



Theses and Dissertations

2011-07-05

Elicited Imitation and Automated Speech Recognition: Evaluating Differences among Learners of Japanese

Shinsuke Tsuchiya
Brigham Young University - Provo

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



Part of the [Other Languages, Societies, and Cultures Commons](#)

BYU ScholarsArchive Citation

Tsuchiya, Shinsuke, "Elicited Imitation and Automated Speech Recognition: Evaluating Differences among Learners of Japanese" (2011). *Theses and Dissertations*. 2782.

<https://scholarsarchive.byu.edu/etd/2782>

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

Elicited Imitation and Automated Speech Recognition:
Evaluating Differences among
Learners of Japanese

Shinsuke Tsuchiya

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

Robert A. Russell, Chair
Masakazu Watabe
Dan P. Dewey

Center for Language Studies
Brigham Young University

August 2011

Copyright © 2011 Shinsuke Tsuchiya

All Rights Reserved

ABSTRACT

Elicited Imitation and Automated Speech Recognition:
Evaluating Differences among
Learners of Japanese

Shinsuke Tsuchiya
Center for Language Studies, BYU
Master of Arts

This study addresses the usefulness of elicited imitation (EI) and automated speech recognition (ASR) as a tool for second language acquisition (SLA) research by evaluating differences among learners of Japanese. The findings indicate that the EI and ASR grading system used in this study was able to differentiate between beginning- and advanced-level learners as well as instructed and self-instructed learners. No significant difference was found between self-instructed learners with and without post-mission instruction. The procedure, reliability and validity of the ASR-based computerized EI are discussed. Results and discussion will provide insights regarding different types of second language (L2) development, the effects of instruction, implications for teaching, as well as limitations of the EI and ASR grading system.

Keywords: Japanese, elicited imitation, automated speech recognition, computerized language test, SLA, second language studies

ACKNOWLEDGMENTS

I would like to extend my appreciation to all of the faculty members, family members, friends, and students who have helped me complete this thesis. I am especially grateful for Hitokazu Matsushita for not only creating the wonderful and reliable ASR system, but also for inviting me to join this research project and his patience and diligence in working with me. I would also like to express my gratitude for Drs. Robert A. Russell, Masakazu Watabe, and Dan P. Dewey for their guidance and understanding throughout my study at BYU, and for many wonderful opportunities to learn, teach, and smile. I am also grateful for Dr. Paul Warnick for his kindness and wonderful training in helping me become a better teacher. I would also like to express my thanks for my wonderful colleagues, Corey Petersen, Jordan Bledsoe, Joy Palmer, and Meg Bush for their valuable friendship, their inspiring examples, and good times in class. I am grateful for my students and friends for their willingness in helping me achieve what I have accomplished at BYU. I am thankful for my mom and dad for their understanding and for allowing me to receive education in the U.S. Finally, I am grateful for my Heavenly Father for his guidance in my life.

Table of Contents

Chapter 1: Introduction	1
1.1 Definitions of Instructed and Self-instructed Learners	4
1.1.1 Instructed learners.....	4
1.1.2. Self-instructed learners with and without post-mission instruction.	5
1.2 Research Questions	6
Chapter 2: Review of Literature	8
2.1 The Effects of Different Learning Environments	9
2.1.1 Self-instructed Learners and Naturalistic Learning.....	9
2.1.2 Instructed Learners and the Effects of Instruction.....	12
2.1.3 The Effects of Instruction on Self-instructed Learners at BYU.	13
2.2 Japanese EI Study by Yoshioka.	14
2.3 Japanese EI Procedure.....	18
2.3.1 Overview.	18
2.3.2 Testing environment.	19
2.3.3 Demographic data survey.	19
2.3.4 Training prior to test.	20
2.3.5 The EI test.....	20
2.3.6 Sentence selection.....	21
2.3.7 Post-test survey.....	23

2.4 ASR Grading Procedure.....	24
2.5 Development and Reliability of ASR Grading	25
2.6. Validation of EI.....	27
2.6.1 Rote repetition abilities and short-term memory.....	28
2.6.2 Implicit knowledge and meaning.	30
2.6.3 Linguistic and task difficulty.....	31
2.7 Implications of this Study	32
2.8 Research Questions and Hypotheses.....	34
Chapter 3: Procedure.....	37
3.1 Participants	37
3.2 Statistical Analysis	40
Chapter 4: Results.....	43
4.1 Classroom Instructed Learners.....	43
4.1.1 Overall EI score results.....	43
4.1.2 Discriminant item results.....	44
4.2 Instructed Learners and Self-instructed Learners.....	45
4.2.1 Overall EI score results.....	45
4.2.2 Discriminant item results.....	46
4.3 Self-instructed Learners with and without Post-mission Instruction	47
4.3.1 Overall EI score results.....	47

4.3.2 Discriminant item results.....	48
Chapter 5: Analysis.....	50
5.1 Classroom Instructed Learners.....	50
5.2 Instructed Learners and Self-instructed Learners.....	55
5.3 Self-instructed Learners with and without Post-mission Instruction	57
Chapter 6: Conclusion.....	60
6.1 Implications.....	62
6.2 Future Work	64
References.....	66
Appendix I: Consent Form.....	72
Appendix II: Translations of Discriminant EI sentences (Romanized).....	73

List of tables

Table 1: ASR scoring system.....	25
Table 2: Cohen’s kappa (comparison between ASR and human scoring).....	26
Table 3: Human-ASR correlation	27
Table 4: Average EI scores in pilot study	29
Table 5: Number of instructed learners	39
Table 6: Number of instructed and self-instructed learners.....	39
Table 7: Number of self-instructed learners with and without instruction	40
Table 8: Discriminant analysis with instructed learners (overall EI scores)	44
Table 9: Discriminant items and average scores of instructed learners.....	44
Table 10: Discriminant analysis of instructed learners (discriminant items scores)	45
Table 11: Discriminant analysis of instructed and self-instructed learners (overall EI scores)....	46
Table 12: Discriminant items and average scores of instructed and self-instructed learners	46
Table 13: Discriminant analysis of instructed and self-instructed (discriminant items)	47
Table 14: Discriminant analysis of self-instructed learners (overall EI scores)	48
Table 15: Discriminant items and average scores of self-instructed learners.....	48
Table 16: Discriminant analysis of self-instructed learners (discriminant items)	49

Chapter 1: Introduction

Measuring oral proficiency is a difficult and time-consuming task, which often yields subjective and unreliable results due to raters' preferences and experience (Hameyer, 1980). Moreover, data from free conversation tests or Oral Proficiency Interviews (OPI) are problematic for second language acquisition (SLA) studies due to test-takers' preference to avoid certain grammatical patterns (Chaudron, 1983; Russell, 2011, in press; Schachter, 1974). In order to account for these challenges in oral communication testing, elicited imitation (EI) tasks have been used in SLA studies for about 40 years to provide objective and consistent results that can provide useful feedback and information for learners and researchers (Bley-Vroman & Chaudron, 1994; C. R. Graham, Lonsdale, Kennington, Johnson, & McGhee, 2008; Jessop, Suzuki, & Tomita, 2007; Naiman, 1974).

EI is a research technique or testing method that requires participants to hear and repeat sentences in order to test oral proficiency or grammatical ability. EI is assumed to be reconstructive in the sense that participants process a stimulus sentence and reconstruct it with their own grammar when reproducing what they hear (Erlam, 2006). EI is usually conducted in a laboratory setting and EI prompts can be delivered orally or via a recording although recent studies prefer the latter medium because recording ensures controlled, consistent, and timed delivery.

Researchers use EI tests in SLA studies because they can easily incorporate and capture grammatical patterns or sounds that they wish to study instead of recording hours of spontaneous speech (Jessop, et al., 2007; Naiman, 1974). They argue that when learners' capacity is strained, their errors will reveal their underlying rule system.

Some researchers criticize EI because imitation tasks give the impression that they are only testing participants' short-term memory capacity instead of oral communicative ability, but recent studies have indicated that EI can be considered a viable oral testing method if done properly (Bley-Vroman & Chaudron, 1994; Erlam, 2006; C. R. Graham, et al., 2008). Another researcher (Ellis, 2005) claims that EI tests some important constructs of implicit knowledge such as (a) response according to feel rather than declarative rules, (b) spontaneity rather than planned behavior, and (c) a primary focus on meaning before forms (p. 152).

EI researchers have used various methods to ensure that EI tasks are not testing short-term memory by providing items that are long enough, and presenting them in a randomized order to make sure that certain orders of sentences do not affect their performance. Furthermore, some researchers have required their participants to spontaneously correct ungrammatical EI stimuli to show that EI is reconstructive and avoid rote repetition (Erlam, 2006). Others have required participants to count to 3 after each test item before responding (Mackey & Gass, 2005). Based on their results, Bley-Vroman and Chaudron stated that EI is a reasonable measure of global proficiency because their research results indicate that "the more you know of a foreign language, the better you can imitate the sentences of the language" (1994, p. 247).

With the advancement of technology, the Brigham Young University Pedagogical Software and Speech Technology research group has been developing computerized EI using automated speech recognition (ASR). They have conducted over 1,600 tests with more than 1,350 English as a second language (ESL) learners using EI and ASR, and have reported that overall comparisons between ESL learners' EI scores and their scores on other measures of oral language proficiency are promising (Hansen, Graham, Brewer, Brewer, & Tieocharoen, 2007), and that EI scores can predict Oral Proficiency Interview (OPI) scores within two levels of a

scale of ten (C. R. Graham, 2006; Lonsdale, Graham, & Madsen, 2005). They reported correlations between an ASR and human grading ranging from 0.85 to 0.88.

In terms of Japanese EI testing with an ASR grading system, a pilot study (Matsushita & LeGare, 2010) was conducted with over 100 learners of Japanese, of whom 28 took the OPI for comparison with their EI scores. It was reported that the correlation between sentence-level Japanese EI scores and OPI scores was $r=0.77$ ($p < 0.30$), and the correlation between human grading of EI tests and ASR grading was $r=0.84$ ($p < 0.001$). Although not statistically significant, Matsushita and LeGare reported that EI can be a promising testing method to measure L2 Japanese oral proficiency, but the system needs to be improved for practical use. They also suggested that EI can be used for placement and diagnostic purposes and supplement the oral portion of the tests. Since then, Matsushita has revised ASR from grammar-based recognition (Matsushita & LeGare, 2010) to statistical language models trained with a Corpus of Spontaneous Japanese (Matsushita, Dewey, & Lonsdale, 2010), and incorporated an exemplar-based machine learning system (Matsushita, 2011) to improve its accuracy rate. He has also studied the effects of repeated exposure to EI tests as well as prompt speed in order to adjust the EI task appropriately, reporting that both repeated exposure and prompt speed affect EI performance. In his most recent study, Matsushita reported that the correlation between human scoring and ASR scores is promising for real-time grading of adaptive testing and he is planning to develop an online EI-based testing system (Matsushita, 2011).

Although machine-based tools such as ASR are becoming more widely available in the language pedagogy field, language teachers and learners are often concerned with the reliability (or the consistency of test scores) of machine-scored language proficiency measures (Kim, 2006). However, an EI + ASR grading system could be a useful research tool in SLA studies because it

has the potential of providing consistent and reliable results much faster than human graders. Moreover, participants' utterances are recorded and these recordings can provide meaningful feedback not only for learners and teachers, but also for SLA researchers who are interested in L2 language development. Thus it is crucial to examine the reliability of the scoring system in question by conducting further analysis to see if the ASR scoring is providing test scores that are consistent.

The purpose of this thesis is to evaluate differences among learners of Japanese, particularly in relation to instructed and self-instructed learners that participated in the Japanese EI study at Brigham Young University (BYU). Their ASR-graded EI performance will be analyzed to evaluate differences among them further below to examine the usefulness of ASR grading scores obtained through EI testing as a research tool. Note that the definition of usefulness in this thesis is the quality of having utility and especially practical worth or applicability of ASR scores for SLA studies.¹ Definitions of the categories of learners will be discussed in the following section.

1.1 Definitions of Instructed and Self-instructed Learners

1.1.1 Instructed learners. The definition of instructed learners for this thesis is college-level classroom students who are taking Japanese courses in the U.S. At BYU, there are typical classroom-instructed learners who learn Japanese in a classroom setting and have little or no experience residing in Japan. Many of these instructed learners are either in the beginning- or intermediate-level classes, although some progress to advanced-level classes as well. For the purposes of this thesis, first-year students will be referred to as beginning-level learners, second-year students as intermediate-level learners, and third- and fourth-year students will be referred

¹ For Bachman and Palmer (Bachman & Palmer, 1996), the definition of usefulness in language testing is a kind of metric to evaluate all aspects of test development and use, which is different from the definition of usefulness in this thesis in that the focus is to examine the practical worth or applicability of ASR scores for SLA studies in particular.

to as advanced-level students. Their EI performance across levels will be compared and analyzed with the data acquired from the EI + ASR grading system to evaluate differences that might be useful for studies related to Japanese as a second language (JSL).

