-.157	.055	-.710	-2.879	.004
numPauses	-.023	.015	-.249	-1.576	.116
EI ASR	.008	.001	.398	9.914	.000

Table 12: Regression model statistics for individual variables of the ASR-based fluency features combined with EI scores

The significant improvement of the regression model after the addition of the EI scores is evidence of the utility of the model in correctly predicting sLAT scores. Figure 8 plots the regression model predictions by their actual hand-graded sLAT results. Once again, the variance still present in the model is somewhat representative of human upper-bound on grading consistency for spontaneous speech tests, such as the sLAT. Despite the advanced statistical processing of the sLAT human-assigned grades which produced the FAIR results being used in

this analysis, a degree of variance is still to be expected because of the variance inherent in the human scoring.

5.3 Implications and discussion

The results from both the TiMBL ML system and the regression model validate for the English language the work done by Matsushita (2011) in Japanese. Although the process used for validation of EI and SS fluency features is considerably different than that used by Matsushita, the results are no less promising or relevant. The use of the fluency features and EI in this work to predict the FAIR sLAT score is comparable to their use for predicting OPI scores.

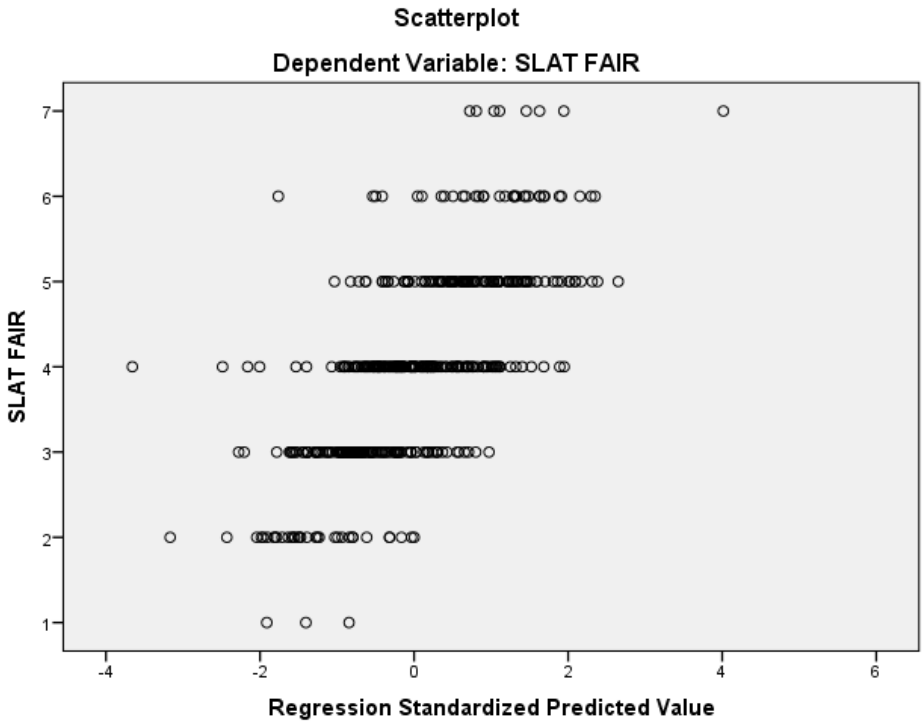


Figure 8: Scatterplot of regression model predicted values and sLAT FAIR scores

Both regression and ML appear to provide more than adequate means of utilizing and compiling the features and EI scores to produce a model that can be used to predict sLAT scores with accuracy that neither test could achieve independently. This independence from modeling technique further serves to validate these features as accurate broad-spectrum measures of global oral proficiency. Despite variable orderings of the features in their significance for the model, a representative model of oral proficiency can be created from fluency features and EI. Though fluency features for this analysis were limited to those identified in other studies, it seems clear from other research in area that other fluency features can be extracted and used successfully in the measurement of oral proficiency. The combination of these other features and EI may yet yield significantly better results. Because the results account for less than 50% of the variance in sLAT scores, additional work must be done to further identify significant features. The within-one scores do demonstrate, however, that the EI results and fluency features do give a good approximation of oral proficiency which could be used in lower-stakes testing scenarios.