1.1.2. Self-instructed learners with and without post-mission instruction. The definition of self-instructed learners, on the other hand, is college-level students who have learned Japanese on their own during their 1.5-1.8-year service missions in Japan. In addition to typical instructed learners, BYU has a large number of advanced-level JSL learners, most of whom are self-instructed learners who served their two-year LDS church missions in Japan during their undergraduate education or before coming to BYU. Although these missionaries received at least two months of instruction in the Missionary Training Center prior to going to Japan, they received no formal Japanese language instruction during their 1.5-1.8-year mission experience in Japan. Instead they learned the language largely naturalistically through interaction with others and by studying Japanese on their own.

Currently, these self-instructed learners begin their Japanese study as third-year students (i.e., advanced-level) along with advanced-level classroom-instructed learners. Since these learners are placed in the same upper-level classes, it would be vital for SLA researchers to evaluate differences among them that EI and ASR might be able to reveal. In this study, those who are enrolled in their first post-mission course at the third-year level will be considered self-instructed learners without post-mission instruction for purposes of comparison with advanced-level instructed learners, because of the minimal amount of formal instruction received by the former compared to the latter.

Many of the self-instructed learners in this study continue to take more advanced-level Japanese classes hoping to improve or keep the Japanese skills acquired during their mission. In

order to show that these self-instructed learners are benefitting from post-mission instruction, it would be important to compare these self-instructed learners with post-mission instruction and the other self-instructed learners whose experience with classroom instruction is minimal. In this thesis, those who have taken two or more-semester of instruction will be referred to as learners with post-mission instruction. If the findings can show that self-instructed learners are benefitting from their post-mission classroom experience, it will be useful in motivating more self-instructed learners to take Japanese classes.

Now that the definitions of instructed and self-instructed learners for this study have been discussed, the following section will list the research questions concerning these participants.

1.2 Research Questions

The purpose of this thesis is to examine the usefulness² of ASR graded scores obtained through EI testing as an SLA research tool, and to evaluate differences between instructed and self-instructed learners at BYU. In particular, the following three research questions will be addressed:

1. Can EI and ASR grading system differentiate beginning-, intermediate-, and advanced-level instructed learners?
2. Can EI and ASR grading system differentiate advanced-level instructed learners and self-instructed learners without post-mission instruction?
3. Can EI and ASR grading system differentiate self-instructed learners with and without post-mission instruction?

Answering these research questions using EI and ASR will allow the author to discuss the usefulness of EI and ASR in SLA research. Moreover, a number of significant EI sentences will be analyzed to evaluate differences among learners of Japanese.

² See note 1 on page 4 for the definition of usefulness in this thesis.

Because participants' status as instructed or self-instructed learners is a pre-existing condition in relation to this study, and the data has already been collected for analysis, a causal-comparative method will be used to identify possible causes of observed variation in the behavioral patterns of JSL learners. Moreover, this is a cross-sectional study comparing groups of learners at the same time, so only inferences can be made about differences in their language development, inferences that may or may not be corroborated by follow-up longitudinal studies.

Chapter 2: Review of Literature

As mentioned in the introduction, this study uses data acquired from the EI and ASR grading system to compare beginning-, intermediate-, and advanced-level instructed learners, as well as self-instructed learners at BYU. Thus, one of the objectives of this section is to provide a brief history of previous research concerning the effects of the learning environment on self-instructed and instructed learners in order to understand the characteristics of the participants in the research. Specifically, this section will look at the advantages and disadvantages of their learning environments, including references to Russell's studies (2005, 2011, in press) on the effects of post-mission instruction on self-instructed learners in advanced-level Japanese courses at BYU. Understanding the characteristics of participants' L2 will be useful for analysis of their EI performance data as evaluated by the ASR system.

After discussing both self-instructed and instructed learning environments, there will be a review of a Japanese SLA study conducted at the University of Hawaii-Manoa concerning the acquisition of Japanese particles (Yoshioka, 1991). This is a very important study to review in this context, as it appears to be the only one in Japanese SLA studies that has used EI as an instrument to examine JSL learners' language development. Yoshioka's report will be useful in understanding the implementations of EI tasks, and what needs to be considered when creating EI sentences.

After reviewing Yoshioka's Japanese EI study, the EI and ASR scoring system procedure that was used with JSL learners at BYU will be reviewed to discuss its validity and usefulness (Matsushita, 2011). This section will help readers understand what participants' EI scores represent, as measured by the ASR scoring system.

Finally, the implications of this study and its research questions and respective hypotheses will be discussed at the end.

2.1 The Effects of Different Learning Environments

The objective of this section is to discuss the effects of different learning environments on self-instructed and instructed learners along with their typical advantages and disadvantages that might be highlighted by the data obtained from the EI and ASR. A review of self-instructed learners will be conducted first, followed by a discussion of instructed learners and the effects of instruction on self-instructed learners at BYU.

2.1.1 Self-instructed Learners and Naturalistic Learning. Many people believe that they can learn to communicate if they live in the target language community for an extended period. Their assumption is that learners who are immersed in the target culture in the contexts of study abroad programs or missionary service become part of a native speech community, which provides linguistically rich comprehensible input. This belief is related to the Input Hypothesis (Krashen, 1982, 1985), which claims that learners acquire L2 by being exposed to rich comprehensible L2 input for an extensive period. Krashen further argues that those who acquire their L2 in natural settings have the potential to learn like a child, who manages to master the language without the benefit of instruction given ample time.

Krashen's assumptions are supported by an influential study (Carroll, 1967), which compared study abroad students and domestic classroom-instructed students on multiple-choice test scores with 2,782 college seniors majoring in French, German, Italian, and Russian. According to Carroll, one of the major predictors of overall language proficiency is time spent abroad. Although this result is based solely on multiple-choice questions, it suggests that

extensive stay in the target culture can play an important role in improving overall language proficiency.

In other recent studies concerning study abroad students, it has been reported that they gain more vocabulary than regular classroom students (DeKeyser, 1986, 1991), and acquire more sociolinguistic features of the target language (Lafford, 2004; Marriot, 1995; Regan, 1995), as well as oral fluency (Freed, 1995; Lafford, 2004). Self-instructed learners' EI performance may indirectly show these benefits when compared with classroom-instructed learners.

Although study abroad students and self-instructed learners are similar in the sense that both of them are immersed in the target culture, they are different in that self-instructed learners do not receive instruction during their 1.5-1.8-year missions. Similar benefits from their learning environment can be assumed, but there are other findings specific to self-instructed learners. In a discussion of the contexts of a naturalistic environment and their effects on self-instructed learners (Collentine & Freed, 2004), it was reported that self-instructed learners tend to have a basic L2 variety that is morphosyntactically noncomplex and constant across time in terms of their grammatical development (Perde, 1993), probably due to lack of instruction. This idea of a noncomplex L2 variety may indicate that self-instructed learners in this study might only have acquired basic grammatical competency. Hence, if the self-instructed learners relied heavily on this noncomplex L2 during their 1.5-1.8-year missions, they may perform poorly on complex sentence structures.

Moreover, others (Carson & Longhini, 2002) have reported that a naturalistic learning environment influences the types of learning strategies that self-instructed learners develop (e.g. reading a grammar book and writing out verbs). But they report that these strategies appeared not to lead to “intake”—which is defined as a part of input that has actually been processed by the

learner and turned into knowledge of some kind (Corder, 1967). This finding implies that these self-instructed learners may not have benefited from their own language studies. In addition, Carson and Longhini's study also indicated that continuous communicative demands in the target culture might make memorization impractical for self-instructed learners, which may lead to frequent use of compensatory communicative strategies. In other words, self-instructed learners may only have memorized and used basic vocabulary and grammatical patterns during their mission, and relied on compensatory communicative abilities to make up for their lack of vocabulary and grammatical knowledge.

Lastly, with regards to immersion experience, it is reported that study abroad learners are sometimes not necessarily "immersed" in the native speech community (Frank, 1997; Wilkinson, 1998). Consequently, there are those who claim that study abroad students do not differ significantly in their improvement in overall grammatical ability when compared to classroom learners (Freed, 1995; Möhle, 1984; Regan, 1995). Self-instructed learners may also show the same weakness in grammatical ability because of their tendency to rely on a noncomplex L2 system.

In sum, self-instructed, host-culture-resident learners may have the same benefits that study abroad students have from their learning environment, such as those of acquiring more vocabulary, sociolinguistic features, and oral fluency with rich linguistic input in the target culture. However, they may differ from study abroad students in that they typically do not receive instruction during their stay in Japan. Due to lack of instruction, their grammatical competency might still be at a basic level, and they may tend to rely heavily on their basic vocabulary and grammatical knowledge through compensatory communicative strategies. And to

the extent that self-instructed learners have not been immersed in the L2 culture, their language skills might not have benefited from extra exposure to the target culture at all.

2.1.2 Instructed Learners and the Effects of Instruction. Although instructed learners' experience with the target language may be limited in the classroom, there are some advantages in their learning environment (Doughty, 2003). For example, L2 instruction can assist learners by organizing the processing space as well as their input processing to help them use a more efficient acquisition mode. Doughty believes that in order for instruction to be effective, teachers need to direct students' attention to an L2 element in meaningful contexts, so that it will help them organize the processing space and input processing. Doughty's claim is supported by the noticing hypothesis (Schmidt, 2001), which claims that second language acquisition can be facilitated when learners' attention is directed to certain elements of target language input.

Moreover, instructed learners may be able to acquire higher language proficiency than self-instructed learners because instruction directs learners to notice 'marked' elements (Long, 1988). In Long's terms, markedness refers to infrequency, which means that grammatical patterns that are marked do not happen frequently in daily conversation. For example, a causative-passive structure or complex relative clause structure in Japanese might be difficult for self-instructed learners to acquire, because these grammatical patterns are not only complex, but do not occur frequently. Thus, self-instructed learners may never gain sufficient access to marked input and tend to acquire only the unmarked elements (Pavesi, 1986). However, instructed learners may be able to notice marked elements in Japanese if their attention is effectively directed to marked elements in class. Consequently, instructed learners may be able to perform better on EI sentences with marked grammatical patterns.

However, instructed learners' experience with the target language is quite limited in the classroom, which may result in acquisition of vocabulary, sociolinguistic features of the target language, as well as oral fluency that is relatively less than that of self-instructed learners (DeKeyser, 1986, 1991; Freed, 1995; Lafford, 2004; Marriot, 1995; Regan, 1995). Thus, they may not have the oral-aural ability to hear, process, and imitate the sentences as well as self-instructed learners do.

To summarize the relative advantages and disadvantages, instructed learners have the potential to acquire target language more efficiently when classroom instruction presents input in an organized way. In addition, they may have better ability to notice marked elements in the target language. However, their acquisition of vocabulary, sociolinguistic features of the language, and oral fluency may not be as great as that of self-instructed learners' due to their limited experience with the target language in the classroom.

2.1.3 The Effects of Instruction on Self-instructed Learners at BYU. Russell's study on the effects of instruction on self-instructed learners at BYU reported that instruction appears to have a positive effect on self-instructed learners in retaining their oral ability (Russell, 2005, 2011, in press). Russell conducted a 12-year study examining learners' lexical and syntactic ability in free oral production tasks. He reported that there was little significant difference between those with and without post-mission instruction within the first few years after their 1.8-year missions. However, significant differences were found after 10-plus years later, especially in vocabulary size and in syntactic complexity and variety, with those who had received post-mission instruction maintaining their oral proficiency better than those who had not (Russell, 2005, 2011, in press).

The third research question relates to Russell's study in that it compares self-instructed learners based on their post-mission instruction experience. However, it differs from Russell's study in the sense that it is a cross-sectional study comparing different groups of participants at the same time, as opposed to comparing the same sets of students longitudinally. Though the testing constructs are different, findings of this research may indirectly either confirm or contradict Russell's findings about self-instructed learners' L2 oral proficiency and the effects of instruction. The results would be potentially important for language instructors if the effects of instruction can be verified through their EI and ASR scores.

The next section will review an EI-based SLA study related to the acquisition of Japanese particles and a discussion of the effects of instruction.

2.2 Japanese EI Study by Yoshioka.

Despite the fact that SLA studies with EI have been done in other languages, Yoshioka's study (Yoshioka, 1991) seems to be the only one dealing with JSL that has examined the usefulness of EI as an oral production assessment. Yoshioka's study analyzed beginning- and intermediate-level learners' EI performance with a focus on the effects of placement of target particles—namely, *wa*, *ga*, and *o*—as well as the ease of repetition in relation to their location. In this section, the actual implementation of EI tasks in her research and the effects of educational instruction will be discussed. Moreover, important findings about EI sentence creation and how the placement of target grammatical structures affects participants' EI performance will be discussed.

Yoshioka's pilot study incorporated various particles, verb forms, forms of negation, and word order in 40 EI sentences to test students. Based on the results, she narrowed down the target structures to three particles—*wa*, *ga*, and *o*. Then, sentences in the pilot study were

modified into 27 sentences to study these three particles. Although particles have multiple functions, functions in her study were limited to one for each particle—the particle *wa* for topic marking, the particle *ga* for subject marking, and the particle *o* for object marking. These 27 sentences made up a short story and a male Japanese speaker recorded the story.

In terms of sentence length in her EI sentences, Yoshioka reported that the length of each sentence was longer than the average human short-term memory span, which is reported to be seven plus-or-minus two English syllables (Naiman, 1974). Naiman reported that appropriate sentence lengths can minimize the possibility of participants relying solely on their short-term memory without processing the meaning of the sentences.

With regard to placement of the target grammatical structures, Yoshioka placed *wa*, *ga*, and *o* in various locations in the sentences (i.e., sentence-initial, -medial, and -final) on the assumption that the accuracy of repetition of the targeted structures might depend on their placement in a given sentence (Blasdell & Jensen, 1970; Grossman & Scholes, 1971; Naiman, 1974). Thus, the location of *wa*, *ga*, and *o* was noted to investigate the location effect. However, Yoshioka did not explain how her sentences were divided into initial, medial, and final parts.

The subjects were all English native speakers who were taking Japanese courses, and they were divided into four different groups based on the number of semesters they had taken the language in college prior to the study. The study involved 15 first-semester students, 21 second-semester students, 26 third-semester-students, and eight fourth-semester students, including 25 males and 45 females. These participants all used the same textbook in class, which introduced *wa*, *ga*, and *o* in the same lesson in the fourth week of the first semester.

In terms of procedures, the EI test was conducted in the language laboratory. English instruction was given to participants to listen to the sentence twice and then, when signaled by a

tone sound, to repeat as closely as possible the sentence they had heard. Although Yoshioka reported that each test session lasted for approximately 30 minutes, she did not explain how prompts were delivered to participants, and the amount of time that was allowed for them to repeat each individual sentence was also not reported.

With regard to scoring, Yoshioka used correctly repeated particles to calculate accuracy rates, which were compared among four groups of students. The accuracy rates for each particle were then analyzed in relation to placement of the particles in the sentence to determine the location effect. Whether feedback or a report of their performance was given to the participants is unknown.

In terms of repetition of the target particles by different levels of participants, Yoshioka reported that “the target structures showed similar improvement of scores towards higher levels (beginning- to roughly lower intermediate-level students) (p. 136).” To an extent, this finding validates the statement that “the more you know of a foreign language, the better you can imitate the sentences of the language” (Bley-Vroman & Chaudron, 1994, p. 247). Furthermore, the finding also shows the effects of instruction—that participants’ language ability is improving from semester to semester as measured by their ability to accurately repeat target particles in the sentence. However, there were no other treatments conducted to examine the learners’ understanding of these three particles.

On the matter of location effect, Yoshioka reports that the position of targeted grammatical patterns in a sentence determines the ease of repetition. She reported that the initial position showed better performance across particles and levels in general. This finding supports the results of an earlier study (Naiman, 1974), that grammatical patterns located in the sentence-initial position were the easiest to repeat. However, there is a conflict between Naiman’s study

and Yoshioka's finding in terms of grammatical patterns located in the medial and final position with regard to ease of repetition. On one hand, Naiman reported that the most difficult place in the sentence to repeat is in the middle of the sentence—whereas Yoshioka's results indicated that it was actually the sentence-final position that was most difficult. In her analysis, Yoshioka suggested that the canonical order of participants' L1 and the target language played an important role in influencing their repetition performance. For example, in the case of topic marker *wa* and the subject marking *ga*, they are easy to repeat in the sentence initial position, because it matches the participants' expectation that the topic marker and subject marking should occur in the sentence-initial position according to English SVO and Japanese SOV canonical ordering. However, participants would not expect the object marking *o* in the sentential initial position because it is not typical in English or Japanese. Her claim is supported by the results that showed less successful retention and repetition with the sentence-initial *o*. Yoshioka (1991) concludes her study with the remark “that in order to obtain an accurate picture of participants' production grammar, the position of target grammatical patterns needs to be controlled so that they appear equally in, at least, the initial and the final half of the sentences” (p.136).

Although Yoshioka does not mention it in her analysis, the rationale that the sentence-final position was the most difficult part to repeat in her study might be due to the fact that the Japanese predicate usually comes at the end (i.e., SOV), which contains the greatest amount of information in the sentence. English, on the other hand, has the predicate in the middle part of the sentence (i.e., SVO), which is also crucial in determining the meaning of the sentence. Thus, this notion of the placement of the predicate in different languages should also be considered when creating EI sentences.

In summary, Yoshioka conducted a cross-sectional study on students at the University of Hawaii-Manoa to show that the accuracy rates of successful repetition of target particles (*wa*, *ga*, and *o*) improved from semester to semester with the EI test she created. In terms of location effect, she reported that patterns placed in the sentence-initial position were the easiest to repeat. However, her findings and Naiman's findings on the ease of repetition and patterns found in sentence-medial and -final position did not match. Although no further studies have been conducted to examine the ease of repetition between middle and final position, Yoshioka suggested that the canonical ordering of participants' L1 and target language, as well as participants' expectations, might have affected their performance.

Now that the advantages and disadvantages of self-instructed and instructed learners have been discussed and the implementation of EI testing has been reviewed, the following section will review and analyze the EI task and the ASR scoring system that have been used with the JSL participants at BYU to discuss their validity and usefulness for SLA research.

2.3 Japanese EI Procedure

The purpose of this section is to explain the procedure of the EI test used in this study in order to review its validity and usefulness for SLA research.

2.3.1 Overview. The procedure of the actual EI test included participants' reading and filling out a physical consent form with their signature, instructions by lab assistants on how to use the headphones and the microphones, and filling out their demographic information electronically. Then, a brief practice session was given after which participants took the actual test by listening to and repeating 60 sentences from which their utterances were recorded. A post-test survey about the EI test was given at the end. Their utterances were later scored by the ASR grading system developed by Matsushita (2011).

2.3.2 Testing environment. To make sure that the testing environment was comparable for all test takers (Tomita, Suzuki, & Jessop, 2009), all of the EI tasks were conducted in the computer lab using recorded sentences. Upon entering the computer lab, all of the participants were required to read and sign a consent form with an agreement that their personal information would be kept confidential. Then instruction on how to use the headphones and the microphone was given to them by a lab assistant using a visual aid. This was done to make sure that participants used the device properly to prevent technical problems such as data loss, background noise or unrecorded utterances. Then the lab assistant led the participants to a computer available in the room to take the test. There were more than 60 Macintosh desktop computers available for the test and participants were able to take it during the lab hours from 8:00 a.m. to 6:00 p.m. on Monday through Saturday during the data collection period. There were no dividers between computers but participants were placed in different sections of the lab by lab assistants to avoid distraction from each other as much as space permitted.

2.3.3 Demographic data survey. Before taking the EI test, participants filled out their demographic data which included their current Japanese class, name, age, nationality, language background—the number of semesters/years studying Japanese as well as other languages, and a brief description of the learning environment that might have influenced their Japanese ability—family members, friends, experience in Japan, and so forth. Finally, their self-assessment of their speaking, listening, reading, and writing skills was reported on a Likert scale of poor, fair, good, excellent, native-like, and native. All of the instructions were given in English and participants typed in their information in English. Some of the data acquired from this demographic survey was used to group students (e.g., self-instructed and instructed learners and the number of semesters they had received instruction).

2.3.4 Training prior to test. In order to appropriately train and familiarize participants with the EI task, an English instruction on how to do the EI task was given on the computer screen upon completing their demographic information section. The instruction informed students to listen to an utterance once, wait for 3 seconds, and then repeat the sentence exactly as they heard it. They were given 8 seconds to repeat the sentence. To make sure that their headphones and microphones were working properly and their utterances were recorded, they were given a practice sentence in which they were instructed to push a start button to hear a sample sentence given via the computer (e.g. *ohayoo gozaimasu* ‘good morning.’), wait for 3 seconds, and repeat the sentence. After the recording, a pop-up box appeared and asked them to listen to their utterance by clicking a button to make sure that it was recorded before proceeding to the actual test. They were able to raise their hand to be helped by a lab assistant if there were any technical problems, and they could practice the task as much as necessary to familiarize themselves with it.

2.3.5 The EI test. After the practice session, the actual EI test was given with a total number of 60 sentences. As mentioned earlier, each sentence was presented once followed by a 3-second pause with beep sounds. Participants had 8 seconds to record their utterance after which a new item was given without pauses between items. With 60 items, the actual EI test took about 15 minutes to complete. All of the prompts were recorded by both male and female native speakers in the recording lab multiple times prior to the study, and items that had the best sound quality were used for the test. To avoid ordering effect, items were presented in a randomized order, but the program was set to give each participant 30 sentences from a male speaker and the other 30 sentences from a female speaker.

2.3.6 Sentence selection. EI items included 18 beginning-level, 17 intermediate-level, 17 advanced-level and 6 superior-level items—half of which came from patterns used in the textbooks used at BYU (Jordan & Noda, 1987a, 1987b, 1987c; Watabe & Gilbert, 1993a, 1993b). The rationale behind using the patterns in the textbooks was to control grammatical patterns and vocabulary in a way that beginning-level items only dealt with first-year materials, intermediate-level items with second-year materials, and the advanced-level items with third-year materials, and so on. Superior-level items were created to differentiate advanced-level learners and native speakers to avoid a ceiling effect. The sentence length, content, and complexity of the sentence patterns were, of course, modified so as to differ from those presented in the book to prevent any bias towards instructed learners. However, some of the beginning-level items had to be short and simple, so they could not be modified as much. As a result, there may have been some bias towards instructed learners with beginning-level items. The other half of the items came from CSJ (Corpus of Spoken Japanese) with less control on vocabulary and grammatical patterns. However, sentence length, content, and pattern complexity were modified to create level-appropriate items. Unlike Yoshioka's approach, these sentences were not related to each other and were not in the form of a narrative.

In terms of placement of grammatical patterns for each testing item, sentence-initial positions were avoided as much as possible. In other words, grammatical patterns were placed towards the middle or final part of the sentence on the assumption that placement of grammatical patterns toward the middle or final part of the sentence would prevent participants from relying excessively on their short-term memory (Naiman, 1974; Yoshioka, 1991), and that they would be required to process the sentence in order to successfully comprehend and repeat it.

Sentence length ranged from 9 to 31 morae in which length and complexity were adapted to suit each level as discussed previously. The sentence length was based in part on English EI studies, which ranged from about 5 to 25 syllables (C. R. Graham, et al., 2008), and the number of morae as well as *bunsetsu* (i.e., syntactic unit phrases) were used to determine length and complexity.

The content of each sentence was also modified to make sure that the test was fair to all of the participants to avoid any bias in gender, religion, language, race, or ethnicity. In other words, gender-oriented words, numbers, proper nouns, interjections as well as semantically similar sentences were avoided. An effort was made to preclude items that required certain background knowledge for participants to successfully complete the task.

The following section will provide a description of the EI sentences. The number of morae and *bunsetsu* were counted by two Japanese native speakers.

2.3.6-a Beginning-level EI sentences. Beginning-level items included grammatical patterns such as case particles, sentence-final particles, and affirmative and negative sentences in present tense. There were also some difficult expressions such as extended predicates, humble forms and comparatives that could be treated as marked patterns discussed in the review of literature. The average number of morae was 15.5, with a range from 9 to 25. The average number of *bunsetsu* was 3.5, with a range from 2 to 6.

2.3.6-b Intermediate-level EI sentences. Intermediate-level items included grammatical patterns that formed more complex sentences than beginning-level items. Sentences with verbs of giving, relative clauses, honorifics, conditional, adversative passives and causatives were used—all of which could be considered as marked patterns for English native speakers. The

average number of morae was 20.1, with a range from 14 to 25. The average number of *bunsetsu* was 4, with a range from 2 to 6.

2.3.6-c Advanced-level EI sentences. Advanced-level EI sentences were longer and more complex than the intermediate sentences. All of the sentences had more than two clauses that would require participants to be able to discern between independent clauses and subordinate clauses. Conditional sentences were also used in addition to verbs of giving, relative clauses, honorifics, adversative passives and causatives—again, these patterns could be considered marked patterns for English native speakers. The average number of morae was 23.8, with a range from 16 to 27. The average number of *bunsetsu* was 4.5, with a range from 3 to 5.

2.3.6-d Superior-level items. Lastly, superior items consisted of sentences that were even longer and more complex than advanced-level items. Unlike beginning-, intermediate-, and advanced-level items, superior items were not based on the grammatical patterns introduced in the textbooks for each level. Instead, very complex relative clauses were used in order to differentiate native speakers and advanced-level learners. The average number of morae was 28.2, with a range from 25 to 31. The average number of *bunsetsu* was 5.7, with a range from 5 to 7.