Chapter 6 – Conclusions

The goal of this inquiry has been that of advancing linguistic understanding of the role and potential applications of elicited imitation (EI) in English and in second-language testing techniques. To achieve this end, my work focused on four research questions:

1. What information does comparing the results of EI with results of other language tests give in respect to better understanding the role of the EI test in language testing?
2. Which tool is ideal for extracting fluency features from SS test result files?
3. Can machine-learning and statistical techniques utilize the fluency features that are extracted in order to accurately predict holistic SS scores?
4. How do the SS and EI correlate, and does adding automatically extracted fluency features to EI scores better account for a holistic score assigned to an SS test then EI alone?

The research conducted into these four questions has enhanced understanding of EI and of its potential applications. While the study of the EI techniques is not new, significant progress has been made in understanding the advantages of the test and the role it can play in the test community. This research provides additional insight and advances current understanding of the role of the EI test in English as a second-language test. Moreover, by comparing the utility and advantages of feature extraction using different automated tools, this study enables linguists to improve their targeted use of these tools in identifying accurate fluency features of ASR and acoustic analysis respectively. While the correlation and prediction accuracy of the models in this work does not reach the level where an automated exam would suffice for a high-stakes test,

it does demonstrate the potential of using this style of testing battery to identify the approximate oral proficiency of a speaker quickly and efficiently.

6.1 EI as a Production and Comprehension Test

Many researchers have wondered whether EI provides a better measure of a subject's oral production ability or of a subject's language comprehension (Jessop *et al*. 2007). By comparing EI to a simulated speech test (the speaking Language Aptitude Test or sLAT) and to an aural comprehension test, this study made it clear that both aspects are represented in the scores of an EI test. The dual production/comprehension nature of the EI test came into sharper focus when EI results were compared with the results from a reading and aural modality comprehension test, on the one hand, and the results from grammar and oral production tests, on the other. Slightly, though not statistically significant, stronger correlation with a grammar test over a reading test could be an indication of the ability of EI to test syntactic skills and grammar acquisition better than other language skills such as reading but would require additional investigation in order to identify a significant correlation trend.

6.2 The Optimal Fluency Feature Extractor

Both automatic speech recognition (ASR) and signal processing methods have been widely used in studies to assess a subject's oral fluency. This assessment is accomplished by extracting features of speech that are indicative of fluency. By contrasting these two methods of extracting features to quantify fluency, I have established that the features extracted via either system yield comparable results in predicting overall proficiency scores returned by an SS test. This finding is remarkable and significant for a number of reasons. Most notably, in

demonstrating that either system can provide an accurate assessment of the subject's fluency, this study has given a reason for linguists to consider other factors such as speed of use, implementation feasibility, and sound file quality when choosing which system to use in a real-world application scenario. The results have also made it apparent that the added complexity of the ASR system yields little extra information, and in some cases may even distort the results.

In order to compare the work of Matushita (2011) and De Jong and Wempe (2009), this study has deliberately omitted some of the available features from the analysis. The additional features for ASR, such as language model and acoustic model scores, might increase the quality of final model. However, the key feature—number of word types—appears to have been washed out by recognition inaccuracy and become a less accurate word-count metric, a metric which can more reliably be directly mapped to the syllable count extracted by the signal-processing methodology.

6.3 Fluency Features and SS

The fluency features provided a relatively good account of the data. While the prediction accuracy for the ML model of the SS scores was not extremely high, the regression model demonstrated that over a third (approximately 35% - $R^2 = 0.343$ for ASR and $R^2 = 0.356$ for Praat) of the variance in scores can be explained solely by the fluency features extracted. These results identify a relatively strong relationship between fluency and overall SS scores. The quality of automatically extracted fluency features also appears to be high.

6.4 EI and SS Fluency Features Combined

The distinction Matsushita (2011) makes between linguistic fluency and accuracy holds also for English testing. As indicated by Matsushita, the EI test results are a good indication of a subject's accuracy-related language skills and the features extracted from the SS test represent a good approximation of the subject's oral fluency. These conclusions are borne out in the results of both his study, and the work conducted here. The significant improvement of both the ML and regression model with the addition of the EI results clearly demonstrates that differing skills are represented in the respective tests. The increase of over 12% in the explanatory power of the regression model and the 10% jump in the predictive power of the TiMBL model indicate the value of the new information available in the EI test, information not represented in the fluency features.

The advantages of using this dual approach to automated testing over using either method separately are obvious. Because EI can be accurately scored via ASR, and used to target a subject's particular vocabulary, syntactic, and morphological acquisition level, the addition of the automatically extracted fluency features serves to greatly augment the accuracy of any automatically generated scores for test subjects. As both fluency and accuracy are fundamental to the considerations of human graders in the assignment of a grade for an oral proficiency exam, neither EI or fluency features by themselves can give an accurate and complete picture of the global oral proficiency of the speaker.

6.5 Research Limitations

This work on fluency features and EI relies heavily on the use of the SPHINX ASR engine. The SPHINX engine is a good representation of ASR functionality. However, other

engines might provide better results. Similarly, the use of Praat and the script designed for extraction of features reflects a choice of only one of the available tools that could be utilized.

All the tests used for comparison in this study were tests already implemented at the ELC. Consequently, no refinements or adjustment to test items, administration, or scoring were made to the listening, grammar, or reading tests that might have allowed for closer comparison with the EI test. Hand-scored sLAT test items were also not available.

6.6 Future Work

The potential applications of the EI/SS testing battery require further investigation. In previous work, an adaptive EI test was proposed and outlined (Lonsdale and Christensen 2011). The addition of the SS test to this process could lead to even more promising results. More work in the area of test-item generation and engineering for both EI and SS could help linguists better apply these tests to identify the accuracy and fluency, respectively, of a subject's use of a given feature.

Although the results of this study demonstrate the fundamentally dual modality of the EI test, significant further investigation is required to identify how the use of a simultaneous production/comprehension test affects a subject's test scores. As Vinther (2002) suggests, additional understanding of EI can be acquired by giving subjects additional listening training without significant oral instruction or practice and measuring the effects of this training on the test outcome. Targeted grammar tests that correlate with grammar elements in an EI test would also shed further light on the use of EI as a grammar-testing tool.

Language pronunciation is an important element of perceived language ability, yet it has received no attention in this study or in many of the previous studies in automatically scoring EI

tests. Once again, the possibility of predicting the test response *a priori* should allow for use of ASR techniques in scoring pronunciation for the learner on given phonemes. Additional modifications to the acoustic model could be made to account for L1 language backgrounds that would target specific pronunciation errors and so refine the EI test score to reflect these particular errors. The use of pronunciation scoring in the EI scoring currently implemented in the automatic testing procedure also requires attention in order to acquire a more complete picture of global oral proficiency of a test subject.

References

ACTFL. 1999. Oral Proficiency Interview Tester Training Manual. New York: American Council on the Teaching of Foreign Languages.

Ambridge, Ben and Julian Pine. 2006. Testing the Agreement/Tense Omission Model using an elicited imitation paradigm. Journal of Child Language, 33, 879-898.

Bernstein, Jared, Alistair Van Moere, and Jian Cheng. 2010. Validating automated speaking tests. Language Testing 27. 355–377.

Bley-Vroman, Robert, and Craig Chaudron. 1994. Elicited imitation as a measure of second-language competence. Research Methodology in Second-Language Acquisition, ed. by Elaine Tarone, Susan M. Gass, and Andrew D. Cohen, 245–261. Hillsdale, NJ: Lawrence Erlbaum.

Boersma, Paul, and David Weenink. 2005. Praat: doing phonetics by computer (version 4.3.29) [computer program], from http://www.praat.org/. Institute of Phonetic Sciences, Amsterdam.