2.3.7 Post-test survey. After completing the EI test, participants evaluated the test in a post-test survey on the computer by answering four questions and leaving comments or feedback. Participants answered the following question on a Likert scale of strongly agree, agree, somewhat agree, somewhat disagree, disagree, strongly disagree:

1. The native speaker models in the Sentence Imitation test spoke at a speed I was comfortable with.
2. I had enough time to focus on meaning of the sentences in the Sentence Imitation test.

3. The audio for the native speakers' examples was easy to understand.
4. The Sentence Imitation test was a good measure of my Japanese language proficiency.

Although participants' responses were interesting, they will not be discussed in the analysis. Lastly, participants were also able to leave any feedback or comments that they had for the test.

This concludes the description of the EI task procedure. The following section will discuss the procedure and validation of the ASR grading system used to measure participants' EI performance.

2.4 ASR Grading Procedure

Participants' utterances were evaluated with a binary scoring system of 0 and 1 in terms of correct morae (C) and other types of morae such as substituted morae (S), deleted morae (D), and inserted morae (I); from which the number of correct morae was used for score calculation. The unique aspect of this scoring system was the fact that participants' ability to imitate the whole sentence was assessed without any emphasis on grammar. The rationale behind this scoring system is that participants' overall ability to imitate the whole sentence requires comprehensive understanding of grammatical patterns and vocabulary in the sentence. Thus, if participants successfully imitate the sentence, it can be inferred that they have a good grasp of grammatical patterns and vocabulary found in the sentence.

The equation for score calculation is simple: the number of correct morae is divided by the number of total morae; which is multiplied by 100. For example, a sentence *その中に何入ってんの?* (i.e., *sono naka ni nani haitte n no?*) 'what's in it?' has a total number of 13 morae (i.e., *so no na ka ni na ni ha i t e n no*). If the participant's utterance that was recognized by the ASR grading system was *その中何入ってたしよ* (i.e., *so no na ka na ni ha i t e ta syo*) with

the number of 10 correct morae, then this number of correct morae (10) was divided by the total number of morae (13). The value was then multiplied by 100 which yields the score of 76.92 for this particular item.

Although inserted morae, substituted morae, and deleted morae were also recorded, they were not used to count against their overall scores. The rationale is that there are times when participants manage to repeat back most of the morae in the sentence correctly, even though they may pause, restate, use other expressions or omit certain morae. In other words, participants were given points as long as they successfully uttered correct morae in the correct order in any parts of the sentence. See Table 1 for a summary.

Table 1: ASR scoring system

prompt	
ソノナカニナニハイッテンノ 'so no na ka ni na ni ha i t te n no'	Total number of morae in the sentence = 13
Participant's utterance	Correct number of morae = 10
ソノナカナニハイッテタシヨ 'so no na ka na ni ha i t te ta syo'	Substitutions = 2
Evaluation	Deletions = 1
C C C C D C C C C C S S I	Insertions = 1
	Score=76.92
	(10 divided by 13 multiplied by 100)

2.5 Development and Reliability of ASR Grading

This section will briefly discuss the development and reliability of ASR scoring system since the beginning of the Japanese EI + ASR studies in 2009. An ASR engine called Julius (Lee & Kawahara, 2009) has used three different models over time to improve its accuracy in recognizing test-takers' speech.

The first model (System I) used grammar-based speech recognition with the assumption that participants would be able to repeat model prompts as closely as possible (Matsushita & LeGare, 2010). However, the assumption was only true for advanced-level learners and the ASR grading system was not able to recognize lower-level learners' 'interlanguage' as well.

Then the second model (System II) was developed using statistical language models with L1 corpora (Matsushita, et al., 2010). Nevertheless, the system was only able to capture speech errors that L1 speakers would make because of the strong influence from the L1 corpora. Moreover, the grading had a tendency to be harsher than human raters towards interlanguage-influenced speech.

To account for problems that were brought up with the first two models, the current ASR model (System III) was developed using learner corpora, so that the ASR could predict all the possible morpheme outputs that language learners would come up with (Matsushita, 2011). Thus, the language model in the system includes not only EI items and items from the Corpus of Spoken Japanese (CSJ), but also an analogical modeling generated corpus that incorporated learners' EI performance from pilot studies and previous data collection (Matsushita, 2011). It was reported that this system had improved the ASR recognition accuracy and that the human-ASR correlation was higher for system III than for the other two systems. Table 2 shows the Cohen's kappa values of System I, II, and III, which shows the agreement values between the scores rated by the EI + ASR grading system and human raters. Table 3 shows the correlation between human raters and the ASR grading system using System III.

Table 2: Cohen's kappa (comparison between ASR and human scoring)

	Cohen's kappa
System I	0.65
System II	0.64
System III	0.71
Human	0.86

*.61-.81 represents substantial agreement (Landis & Koch, 1977)

Table 3: Human-ASR correlation

	<i>r</i>	<i>r</i> ²	<i>P</i>
Sentence N=13920	0.9088	0.8259	< .0001
Subject N=232	0.9852	0.9705	< .0001

Tables 2 and 3 show that the test scores obtained from the ASR grading are reliable in that ASR is successfully providing consistent test scores comparable to human raters. Matsushita stated that the ASR using System III has yielded promising results for real-time grading for adaptive testing that can be used for an EI-based online testing system (Matsushita, 2011).

2.6. Validation of EI

Now that the procedure and the reliability of the ASR scoring have been reported, this section will discuss the validity of acquired EI and ASR data to see if the scores represent participants' oral-aural abilities. Specifically, this section will follow the five validation criteria for EI tasks as suggested by Tomita and others (Tomita, et al., 2009).

1. EI performance is not greatly influenced by participants' rote repetition abilities.
2. Participants' short-term memory capacity does not greatly influence EI's performance.
3. Participants attend to meaning, not form, during EI tasks.
4. Multiple L2 implicit knowledge measures are used to ensure that EI actually taps into implicit knowledge.
5. EI performance is generally unaffected by a) unbalanced linguistic complexity and/or difficulty; b) instructions about how to do EI tasks; c) task difficulty.

Discussion of rote repetition abilities and short-term memory with regard to the length of EI sentences will be addressed using the first two criteria. Then, discussion of whether participants had focused their attention on meaning as well as implicit knowledge will be

addressed using the third and fourth criteria. The fifth criterion regarding linguistic and task difficulty will be discussed only briefly since these topics have already been addressed in the previous section in connection with EI task procedures and item selection.

2.6.1 Rote repetition abilities and short-term memory. In terms of criteria 1 and 2 regarding rote repetition abilities and short-term memory capacity, it is important to note that participants' rote repetition abilities and short-term memory were not measured during the data collection. This means that there is no solid evidence concerning whether or not participants' rote repetition abilities and short-term memory capacity greatly influenced their EI performance. However, in an attempt to satisfy these criteria, EI sentences were created and refined in pilot studies to come up with appropriate lengths to prevent participants from relying on their rote repetition abilities and short-term memory. The following section will briefly discuss the process of sentence selection and refinement.

In the JSL EI pilot study, 60 sentences were chosen from the Simple Performance-Oriented Test (SPOT) (Kobayashi, Ford-Niwa, & Yamamoto, 1996) by limiting sentence lengths to within a range of 10 to 25 morae. As mentioned earlier, the sentence length was adapted from the English EI research by converting syllable-based length values to mora-based values (C. R. Graham, et al., 2008).

There was a problem with this set of items because they failed to differentiate advanced-level learners and native speakers whose scores were clustered together in the upper range. To avoid this ceiling effect, more than 100 sentences from the textbook and the CSJ were created and recorded, and 60 items were chosen from among them based on item difficulty (i.e., sentence length, item level, grammatical patterns, vocabulary, and the quality of sound recordings). The

maximum sentence length was extended from 25 to 31 to provide challenging items for advanced-level learners to differentiate them from the native speakers.

With the new items from the textbooks and CSJ, another pilot study was conducted to test these items. A total of 56 participants took this set of EI items: 15 native speakers, 31 advanced-level learners and 10 beginning-level learners. These advanced-level learners included anyone enrolled in third- and fourth-year level classes (i.e., including both instructed and self-instructed learners), and all of the beginning-level learners were those who were enrolled in the first-year classes. There were no intermediate-level participants for this data collection because there were no intermediate-level classes offered at BYU during the data collection. Their average scores showed that these items differentiated advanced-level learners and the native speakers well.

Table 4 shows the average EI scores of the native speakers, advanced-level learners, and beginning-level learners as measured by the ASR system II:

Table 4: Average EI scores in pilot study

Level	Number of participants	Average EI scores of 60 sentences as measured by ASR (max=100)
Beginning-level learners	10	38.79
Advanced-level learners	31	72.07
Native speakers	15	96.50

The results clearly showed that the native speakers were able to imitate the newly created items almost perfectly, whereas advanced-level learners failed to do so as accurately. From this result, one can infer that these test items were long enough that advanced-level learners could not rely only on their rote repetition abilities to repeat back the sentences. Moreover, native speakers' average scores show that they had no problem repeating back these sentences. With the fact that native speakers were capable of handling these items, it can be inferred that these sentences were not excessively long, and that their EI performance was not influenced greatly by their short-

term memory capacity. For more on the relationship between short-term memory and EI, see Okura's study on EI performance of learners of English and working memory (referring to structures and processes used for temporarily storing and manipulating information), in which no significant correlation was found (Okura, 2011).

2.6.2 Implicit knowledge and meaning. With regard to criteria 3 and 4 dealing with focusing on meaning and implicit knowledge during the EI task, participants were required to wait for 3 seconds after listening to the sentence with countdown beep sounds. This 3-second pause was adopted with the assumption that it would provide a chance for participants to focus on meaning by allowing them to process linguistic vocabulary and structure they heard. With regard to the 3-second delay, McDade, Simpson, and Lamb had reported that it provided time for participants to process meaning and prevented them from just parroting or relying on their rote repetition abilities. They reported that participants could repeat what they did not understand as long as imitation was immediate, but they were unable to after a 3-second delay (McDade, Simpson, & Lamb, 1982).

Another treatment that indirectly let participants focus on meaning was the placement of grammatical patterns toward the middle and final parts of the sentence. As mentioned previously, the middle part of the sentence in English and the final part of the sentence in Japanese are the most difficult parts to remember (Naiman, 1974; Yoshioka, 1991). The assumption is that successful repetition of the sentence required participants to process the grammatical patterns found in the most difficult places in the sentence. If they are relying only on their short-term memory alone without comprehension, they would not be able to repeat the sentence back as easily. Thus, the appropriate placement of target grammatical structures may encourage participants to focus on meaning.

However, no other measurements of implicit knowledge (e.g., asking comprehension questions or requiring participants to correct ungrammatical sentences) were conducted to encourage participants to attend to meaning. These measurements were not conducted because the primary purpose of the study was to improve the accuracy rate of the ASR scoring system, and this is part of an ongoing research project here at BYU.

Another major limitation of this EI test is the fact that contexts were not given to participants during the EI task. This is problematic because context plays a significant role in determining the meaning of the sentence along with syntax and vocabulary to help learners focus on meaning rather than form. Thus, it is possible that participants might not have focused on meaning as much as expected.

Nonetheless, the theory behind the EI test is that participants' ability to correctly imitate the target language structure assumes that successfully repeated structures must be part of the learner's interlanguage system, and is tapping into implicit knowledge to an extent (Erlam, 2006). Moreover, Sachs reported that while memory of the form of a sentence is lost shortly after the sentence has been comprehended, memory of the meaning of the sentence is retained for a significantly longer period (Sachs, 1967). Thus, it may be inferred that the 3-second pause is useful in differentiating participants who can only remember the form of the sentence from those, who can comprehend the meaning of the sentence. It is assumed that the ones who have the ability to comprehend and process the meaning of the sentence are able to imitate the sentence better than the ones who cannot.

2.6.3 Linguistic and task difficulty. Lastly, to account for linguistic difficulty and complexity discussed in criterion 5, each item was carefully created by selecting and adapting grammatical patterns and vocabulary for beginning-, intermediate-, and advanced-level learners.

Sentence lengths as well as the number of *bunsetsu* were also taken into consideration. The test was probably too difficult for beginning- and intermediate-level learners who had to deal with unknown grammatical patterns and vocabulary found in advanced-level and superior-level items. However, various levels of sentences were needed in order to balance out the difficulty in the test, and beginning- and intermediate-level items were also incorporated equally.

In terms of task difficulty, English instruction on how to use the headphones and microphones was given to participants, and instruction on the EI task was also provided in English as well. Moreover, participants were able to practice the EI task as much as they needed to before taking the actual test. Thus, it was assumed that all of the participants did not have difficulty with the technology and the EI task itself.