Carrow, Elizabeth. 1974. A test using elicited imitations in assessing grammatical structure in children. Journal of Speech and Hearing Disorders 39.437.

Chambers, Francine. 1997. What do we mean by fluency? System 25.535–544.

Chapman, Robin, and Jon Miller 1975. Word order in early two- and three-word utterances: Does production precede comprehension? Journal of Speech and Hearing Research, 18, 355-371.

Chaudron, Craig. 2003. Data collection in SLA research. The Handbook of Second Language Acquisition, ed. by Catherine J. Daughty and Michael H. Long, 762–821. Malden, MA: Blackwell Publishing.

Christensen, Carl, Ross Hendrickson, and Deryle Lonsdale. 2010. Principled construction of elicited imitation tests. Proceedings of the 7th Conference on International Language Resources and Evaluation (LREC ’10), ed. by Nicoletta Calzolari, Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, Mike Rosner, and Daniel Tapias, 233–238. Valetta, Malta: European Language Resources Association (ELRA).

Clark, John L.D., and Ying-che Li. 1986. Development, validation, and dissemination of a proficiency based test of speaking ability in Chinese and an associated assessment model for other less commonly taught languages. Washington, DC: Center for Applied Linguistics.

Cook, Kevin, Jeremiah McGhee, and Deryle Lonsdale. 2011. Elicited imitation for automatic prediction of OPI Test Scores. Proceedings of the Sixth Workshop on Innovative Use of NLP for Building Educational Applications. Portland, OR.: Association for Computational Linguistics. 30-37.

Cucchiarini, Catia, Helmer Strik, and Lou Boves. 2000. Quantitative assessment of second language learners' fluency by means of automatic speech recognition technology. Journal of the Acoustical Society of America 107. 989–999.

Daelemans, Walter, Ko van der Sloot, and Antal van den Bosch. 2010. "TiMBL: Tilburg Memory Based Learner, version 6.3, Reference Guide," Tech. Rep., ILK Research Group Technical Report Series no. 10-01.

De Jong, Nivja H. and Ton Wempe. 2009. Praat script to detect syllable nuclei and measure speech rate automatically. Behavior research methods, 41 (2), 385 - 390.

Ellis, Rod. 1993. The Study of Second Language Acquisition. Oxford University Press.

Feyten, C. M. 1991. The Power of Listening Ability: An Overlooked Dimension in Language Acquisition. The Modern Language Journal, 75: 173–180. doi: 10.1111/j.1540-4781.1991.tb05348.x

Freed, Barbara F., Norman Segalowitz, and Dan P. Dewey. 2004. Context of learning and second language fluency in French: Comparing regular classroom, study abroad and intensive domestic immersion programs. Studies in Second Language Acquisition 26. 275–301.

Fujiki, Martin, and Bonnie Brinton. 1987. Elicited imitation revisited: A comparison with spontaneous production. Language, Speech, and Hearing Services in Schools 18: 301-311.

Ginther, April, Slobodanka Dimova, and Yang Rui. 2010. Conceptual and empirical relationships between temporal measures of fluency and oral English proficiency with implications for automated scoring. Language Testing 27(3), 379-399

Graham, C. Ray, Deryle Lonsdale, Casey Kennington, Aaron Johnson, and Jeremiah McGhee. 2008. Elicited imitation as an oral proficiency measure with ASR scoring. Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC '08), 57–67, Marrakech, Morocco.

Graham, C. Ray, Jeremiah McGhee and Benjamin Millard. 2010. The Role of Lexical Choice in Elicited Imitation Item Difficulty. Paper presented at the 2008 Second Language Research Forum: Exploring SLA Perspectives, Positions, and Practices, Somerville, MA.

Graham, C. Ray. 2006. An analysis of elicited imitation as a technique for measuring oral language proficiency. Selected Papers from the Fifteenth International Symposium on English Teaching, 57–67. Taipei, Taiwan: English Teachers Association.