2.7 Implications of this Study

The purpose of this study is to examine the usefulness³ of the ASR scoring to evaluate differences among learners of Japanese, and thus it is important that the test construct be valid in that the test scores represent their oral-aural ability that EI is trying to measure. To increase the validity of the test construct, the EI sentences were created carefully in terms of their sentence length and the placement of grammatical items to make sure that it is not testing participants' rote repetition ability or short-term memory. Moreover, the EI task required participants to wait for 3 seconds to make sure that they were able to process the sentences with a focus on meaning. The results of the pilot study indicated that these items were able to differentiate native speakers and advanced-level non-native speakers, and it was assumed that the participants were focused on meaning with the test items, and the scores represent their oral-aural ability.

In terms of linguistic and task difficulty, different levels of test items were incorporated equally to balance out the test difficulty, to make sure, that the test was not too difficult or too

³ See note 1 on page 6 for the definition of usefulness for this thesis.

easy. The procedure of the EI task was explained to all of the participants to make sure that they were familiar with the task, and it was assumed that the difficulty of the test itself did not cause any problems that affected their performance on the test.

The biggest limitation with the EI task is the lack of other implicit measurements (i.e., comprehension questions, grammaticality tests etc.) to support the results. Thus, the EI + ASR results would only reflect their ability to comprehend and repeat back the sentence. Moreover, context was not given to participants during the task, which also indicates that participants might not have focused on meaning as much as expected. Thus, one must be cautious regarding the validity of this EI test and any conclusions drawn on the basis of the data obtained from it.

Although an imitation task by itself may not be sufficient to draw any inferences, it is assumed that EI is reconstructive and that participants' ability to imitate represents their underlying interlanguage system. With that assumption in mind, their EI + ASR scores would be meaningful in evaluating differences among learners of Japanese based on their class levels and learning environment. Although technical problems with the ASR scoring can be expected, technology has become a vital part of language instruction and research, and the correlation between the current ASR (system III) and human rating is, at any rate, quite high (See tables 3 and 4) (Matsushita, 2011). If the system proves its reliability and usefulness for SLA studies, then researchers can consider using technology to collect a large amount of data with EI and ASR, which would be much faster, cheaper, more consistent, and more objective compared to human raters and researchers. To this end, further analysis of the participants' scores and demographic data will be conducted to examine the usefulness and reliability of the EI + ASR grading system in evaluating differences among learners of Japanese.

2.8 Research Questions and Hypotheses

As discussed in the introduction, this thesis will evaluate differences among beginning-, intermediate-, and advanced-level instructed learners, advanced-level instructed learners and self-instructed learners without post-mission instruction, as well as advanced-level self-instructed learners with and without post-mission instruction by using the EI + ASR scoring system to discuss its usefulness in SLA research. As reported previously, beginning-level participants are first-year students, intermediate-level participants are second-year students, and advanced-level participants are third- and fourth-year students. Self-instructed learners were divided into two groups: self-instructed learners with and without post-mission instruction. The research questions and their respective hypotheses are discussed below:

1. Can EI and ASR grading differentiate beginning-, intermediate-, and advanced-level instructed learners, leading to its possible use as a placement test?

The null hypothesis is that there is no difference among beginning-, intermediate-, and advanced-level learners in their EI performance on the EI task as measured by ASR grading. If the differences are not significant, then it might be due to misclassified participants or because of technological problems with the EI + ASR grading system. Therefore, participants must be carefully selected using their demographic data to make sure that participants in each group represent their level of proficiency as determined by class-level. Then, the results will be analyzed to evaluate the reliability of the ASR scoring. If the null hypothesis is rejected, then the results may indicate that EI with ASR grading can differentiate beginning-, intermediate-, and advanced-level instructed learners. If this is the case, then one can consider using the EI + ASR grading as a possible tool for placing students in classes according to their proficiency level.

2. Can EI and ASR grading differentiate advanced-level instructed learners and self-instructed learners without post-mission instruction?

The null hypothesis is that there is no difference between advanced-level instructed learners and self-instructed learners without post-mission instruction on their EI performance as measured by the ASR grading. If there is no significant difference, then it would indicate that advanced-level instructed learners and self-instructed learners with post-mission instruction have the same language oral-aural ability as measured by the EI + ASR grading, and that learning environment does not influence language ability. However, if the null hypothesis is rejected, the results might highlight benefits that these learners have acquired from their learning environment. Self-instructed learners who had learned Japanese in a naturalistic learning environment may have advantages in vocabulary acquisition, sociolinguistic features, and oral fluency. On the other hand, instructed learners may be able to better notice EI items with marked elements in Japanese.

3. Can EI and ASR grading differentiate self-instructed learners with and without post-mission instruction?

Assuming hypothetically that all self-instructed learners have the same level of proficiency when they return from their 1.5-1.8-year missions, and their post-mission instruction has no effect on their proficiency, the null hypothesis is that there is no difference between self-instructed learners with and without post-mission instruction. As Russell (2005, 2011, in press) reported on self-instructed learners in his study using a free oral task, there was no significant difference between self-instructed learners with and without post-mission classroom instruction within the first couple of years of observation following their 1.8-year volunteer service in Japan. If there is no significance, then the findings of this study would indirectly support Russell's

findings about the effects of instruction not seen within the first few years after their mission experience. However, if there is a significant difference between self-instructed learners with and without post-mission instruction, then something about the effects of instruction can be said depending on which group outperforms the other.

EI is different from free oral tasks in that participants cannot avoid difficult or less frequent grammatical patterns. Thus, results from the EI task and ASR grading may yield different findings. If the null hypothesis is successfully rejected, then their difference will indicate either signs of language development or loss after leaving their naturalistic learning environment.

Chapter 3: Procedure

This section will discuss decisions concerning how participants were selected, excluded, and grouped together using their demographic data. The participants had already taken the EI test in the computer lab. While there, they were instructed to listen to and repeat back 60 Japanese sentences as closely as possible. Their recorded utterances were graded by the ASR grading after the data collection. Their overall EI scores as well as individual item scores were used for statistical analysis to see if the system successfully differentiated beginning-, intermediate-, advanced-level instructed learners; advanced-level instructed learners and self-instructed learners without post-mission instruction; and self-instructed learners with and without post-mission instruction. Descriptions of the statistical analysis used in this project will be discussed. Further discussion of the procedures used in the EI task and ASR scoring will not be given here because they have been explained in detail in the review of literature.

3.1 Participants

All Japanese language students in the program were required to take the test and announcements were made in every class in the Japanese department. The data collection was held at the end of fall semester in 2010, in November. Although not everyone participated, there was a total of 238 students who took the EI test.

According to participants' demographic data collected in the pre-test survey, there were 224 English native speakers, 8 Japanese native speakers, 4 Spanish native speakers, 3 Mandarin native speakers, 6 Korean native speakers and 1 Russian native speaker. Eight participants reported that they had two native languages and thus the sum of the native speakers exceeds the total number of participants. There were 136 male and 102 female participants.

Although instructions on how to use headphones and a microphone and practice sessions were given to all of the participants, eight of the participants' data were not successfully recorded because these participants could not finish the test due to schedule conflicts, and some of the participants' recordings were not loud enough. Thus, analysis was conducted using the utterances of the remaining 230 participants.

Participants were divided into groups according to their class levels and based on their 1.5-1.8-year mission experience in Japan as well as the number of semesters of instruction they had received. The research questions for this thesis and grouping criteria for each question are discussed below:

1. Can EI and ASR grading differentiate beginning-, intermediate-, and advanced-level instructed learners, leading to its possible use as a placement test?

To address this research question, instructed learners were divided according to class level for comparison. Beginning-level participants included participants who were enrolled in the first-year Japanese classes. There were two truly bilingual heritage learners in the first-year classes who indicated that they spoke Japanese with family members growing up, and they were excluded from the data to avoid skewed results. Intermediate-level participants included students who were taking second-year Japanese classes. There were no false beginners. Advanced-level participants included students who were taking third-year and fourth-year Japanese classes. Advanced-level instructed learners who were enrolled in the fourth-year Japanese classes were also included in this group due to the limited number of advanced-level participants in the third-year classes. Any Japanese native speakers, truly bilingual heritage learners or primarily self-instructed learners were excluded from this group. Table 5 shows the number of participants in each group.

Table 5: Number of instructed learners

Groups	Number of participants
1 st year beginning-level instructed learners	79
2 nd year intermediate-level instructed learners	34
3 rd & 4 th year advanced-level instructed learners	17

2. Can EI and ASR grading differentiate advanced-level instructed learners and self-instructed learners without post-mission instruction?

Instructed and self-instructed learners in advanced-level classes were grouped according to their class level and learning environment. Advanced-level instructed learners were the same group as the ones in the first research question. Self-instructed learners, on the other hand, were those who were enrolled in their first post-mission course at the third-year level. Thus, self-instructed learners who had two or more semesters of instruction had been excluded from the group to minimize the classroom instruction influence. Truly bilingual heritage speakers and those who lived in Japan and received formal instruction for an extended period of time (i.e., more than a year) were also excluded from this group. Table 6 summarizes the number of participants in each group:

Table 6: Number of instructed and self-instructed learners

Groups	Number of participants
3 rd & 4 th year advanced-level instructed learners	17
3 rd year 1st semester advanced self-instructed learners	38

3. Can EI and ASR grading differentiate self-instructed learners with and without post-mission instruction?

Self-instructed learners were divided into two groups based on the number of semesters of instruction they had received on the assumption that they had comparable language proficiency upon returning from their 1.5-1.8-year missions in Japan. Those who were in their

first post-mission course were the same group as in research question 2 because they had minimal influence from classroom instruction. The other group was made up of advanced-level self-instructed learners who had received two or more semesters of instruction. All of the self-instructed learners who had taken two or more semesters of instruction were grouped together and included in the second group. Classroom-instructed learners, native speakers, truly bilingual heritage learners and those who had lived in Japan and received extensive instruction were excluded from the group. Table 7 shows the number of participants in each group:

Table 7: Number of self-instructed learners with and without instruction

Self-instructed learners at advanced-level	Number of participants
1st semester advanced-level self-instructed learners	38
2 or more semesters advanced-level self-instructed learners	52

3.2 Statistical Analysis

The purpose of this study is to discuss the usefulness of EI and ASR for SLA studies by investigating its ability to differentiate participants based on their class level and learning environment to evaluate differences in their language ability. Thus, participants were placed in different groups based on their membership determined by their class level, learning environment, and the number of semesters of instruction they had received. Since the criterion variable is categorical predetermined by their class level and learning environment as discussed in the previous section, linear discriminant analysis will be utilized with the hypothetical assumption that participants' language proficiency is equal, and they are equally distributed in each group. Although it is difficult to assume that students were perfectly grouped into their level with equal proficiency, participants were grouped carefully, and those that would skew the results have been excluded as discussed previously.

Discriminant analysis will take participants' overall EI scores as well as their individual item scores into consideration to classify participants into appropriate groups according to their performance. In other words, if all of the participants were placed in the appropriate group, then they will be classified back into their own group with a 0 percent error rate. However, if their performance is not in accordance with the rest of the group to which they belong, then they will be misclassified into the groups that their EI scores would indicate. If this is the case, then the error rate will rise indicating that these participants were not placed in the right group. Thus, in order for discriminant analysis to successfully differentiate participants based on their class level, learning environment, and semesters of instruction they received, it has to be assumed that these participants have equal ability within their own group. Higher error rates would indicate that there are problems with either the test construct or grouping of participants. An error rate lower than 0.67 for three groups, or 0.50 for two groups, would show that the system is capable of predicting behavior, and that the grouping was not by chance.

Moreover, discriminant analysis is useful in selecting sets of significant items that best discriminate the participants according to their categories by using a stepwise procedure. These specific items play a significant role in lowering the error rate, and it is likely that they would yield a lower error rate by themselves than with the overall EI scores with all items taken into consideration. Thus, further qualitative item analysis on these significant items will yield insightful findings for evaluating differences in participants' language abilities.

The error rate calculated by discriminant analysis and the characteristics of misclassified participants will be analyzed in relation to the demographic data to see if the findings are accurate. Although imitation by itself may not be sufficient to evaluate language differences, the

results might be useful in identifying certain patterns of development that might be valuable for instructors as well as SLA researchers conducting studies in the future.

Chapter 4: Results

Discriminant analysis was conducted using the overall EI performance scores as measured by ASR to compare 1) beginning-, intermediate-, and advanced-level instructed learners, 2) advanced-level instructed learners and self-instructed learners without post-mission instruction, and 3) advanced-level self-instructed learners with and without post-mission instruction. Significant items that discriminated participants were selected using a stepwise procedure and further discriminant analysis was conducted using the participants' performance on these items. This section will list the number of observations classified into the groups to which discriminant analysis determined these participants actually belong. An error rate or percentage of misplacement of participants calculated by discriminant analysis will be provided. There will be two tables for each research question using 1) the overall EI scores and 2) discriminant item scores. Research questions and the results are listed together as follows.

4.1 Classroom Instructed Learners

1. Can EI and ASR grading differentiate beginning-, intermediate-, and advanced-level learners, leading to its possible use as a placement test?

4.1.1 Overall EI score results. The overall average EI scores of beginning-, intermediate-, and advanced-level learners were 34.71, 45.28, and 52.22 respectively. Based on the results produced by discriminant analysis using participants' overall EI scores calculated by ASR, out of 79 beginning-level learners, 63 of them were classified as beginning-level, 12 of them as intermediate-level, and 4 of them as advanced-level. Out of 34 intermediate-level participants, 12 of them were classified as beginning-level, 10 of them as intermediate-level, and 12 as advanced-level. Out of 17 advanced-level participants, none of them were classified as beginning-level, 6 of them were classified as intermediate-level, and 11 of them were classified as advanced-level.

The error rate or percentage of misplacement of participants was 0.42. Table 8 provides a summary:

Table 8: Discriminant analysis with instructed learners (overall EI scores)

Level	Classified as beginning-level	Classified as intermediate-level	Classified as advanced-level	Total
Beginning-level participants	63	12	4	79
Intermediate-level participants	12	10	12	34
Advanced-level participants	0	6	11	17
Total	75	28	27	130

*Overall average EI scores: beginning 34.71, intermediate 45.28, advanced 52.22.

*Error rate=0.42

4.1.2 Discriminant item results. According to discriminant analysis, items that best discriminated beginning-, intermediate-, and advanced-level learners were items 35, 42, and 15. Items are placed in the order of significance. Table 9 provides the average scores and item levels as discussed in the review of literature:

Table 9: Discriminant items and average scores of instructed learners

Items*	Beginning	Intermediate	Advanced	Item level
35. 実は仕事、辞めさせられちゃったんです。	20.32	48.30	66.87	intermediate
42. これを見ると面白いことに気が付くはずです。	23.65	24.27	58.29	advanced
15. 明日はいい天気になるね。	60.56	87.78	85.97	beginning

*See Appendix II for Romanized version and English translations.

Participants' average scores show constant improvement across levels with item 35. Item 42 shows a plateau between beginning and intermediate-level learners but significant improvement is seen with advanced-level learners. Item 15, on the other hand, shows significant

improvement between beginning- and intermediate-level, but no further improvement is seen between intermediate- and advanced-level learners.

With participants' scores on these three items, out of 79 beginning-level learners, 66 of them were classified as beginning-level, 11 of them as intermediate, and 2 of them as advanced-level. Out of 34 intermediate-level participants, 8 of them were classified as beginning-level, 19 as intermediate, and 7 as advanced-level. Out of 17 advanced-level learners, 1 was classified as beginning-level, 4 as intermediate-level, and 12 were classified as advanced-level. The error rate was 0.30 with these three items. See Table 10 for a summary:

Table 10: Discriminant analysis of instructed learners (discriminant items scores)

	Classified as beginning-level	Classified as intermediate	Classified as advanced	Total
Beginning-level learners	66	11	2	79
Intermediate-level learners	8	19	7	34
Advanced-level learners	1	4	12	17
Total	75	34	21	130

*Error rate = 0.30

4.2 Instructed Learners and Self-instructed Learners

2. Can EI and ASR grading differentiate advanced-level instructed learners and self-instructed learners without post-mission instruction?

4.2.1 Overall EI score results. The overall average EI scores of advanced-level instructed learners and self-instructed learners without post-mission instruction were 54.22 and 62.94. Based on the results produced by discriminant analysis using participants' overall EI scores calculated by ASR, out of 17 instructed learners, 12 of them were classified as instructed learners, and 5 of them as self-instructed learners. Out of 38 self-instructed learners, 11 of them

were categorized as instructed learners, and 27 of them as self-instructed learners. The error rate or percentage of misplacement of participants was 0.29. Table 11 provides a summary:

Table 11: Discriminant analysis of instructed and self-instructed learners (overall EI scores)

	Classified as Instructed learners	Classified as Self- instructed learners	Total
Instructed learners	12	5	17
Self-instructed learners	11	27	38
Total	23	32	55

*Overall average EI scores: instructed 54.22, uninstructed 62.94

*Error rate: 0.29

4.2.2 Discriminant item results. According to discriminant analysis, items that best discriminated instructed and self-instructed learners were items 33, 2, 1, 59, 47, 34, and 36. Again, items are placed in the order of significance. Out of seven items, self-instructed learners outperformed instructed learners with six of them. The only item on which instructed learners outperformed self-instructed learners was item 2. See Table 12 for their average scores on discriminant items:

Table 12: Discriminant items and average scores of instructed and self-instructed learners

Items*	Instructed	self- instructed	Item level
33. 調べたいことがあるときはどうしたらいい。	48.53	72.89	intermediate
2. これは、手紙じゃありません。	85.07	68.42	beginning
1. 明日行きませんか。	81.70	97.08	beginning
59. 知られてはいるんですが、見た人はあまりいないようです。	26.92	47.17	superior
47. お金に恵まれていたら、幸せに決まってるでしょう。	46.61	54.66	advanced
34. 難しく考えることはありません。	70.59	79.97	advanced
36. 冷めないうちに召し上がってください。	52.25	71.83	intermediate

*See Appendix II for Romanized version and English translations.

With participants' scores on these seven items, out of 17 instructed learners, 14 were classified as instructed learners, and 3 as self-instructed learners. Out of 38 self-instructed learners, 2 were classified as instructed learners, and 36 as self-instructed learners. The error rate was 0.11 with these seven items. See Table 13 for a summary:

Table 13: Discriminant analysis of instructed and self-instructed (discriminant items)

	Classified as Instructed learners	Classified as Self- instructed learners	Total
Instructed learners	14	3	17
Self-instructed learners	2	36	38
Total	23	32	55

*Error rate: 0.11

4.3 Self-instructed Learners with and without Post-mission Instruction

3. Can EI and ASR grading differentiate self-instructed learners with and without post-mission instruction?

4.3.1 Overall EI score results. The overall average EI scores of advanced-level self-instructed learners without instruction and with instruction were 62.94 and 64.08. Based on the results produced by discriminant analysis using the overall EI scores calculated by ASR, out of the 38 self-instructed learners without instruction, 16 were classified as self-instructed learners without instruction, and 22 as self-instructed learners with instruction. Out of the 52 self-instructed learners, 23 were placed as self-instructed learners without instruction, and 29 as self-instructed learners with instruction. The error rate was 0.51. Table 14 provides a summary:

Table 14: Discriminant analysis of self-instructed learners (overall EI scores)

	Classified as Without post-mission instruction	Classified as With post-mission instruction	Total
Without post-mission instruction learners	16	22	38
With post-mission instruction learners	23	29	52
Total	39	51	90

*Overall average EI scores: without post-mission instruction 62.94, with post-mission instruction 64.08

*Error rate: 0.51

4.3.2 Discriminant item results. According to discriminant analysis, the items that best discriminated instructed and self-instructed learners were items 12, 42, 32, 55, 56, 45. As with previous cases, items are placed in order of significance. Their average scores showed that both groups outperformed the other group with three items each. Table 15 shows the average scores and item difficulty:

Table 15: Discriminant items and average scores of self-instructed learners

Items*	Without post-mission instruction	With post-mission instruction	Item level
12. 児玉と申しますが、また後でご連絡いたします。	37.58	50.54	beginning
42. これを見ると面白いことに気が付くはずです。	66.42	57.92	advanced
32. この翻訳全部あなたがさせられたんだそうですね。	47.26	58.46	intermediate
55. 日本がどんな国なのか知りたがっている人がたくさんいます。	62.19	56.09	superior
56. 学生にとって十分に寝る時間を取るのは大変でしょう。	60.69	58.35	superior
45. いつの間にかよく話し掛けられるようになりました。	40.46	48.24	advanced

*See Appendix II for Romanized version and English translations.

With participants' scores on these six items, out of the 38 self-instructed learners without instruction, 31 were classified as self-instructed learners without instruction, and 7 as self-

instructed learners with instruction. Out of the 52 self-instructed learners with instruction, 14 were classified as self-instructed learners without instruction, and 38 as self-instructed learners with instruction. The error rate was 0.23 with these six items. See table 16 for a summary:

Table 16: Discriminant analysis of self-instructed learners (discriminant items)

	Classified as Without post-mission instruction	Classified as With post-mission instruction	Total
Without post-mission instruction learners	31	7	38
With post-mission instruction learners	14	38	52
Total	39	51	90

*Error rate: 0.23

Chapter 5: Analysis

The purpose of this section is to provide an analysis of the results discussed in the previous section in relation to the following research questions: 1) Can EI and ASR grading differentiate beginning-, intermediate-, and advanced-level instructed learners, leading to its possible use as a placement test?; 2) Can EI and ASR grading differentiate advanced-level instructed learners and self-instructed learners without post-mission instruction?; and 3) Can EI and ASR grading differentiate self-instructed learners with and without instruction?

5.1 Classroom Instructed Learners

1. Can EI and ASR grading differentiate beginning-, intermediate-, and advanced-level instructed learners, leading to its possible use as a placement test?

The null hypothesis was that there is no difference among beginning-, intermediate-, and advanced-level learners with their EI performance as measured by the ASR grading. The error rate was 0.42, which is lower than 0.67 and thus, it can be stated that the system was able to predict, to an extent, different behaviors across the levels of participants. It indicates that the system was able to differentiate participants based on their class-level with a 0.42 error rate, which is not very low. Whether this error rate was due to grouping problems or problems in the ASR grading is unknown at this point.

The most problematic set of participants was the intermediate-level. Participants within that level were classified into beginning-, intermediate-, and advanced-level almost equally (i.e., 12, 10, and 12 respectively). This is an interesting phenomenon because it probably indicates that intermediate learner performance ranges from that of beginning-level to advanced-level students because their language development is at various stages. Thus, it is probably safe to say that the assumption that these intermediate learners had equal Japanese ability was wrong. Their

demographic data give further insights that are fascinating. Characteristics of those who were classified higher included Korean native speakers, a fluent Chinese speaker, or participants who had had frequent contact with Japanese speakers (e.g., relatives and friends), and language-related majors such as Asian studies and linguistics. On the other hand, those who were classified lower seemed to struggle with the imitation task, especially with longer sentences. In fact, one participant just gave up and did not repeat back anything with longer sentences. However, their scores did seem to reflect performance in general.

The distribution of beginning-level participants is also interesting because the majority of participants were accurately placed in the appropriate level with some exceptions. While 63 of them were classified as beginning-level, 12 of them were classified as intermediate-level, and 4 of them as advanced-level. It is interesting to point out that 2 of the 4 participants who were classified as advanced-level were native Korean speakers. As with the other Korean speakers in the intermediate level that were classified as advanced-level learners, it can be assumed that Korean speakers may have an advantage over English native speakers. This is probably due to the fact that Korean and Japanese have similar language structure. However, there were other Korean speakers who did not perform as well, so further analysis with Korean speakers is needed to confirm this assumption.

The other two beginning-level learners who were classified as advanced-level were English native speakers. It is interesting to note that one of them actually went to Japan for 11 months as an exchange student, which was not considered to be “extensive” (i.e., more than one year) in the grouping process. The other one mentioned growing up with a Japanese mother and Japanese-fluent American father. This particular participant had an English name and it was the author’s oversight to not have excluded her from the group. However, the fact that this

participant was classified as an advanced-level learner supports the reliability of the EI + ASR grading system. Based on this result, these four beginning-level participants may need to be placed in higher-level classes in terms of their oral-aural ability because their performance on this test shows that their ability is as good as that of advanced-level learners.

It is interesting to point out the characteristics of beginning-level learners who were classified as intermediate-level. These participants were: multiple-language speakers (e.g., Spanish and Mandarin), language-related major students (e.g., Asian studies, Japanese, Linguistics), ones with frequent contact with Japanese speakers (e.g., relatives), and those with some experience with Japan (e.g., living in Japan for six months, taking Japanese classes before college, etc.).

In terms of advanced-level learners, there were six participants who were classified as intermediate-level. Their recording seemed to reflect their scores calculated by ASR, as they struggled to imitate longer and more complicated sentences. Moreover, they tended to pause often or only utter individual words they remembered instead of attempting to articulate longer sentences.

A fascinating phenomenon is observed with the three discriminant items selected by the stepwise procedure. The error rate is brought down to 0.30 with these three items due to the fact that more intermediate-level participants were classified as intermediate-level when compared with the result obtained from the overall EI scores. This is noticeably lower than that of the overall EI scores. It is also interesting to note the distribution of the item difficulty of selected discriminant items, which consisted of a beginning-level item, an intermediate-level item, and an advanced-level item—and average scores on these items further show interesting language phenomena.