Hakansson, Gisela, and Kristina Hansson. 2000. Comprehension and production of relative clauses: a comparison between Swedish impaired and unimpaired children. Journal of Child Language 27:313-333.

Hendrickson, Ross, Meghan Eckerson, Aaron Johnson, and Jeremiah McGhee. 2008. What makes test items difficult? – A syntactic, lexical and morphological study of elicited imitation test items. Proceedings of the Second Language Research Forum (SLRF).

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

Higgins, Derrick, Xiaoming Xi, Klaus Zechner, and David M.Williamson. 2011. A three stage approach to the automated scoring of spontaneous spoken responses. Computer Speech and Language 25.282–306.

Hood, Lois and Patsy Lightbrown. 1978. What children do when asked to "say what I say": does elicited imitation measure linguistic knowledge. Reprints from Allied Health and Behavioral Sciences 1. 195.

Housen, Alex and Folkert Kuiken. 2009. Complexity, accuracy and fluency in second language acquisition. Applied Linguistics 30:4.461-473.

Jessop, Lorena, Wataru Suzuki, and Yasuyo Tomita. 2007. Elicited imitation in second language acquisition research. The CanadianModern Language Review 64.215–220.

Keller-Cohen, Deborah. 1981. Elicited imitation in lexical development: evidence from a study of temporal reference. Journal of Psycolinguistic Research, 10(3), 273–288.

Koike, Dale. A. 1998. What happens when there's no one to talk to? Spanish foreign language discourse in simulated oral proficiency interviews. Talking and testing: Discourse approaches to the assessment of oral proficiency, ed. by Richard Young and Agnes W. He, 70–98. Amsterdam, Netherlands: John Benjamins Publishing Company.

Koponen, Matti, and Heidi Riggenbach. 2000. Overview: Varying perspectives on fluency. Perspectives on Fluency, 5–24. The University of Michigan Press.

Lee, Akinobu, and Tatsuya Kawahara. 2009. Recent development of open-source speech recognition engine Julius. Asia-Pacific Signal and Information Processing Association Annual Summit and Conference.

Laver, John. 1994. Principles of phonetics. Cambridge, UK: Cambridge University Press.

Lazaraton, Anne. 2002. A qualitative approach to the validation of oral language tests. Cambridge University Press, Cambridge.

Lee, Kai-Fu. 1989. Automatic Speech Recognition: The Development of the SPHINX System. Kluwer Academic Publishers, Boston, MA.

Lonsdale, Deryle, and Carl Christensen. 2011. Automating the scoring of elicited imitation tests. Proceedings of the ACL-HLT/ICML/ISCA Joint Symposium on Machine Learning in Speech and Language Processing. (5 pages).

Lonsdale, Deryle, Dan P. Dewey, Jeremiah McGhee, Aaron Johnson, and Ross Hendrickson. 2009. Methods of scoring elicited imitation items: an empirical study. Paper presented at American Association for Applied Linguistics (AAAL), Denver, CO.

Malabonga, Valerie, Dorry M. Kenyon, and Helen Carpenter. 2005. Self-assessment, preparation and response time on a computerized oral proficiency test. Language Testing 22. 59–92.

Malone, Margaret. 2007. Oral Proficiency Assessment: The Use of Technology in Test Development and Rater Training. In October 2007 CALdigest.

Matsushita, Hitokazu, and Matthew LeGare. 2010. Elicited imitation as a measure of Japanese L2 proficiency. Paper presented at Association of Teachers of Japanese (ATJ), Philadelphia, PA.

Matsushita, Hitokazu. 2011. Computerized Oral Proficiency Test for Japanese: Measuring Second Language Speaking Ability with ASR Technology. Unpublished MA Linguistics thesis, Brigham Young University.

Matsushita, Hitokazu and Deryle Lonsdale. 2012. Item Development and Scoring for Japanese Oral Proficiency Testing. Proceedings of 8th Language Resources and Evaluation Conference. 2682-2689.

McDade, Hiram L., Martha A. Simpson, and Donna Elmer Lamb. 1982. The use of elicited imitation as a measure of expressive grammar: A question of validity. Journal of Speech and Hearing Disorders 47. 19–24.