First of all, a notable improvement in their average scores was found between beginning-level and intermediate-level learners with beginning-level item 15, which is characterized by a simple topic and predicate with a sentence final particle *ne*. The recorded utterances showed that beginning-level learners repeated back the sentence fairly well, but they tended to leave out some parts of the sentence that they did not seem to understand in general. On the other hand, most of the intermediate- and advanced-level learners successfully reproduced the sentence with a minor mistake of just dropping the particle *ni*. Thus, there was plateauing between intermediate- and advanced-level learners with this particular item. This result may signify that notable growth between beginning and intermediate levels may be found in sentences with simple constructions.

Secondly, constant improvement across levels was seen with intermediate-level item 35, which is characterized by topicalization followed by the causative-passive construction with an extended predicate. It was interesting listening to their recordings because beginning-level learners were generally only able to utter the first and the last part of the sentence, whereas intermediate- and advanced-level learners were able to articulate the middle part. What made the difference between intermediate- and advanced-level learners was the notion of causative and passive that most intermediate-level learners were not able to repeat back.

Thirdly, plateauing between beginning- and intermediate-level was seen with the advanced-level item 42, which is characterized by the conditional followed by a general expectation. Recorded utterances revealed the beginning- and intermediate-level learners' tendency to only recognize familiar words in the sentence (i.e., *miru* 'see' and *omosiroi* 'interesting' and *desu* 'is'). Some seemed not to understand anything and just gave up. Advanced-level learners struggled with this item as well, but generally they were able to articulate back the conditional form and the abstract nominal *koto*, as well as the general

expectation form *hazu*. From this result, it can be assumed that there was no notable growth between beginning-level and intermediate-level learners, perhaps because they were not ready to process this construct found in the advanced-level item.

Speaking of misclassified participants in Table 15 with discriminant items, one interesting phenomenon is the population of beginning- and intermediate-level participants who were classified as advanced-level. There were 2 from the beginning-level and 7 from the intermediate-level learners. Out of these 9 participants, 3 of them were, again, Korean native speakers and one of them was a fluent Chinese speaker. Others were native English speakers with rich experience with Japanese people and culture (e.g., exchange student, frequent contact with Japanese speakers, and watching anime).

One advanced-level participant, on the other hand, was classified as beginning-level because of poor performance with these sets of items. The recorded utterance had some of the beginning-level-like characteristics such as leaving words out and mumbling, although the utterances did seem to be faster than typical beginning-level learners. This result may indicate that the appropriateness of the participant's enrollment in advanced-level classes might be questionable. Moreover, there were four other advanced-level participants who were classified as intermediate-level. Their recordings revealed that they were not able to do well with these items either.

In conclusion, although the error rate was 0.42 with overall EI scores of beginning-, intermediate-, and advanced-level learners, it can be assumed that the EI and ASR grading system successfully classified participants according to their performance on the EI test, with the evidence that the participants were, in general, appropriately classified according to their EI performance—based on those who were misclassified as well (e.g., Korean speakers in the

beginning-level classes etc.). Intermediate-level participants were the most problematic because their interlanguage development is at varying stages, which caused discriminant analysis to classify them across all levels. However, misclassified participants' demographic data and performance revealed that these learners were actually placed appropriately in general, which indirectly confirms the reliability of the EI + ASR grading. To say the least, the EI + ASR grading was clearly successful in differentiating beginning- and advanced-level learners. Thus, the EI + ASR grading system can be used to at least identify individuals who have beginning-level-like or advanced-level-like oral proficiency.

Moreover, three discriminant items were effective in showing participants' language development across levels. A notable improvement in their average scores was seen between beginning- and intermediate-level learners with the beginning-level item, and between intermediate- and advanced-level learners with the advanced-level item. If their scores did not improve and seemed to be plateauing, the assumption was that there was no notable difference in their language development. However, further analysis of their comprehension with these sentences is needed before making any generalizations.

5.2 Instructed Learners and Self-instructed Learners

2. Can EI and ASR grading differentiate advanced-level instructed learners and self-instructed learners without post-mission instruction?

The null hypothesis was that there is no difference between advanced-level instructed and self-instructed learners with their EI performance as measured by the ASR grading. The error rate calculated by discriminant analysis was 0.29, which is fairly far from 0.50. This result allows an inference with greater confidence that the difference between instructed and self-instructed learners was most likely not by chance, and that they were classified appropriately.

Their average overall EI scores also revealed that self-instructed learners outperformed instructed learners in this study. Furthermore, self-instructed learners' average scores on discriminant items were higher than those of the instructed learners in general. Out of seven discriminant items, self-instructed learners outperformed instructed learners on six of them in terms of their average scores. The error rate with these seven discriminant items declined to 0.11, which is significantly low. This result seems to indicate that, among the participants in this study, self-instructed learners had a higher proficiency level than instructed learners. This finding may support the notion that self-instructed learners are better at acquiring vocabulary and that they have better oral fluency than instructed learners. At the very least, they may have probably acquired better oral-aural ability than instructed learners who have had less experience of being immersed in the target culture.

However, it is important to note that instructed learners had not had as much time studying and using the language compared to self-instructed learners, who had had 1.5-1.8 years to focus on improving their language ability. In other words, the assumption that their proficiency was at an equal level was disconfirmed. Contributing to the differences between the two groups may be the fact that instructed learners have time off between semesters and that they have limited exposure to the language in a classroom setting when they study. Although the difference between the two was statistically significant, it is noteworthy to mention that instructed learners' overall average score is not so far from that of self-instructed learners—which may indicate that instructed learners might have acquired the target language in a more efficient way with limited amounts of exposure to the target language. Further analysis with other measurements of implicit knowledge is needed to study the effects of learning environment on this note.

In terms of other effects of instruction, there was one discriminant item on which instructed learners outperformed self-instructed learners (item 2). This item is a beginning-level item and has a very simple construct of a deictic expression (i.e., *kore* ‘this’) with a topic particle *wa* and a negation. One interesting note about this sentence is that it does not have marked elements that self-instructed learners might not have understood. Instead, this item actually stemmed from the textbook and there is a possibility that these instructed learners might have remembered this sentence structure from their classroom experience. As indicated in the item selection discussion of this thesis, beginning-level items were particularly difficult to modify because they were already simple and it was difficult to make many changes.

In sum, advanced-level self-instructed learners outperformed instructed learners and their difference was significant according to the results obtained from the EI + ASR system and discriminant analysis used in this study. Notable effects of instruction were not found with instructed learners’ performance although they outperformed self-instructed learners with one of the seven discriminant items selected. However, that may be due to the fact that the item stemmed from the textbook.

5.3 Self-instructed Learners with and without Post-mission Instruction

3. Can EI and ASR grading differentiate self-instructed learners with and without post-mission instruction?

The null hypothesis was that there is no difference between self-instructed learners with and without significant post-mission classroom instruction. Although self-instructed learners with post-mission classroom experience had slightly higher overall EI scores than those without post-mission instruction, the error rate calculated by discriminant analysis with the overall EI scores was 0.51, which is above 0.50. Strictly speaking, this indicates that the system was not

able to discriminate these two groups with the overall EI scores. This result indirectly confirms Russell's finding that there is no significant difference in the oral communication skills of these two groups within two years after their naturalistic learning experience.

The error rate calculated with the discriminant items is 0.23, which is far below 0.50, and these items may be useful in describing their language difference. There were six items in which both groups outperformed the other with three items each. In other words, they are competing against each other equally on these six discriminant items. The order of significance and outperforming group in terms of their average scores is 1) item 12 (with instruction group), 2) item 42 (without instruction group), item 32 (with instruction group), item 55 (without instruction group), item 56 (without instruction group) and item 45 (with instruction group). Although one of the groups outperforms the other with their average scores on each of these discriminant items, their scores on these items are actually quite similar, and it is difficult to draw any generalizations about differences in their language development from the data set at this point.

In terms of item analysis, items 12 and 32 are particularly interesting because these items include 'marked' elements of Japanese language (e.g., honorifics, polite speech, causative-passive), and self-instructed learners with post-mission instruction outperformed the ones without post-mission instruction. However, from listening to their recordings, participants from both groups seemed to struggle with these particular items. It seemed like they either rephrased the sentences in a simpler form or were not able to repeat anything back. It is interesting to note that the notion of self-instructed learners' tendency to rely on their noncomplex L2 system was seen with their behavior in rephrasing the sentences in a simpler form with this item. Perhaps those with post-mission instruction were able to repeat back utterances that were closer to the

original utterance. Thus, scores of self-instructed learners without post-mission instruction on these items were lower, perhaps, because they were more relying heavily on their noncomplex L2 system. Further research is needed here to differentiate these two groups.

It is difficult to assume that self-instructed learners are benefiting from their post-mission classroom experiences at this point, at least with regard to the skills being tested by EI. It is challenging because self-instructed learners are no longer immersed in the language culture after their mission, and they may have lost some proficiency over time because their L2 exposure time in the classroom is limited. The results can at least be used, however, to infer that post-mission instruction is helpful in maintaining their language proficiency.

Chapter 6: Conclusion

The purpose of this thesis was to examine the usefulness of ASR-graded scores obtained through EI testing as an SLA research tool and to evaluate differences between instructed and self-instructed learners at BYU. In particular, the following three research questions were addressed: 1) Can EI and ASR grading differentiate beginning-, intermediate-, and advanced-level instructed learners, leading to its possible use as a placement test?; 2) Can EI and ASR grading differentiate advanced-level instructed learners and self-instructed learners without post-mission instruction?; and 3) Can EI and ASR grading differentiate self-instructed learners with and without post-mission instruction? Participants' overall EI performance as well as individual item performance were measured by ASR scoring and analyzed using linear discriminant analysis.

In relation to the first research question, the results indicated that the EI + ASR grading system could be used to differentiate beginning- and advanced-level learners with an error rate of 0.42. One major problem was intermediate-level learners whose proficiency was too variable and the presence of outliers that were not excluded. However, further analysis of demographic data of the participants supported the reliability of the EI + ASR in that the system was able to identify those outliers and differentiate participants based on their oral-aural ability. Thus, it can be assumed that the system can be used to measure participants' oral-aural abilities in order to put them into appropriate categories according to proficiency as measured by ASR grading. The analysis of discriminant items was useful in locating significant improvement or plateauing that had occurred among beginning-, intermediate-, and advanced-level learners. The three discriminant items reduced the error rate to 0.11, which indicates that these three items clearly differentiate beginning-, intermediate-, and advanced-level learners.

In terms of the second research question, the results indicated that self-instructed learners outperformed instructed learners, and that the EI + ASR grading was successful in differentiating them with an error rate of 0.29. Selected discriminant items reduced the rate to 0.11. Although the self-instructed learners did not have formal instruction during their two-year missions, their naturalistic learning environment in the target culture probably refined their oral-aural abilities. In terms of the effects of instruction, there was no significant finding in instructed learners' oral-aural ability to recognize marked elements of the target language, although they outperformed self-instructed learners on one of the discriminant items. However, the fact that instructed learners had had less time practicing Japanese has to be considered when interpreting the data, and other measurements of implicit knowledge are clearly needed for further investigation.

With regard to the third research question, no significant difference was found between advanced-level self-instructed learners with and without post-mission classroom instruction on their overall EI scores, with an error rate of 0.51. This finding aligns with Russell's studies on the effect of instruction, in that no significant difference was found between groups of self-instructed learners with and without post-mission instruction within the first few years after their mission (Russell, 2005, 2011, in press). Although discriminant items came up with a lower error rate of 0.29 and self-instructed learners performed better with EI items that contained 'marked' elements, their average scores on those items were quite close. However, item analysis revealed self-instructed learners' tendencies to rely on their noncomplex L2 system with some of the difficult items within both groups of participants. Further investigation is needed to examine the characteristics of those who rely heavily on their noncomplex L2 system. It can at least be inferred from the results, however, that self-instructed learners' Japanese oral-aural proficiency

can be maintained or slightly improved through post-mission classroom instruction as measured by the EI + ASR system.

In conclusion, EI and ASR could be useful for SLA research if the task includes other measurements of implicit knowledge to verify the EI results. Grammatical patterns could be incorporated more effectively for SLA research because test-takers cannot avoid using them when elicited. Sentence length and the placement of targeted grammatical patterns should be considered when creating sentences for EI. Furthermore, context should be provided to make sure that learners are focusing on meaning and not just the form in a sentence.

In terms of language pedagogy, it seems that EI and ASR can be a useful tool in tracking students' progress or in the context of program review, as technology becomes more and more useful in providing feedback that is fast, objective, consistent, and reliable. The reliability of ASR is promising and it is possible to have EI-based online tasks in the near future that can allow test takers to receive consistent, meaningful, and nearly instantaneous feedback on their language ability at any time. The advancement of this type of technology can encourage more learning, and it is the author's hope that EI and ASR will play an important role in SLA studies as an effective tool in collecting a large amount of meaningful data.