Menyuk, Paula. 1964. Comparison of grammar of children with functionally deviant and normal speech. Journal of speech and hearing research 7. 109.

Millard, Benjamin and Deryle Lonsdale. 2011. Developing French Sentences for Use in French Oral Proficiency Testing. Paper presented at the Linguistic Symposium on Romance Linguistics, University of Ottawa, Canada.

Mostow, Jack, and Gregory Aist. 1999. Giving help and praise in a reading tutor with imperfect listening -- because automated speech recognition means never being able to say you're certain. CALICO Journal, 16(3), 407-424.

Müller, Pieter, Febe de Wet, Christa van der Walt, and Thomas Nielser. 2009. Automatically assessing the oral proficiency of proficient L2 speakers. Proceedings of SLaTE 2009.

Naiman, Neil. 1974. The use of elicited imitation in second language acquisition research. Working Paper on Bilingualism 2.1–37.

Natalica, Diana. 1976. Sentence Repetition as a Language Assessment Technique: Some Issues and Applications. Paper presented at the Annual Meeting of the American Educational Research Association.

Neumeyer, Leonardo, Horacio Franco, Vassilios Digalakis, and Mitchel Weintraub. 2000. Automatic Scoring of Pronunciation Quality. Speech Communications, 30(2–3):83–93, Feb.

O'Loughlin, Kieran J. 2001. The equivalence of direct and semi-direct speaking tests. Studies in Language Testing. Cambridge, UK: Cambridge University Press.

Okura, Eve and Deryle Lonsdale. 2012, in print. Working memory's meager involvement in sentence repetition tests. In: Proceedings of the 34th Annual Meeting of the Cognitive Science Society.

Payne, J. Scott, and Paul Whitney. 2002. Developing L2 oral proficiency through synchronous CMC: Output, working memory and interlanguage development. CALICO Journal, 20, 7 – 32.

Préfontaine, Yvonne. 2010. Differences in Perceived Fluency and Utterance Fluency across Speech Elicitation Tasks: A Pilot Study. Papers from LAEL PG 2010. Edited by Kathrin Kaufhold, Sharon McCulloch, Ana Tominc. Vol 5.134-154.

Segalowitz, Norman. 2010. Conceptual and empirical relationships between temporal measures of fluency and oral English proficiency with implications for automated scoring. Language Testing 27. 379-399.

Shohamy, Elena, Chambers Gordon, Dorry M. Kenyon, and Charles W. Stansfield. 1989. The development and validation of a semi-direct test for assessing oral proficiency in Hebrew. Bulletin of Hebrew Higher Education 4.4–9.

Tomita, Yasuyo, Wataru Suzuki and Lorena Jessop. 2009. Elicited Imitation: Toward Valid Procedures to Measure Implicit Second Language Grammatical Knowledge. TESOL Quarterly: A Journal for Teachers of English to Speakers of Other Languages and of Standard English as a Second Dialect 43.6.

Valian, Virgina, Sandeep Prasada, and Jodi Scarpa. 2006. Direct object predictability: Effects on young children's imitation of sentences. Journal of Child Language, 33:247–269.

Vinther, Thora. 2002. Elicited imitation: A brief overview. International Journal of Applied Linguistics 12.54–73.

Weitze, Malena, and Deryle Lonsdale. 2011. The effect of syntax on English language learning. LACUS Forum XXXVI. Linguistics Association of Canada and the U.S. 309-315.

Xi, Xiaoming, Derrick Higgins, Klaus Zechner, and David M. Williamson. 2008. Automated scoring of spontaneous speech using SpeechRater v1.0. ETS Research Report No. RR-08-62. Princeton, NJ: Educational Testing Service.

Zhang, Xiaonan, Jack Mostow, and Joseph E. Beck. 2007. Can a Computer Listen for Fluctuations in Reading Comprehension? Proceedings of the 13th International Conference on Artificial Intelligence in Education, Los Angeles, CA, 495-502.