6.1 Implications

This current model of EI and ASR can be used in many ways. For example, students can take the test every semester to track their progress and identify their level of proficiency. As discussed in the study, scores can, to some extent, be used to place students at appropriate level based on their oral-aural ability. Their scores and recorded sounds can be used to identify and diagnose their strengths and weaknesses that teachers can address with them. Scores can be used and kept to conduct program review to see how students are doing overall within a language

program. Different language programs and institutions can be compared by using the scores as well. If other implicit measurements are incorporated, a large amount of SLA data can be collected in an efficient, timely, objective, and consistent manner—which will be useful for language teachers. With the advancement of technology, EI and ASR will probably be available online as a computer adaptive test and anyone who has access to the Internet and appropriate headsets will be able to get feedback on their language ability.

With regard to sentence length and complexity, it is important for language instructors to teach their students with materials that are appropriate for their level of proficiency. If difficult sentence patterns are introduced in the first and second years, students will most likely not be able to process meaning quickly, as demonstrated by their performance on the advanced-level items in this study, and it may be necessary to give them more time or to simplify sentence patterns. However, the idea of elicited imitation can be incorporated in approximating the proficiency of language learners. Some of the discriminant items discussed in this thesis can be used to test students, and their performance can quickly highlight their levels of language proficiency to an extent.

In terms of application to students at BYU, it may be necessary to provide some kind of support for advanced-level instructed learners whose oral-aural ability is not as high as that of self-instructed learners in class. Both instructed and self-instructed learners are currently enrolled in the same classes at BYU, and instructed learners may feel out of place because the majority of the students are self-instructed learners whose proficiency seems to be higher than theirs (as seen in the number of participants in this study).

6.2 Future Work

First, the greatest limitations to this study were technological problems and it is important to prevent them as much as possible in the future research. Some of the results may have been skewed because of badly recorded utterances that may have affected overall scores significantly. Unfortunately, some of the participants' tests were recorded poorly, and their scores on all of the items were very low. Problems with voice quality (e.g., mumbling, unclear articulation, low volume, unrecognizable utterances, etc.) were seen in those who were classified into a lower level. There were also problems with some continuous buzzing sounds and background noise that might have affected participants' scores. In terms of scoring accuracy, however, their scores seemed to be fairly accurate for the most part, but sometimes there were scores that did not seem to reflect their utterance. It might be important to check the computer system to observe the number of correct morae, inserted morae, deleted morae, and substituted morae for those individual items to locate any technological problems.

Second, it is crucial to incorporate other implicit knowledge measurements. In terms of the test construct, the EI test used in this study was only testing participants' aural and oral ability because there were no other implicit knowledge measurements to support the EI-based findings. Although EI is assumed to be reconstructive, it would be important to have other types of measurements along with the EI task to increase its face validity. Perhaps the effects of instruction could have been more apparent with other types of measurements such as reading and writing. Moreover, students may become more motivated to take the test if other measurements of L2 knowledge are provided along with the task. Listening to 60 sentences can be tedious especially for beginning- and intermediate-level learners who have to deal with many test items

that they cannot process well. In terms of item selection, there may have been a bias toward instructed learners whose scores might have benefitted from their experience with textbook items.

Third, selection and grouping of participants have to be carefully constructed. Grouping of intermediate-level learners was problematic and it was particularly difficult to exclude heritage learners whose language abilities varied considerably. As evident in the results from discriminant analysis, these misclassified participants might have skewed the analysis process, resulting in higher error rates. Another problem was that some of the participants did not provide sufficient information, and there were a few participants who left some of the answer boxes blank in their demographic survey. Moreover, there were other factors that may have to be considered in the future in grouping participants (such as participants' native language, language-related majors, and the number of languages they speak, etc.). Lastly, more studies need to be conducted with a larger number of participants, especially in the category of advanced-level instructed learners.

References

- Bachman, L., & Palmer, A. (1996). *Language testing in practice*. Oxford: Oxford University Press.
- Blasdell, R., & Jensen, P. (1970). Stress and word position as determinants of imitation in first-language learners. *Journal of Speech and Hearing Research, 13*, 193-202.
- Bley-Vroman, R., & Chaudron, C. (1994). Elicited imitation as a measure of second language. In E. Tarone, Gass, S., Cohen, A. (Ed.), *Research methodology in second language acquisition* (pp. 245-261). Mahwah, NJ: Erlbaum.
- Carroll, J. B. (1967). Foreign language proficiency levels attained by language majors near graduation from college. *Foreign Language Annals, 1*, 131-151.
- Carson, J. G., & Longhini, A. (2002). Focusing on learning styles and strategies: A diary study in an immersion setting. *Language Learning, 52*, 401-438.
- Chaudron, C. (1983). Research on metalinguistic judgements: A review of theory, methods, and results. *Language Learning, 33*, 343-377.
- Collentine, J., & Freed, B. F. (2004). Learning context and its effects on second language acquisition. *Studies in Second Language Acquisition, 26*, 153-171. Retrieved from doi:10.1017/S0272263104262015
- Corder, S. P. (1967). The significance of learner's errors. *International Review of Applied Linguistics in Language Teaching, 5*, 161-170.
- DeKeyser, R. M. (1986). *From learning to acquisition? Foreign language development in a US classroom and during a semester abroad*. (Unpublished doctoral dissertation), Stanford University.

- DeKeyser, R. M. (1991). Foreign language development during a semester abroad. In B. Freed (Ed.), *Foreign language acquisition and the classroom* (pp. 104-118). Lexington, MA: D.C. Heath.
- Doughty, C. J. (2003). Instructed SLA: Constraints, compensation, and enhancement. In C. J. Doughty & M. Long (Eds.), *The handbook of second language acquisition* (pp. 313-348). Malden, MA: Blackwell.
- Ellis, R. (2005). Measuring implicit and explicit knowledge of a second language: A psychometric study. *Studies in Second Language Acquisition*, 27, 141-172.
- Erlam, R. (2006). Elicited imitation as a measure of L2 implicit knowledge: An empirical validation study. *Applied Linguistics*, 27, 464-491.
- Frank, V. (1997). *Potential negative effects of homestay*. Paper presented at the Middle Atlantic Conference of the American Association for the Advancement of Slavic Studies, Albany, NY.
- Freed, B. (1995). What makes us think that students who study abroad become fluent? In B. Freed (Ed.), *Second language acquisition in a study abroad context* (pp. 123-148). Amsterdam: John Benjamins.
- Graham, C. R. (2006). An analysis of elicited imitation as a technique for measuring oral language proficiency. *Selected Papers from the Fifteenth International Symposium on English Teaching* (pp. 57-67). Taipei, Taiwan: English Teachers' Association.
- Graham, C. R., Lonsdale, D., Kennington, C., Johnson, A., & McGhee, J. (2008). *Elicited imitation as an oral proficiency measure with ASR scoring*. Paper presented at the European Language Resources Association (ELRA), Marrakech, Morocco.
- <http://www.lrec-conf.org/proceedings/lrec2008/summaries/409.html>

- Grossman, P. F., & Scholes, R. J. (1971). The role of grammaticality and intonation in imitation of word strings by Hebrew-speaking children *Communication Science Lab Quarterly Report* (Vol. 9, pp. 21-32).
- Hameyer, K. (1980). Testing oral proficiency via elicited imitation. *Revue de Phonetique Appliquee*, 53(1), 11-24.
- Hansen, L., Graham, R. C., Brewer, J., Brewer, R., & Tieocharoen, W. (2007). *The assessment of English oral proficiency: Alternative measures*. Paper presented at the China English Language Education Association, Beijing.
- Jessop, L., Suzuki, W., & Tomita, Y. (2007). Elicited imitation in second language acquisition research. *The Canadian Modern Language Review*, 64(1), 215-238.
- Jorden, E. H., & Noda, M. (1987a). *Japanese: The spoken language, Part I*. New Haven, CT: Yale University Press.
- Jorden, E. H., & Noda, M. (1987b). *Japanese: The spoken language, Part II*. New Haven: Yale University Press.
- Jorden, E. H., & Noda, M. (1987c). *Japanese: The spoken language, Part III*. New Haven: Yale University Press.
- Kim, I. (2006). Automatic speech recognition: Reliability and pedagogical implications for teaching pronunciation. *Educational Technology & Society*, 9(1), 322-334.
- Kobayashi, N., Ford-Niwa, J., & Yamamoto, K. (1996). Nihongo nouryoku no atarashii sokuteihou [A new method for measuring Japanese proficiency] (SPOT). *Sekai no Nihongo Kyouiku [Journal of the Language Institute of the Japan Foundation]*, 6, 201-218.

- Krashen, S. (1982). *Principles and practice in second language acquisition*. Oxford, England: Pergamon.
- Krashen, S. (1985). *The input hypothesis: Issues and implications*. London, England: Longman.
- Lafford, B. (2004). The effect of the context of learning on the use of communication strategies by learners of Spanish as a second language. *Studies in Second Language Acquisition*, 26, 201-225.
- Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, 33(1), 159-174. Retrieved from doi:10.2307/2529310
- Lee, A., & Kawahara, T. (2009). *Recent Development of Open Source Speech Recognition Engine Julius*. Paper presented at the Asia-Pacific Signal and Information Processing Association Annual Summit and Conference, Sapporo, Japan.
- Long, M. H. (1988). Instructed interlanguage development. In L. Beebe (Ed.), *Issues in second language acquisition: multiple perspectives* (pp. 115-141). Rowley, MA: Newbury House.
- Lonsdale, D., Graham, C. R., & Madsen, R. (2005). Learner centered language programs: Integrating disparate resources for task-based interaction *User Centered Computer Aided Language Learning* (pp. 116-132). Hershey, PA: Information Science Publishing.
- Mackey, & Gass. (2005). *Second language research: Methodology and design*. Mahwah, NJ: Lawrence Erlbaum.
- Marriot, H. (1995). Acquisition of politeness patterns by exchange students in Japan. In B. Freed (Ed.), *Second language acquisition in a study abroad context* (pp. 197-224). Philadelphia, PA: John Benjamins.

- Matsushita, H. (2011). *The development of effective language models for an EI-based L2 speaking test: Capturing Japanese interlanguage phenomena with ASR technology*. Paper presented at the American Association for Applied Linguistics, Chicago.
- Matsushita, H., Dewey, D. P., & Lonsdale, D. (Eds.). (2010). *Japanese elicited imitation: ASR-based oral proficiency test and optimal item creation*. Tokyo: Kobe University Press.
- Matsushita, H., & LeGare, M. (2010). *Elicited imitation as a measure of Japanese L2 proficiency*. Paper presented at the Association of Teachers of Japanese, Philadelphia
- McDade, H., Simpson, M., & Lamb, D. (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.
- Möhle, D. (1984). A comparison of the second language speech of different native speakers. In H. Dechert, D. M. Möhle & M. Raupach (Eds.), *Second language productions* (pp. 26-49). Tübingen, Germany: Narr.
- Naiman, N. (1974). The use of elicited imitation in second language acquisition research. *Working Papers on Bilingualism*, 2, 1-37.
- Okura, E. (2011). *The effects of working memory on elicited imitation assessment of second language oral proficiency*. MA Thesis, Brigham Young University, Provo.
- Pavesi, M. (1986). Markedness, discoursal modes, and relative clause formation in a formal and an informal context. *Studies in Second Language Acquisition*, 81, 38-55.
- Perde, C. (1993). *Adult language acquisition: Crosslinguistics: The results* (Vol. 2). New York: Cambridge University Press.

- Regan, V. (1995). The acquisition of sociolinguistic native speech norms: Effects of a year abroad on second language learners of French. In B. Freed (Ed.), *Second language acquisition in a study abroad context* (pp. 245-268). Philadelphia, PA: John Benjamins.
- Russell, R. A. (2005). Acquisition and attrition of -wa and -ga in Japanese as a second language. In J. Cohen, K. T. McAlister, K. Rolstad & J. MacSwan (Eds.), *ISB4: Proceedings of the 4th International Symposium on Bilingualism* (pp. 2020-2036). Somerville, MA: Cascadilla Press.
- Russell, R. A. (2011, in press). Syntactic attrition in L2 Japanese missionary language. In L. Hansen (Ed.), *Second language acquisition abroad: The LDS missionary experience*.
- Sachs, J. (1967). Recognition memory for syntactic and semantic aspects of connected discourse. *Perception and Psychophysics*, 2, 437-442.
- Schachter, J. (1974). An error in error analysis. *Language Learning*, 24, 205-214.
- Schmidt, R. W. (2001). Attention. In P. Robinson (Ed.), *Cognition and second language instruction* (pp. 3-32). Cambridge, MA: Cambridge University Press.
- Tomita, Y., Suzuki, W., & Jessop, L. (2009). Elicited Imitation: Toward valid procedures to measure implicit second language grammatical knowledge. *TESOL Quarterly*, 43(1), 345-350.
- Watabe, M., & Gilbert, K. S. (1993a). *Toward better Japanese, Part I*. Provo, UT: Bunkyoasha.
- Watabe, M., & Gilbert, K. S. (1993b). *Toward better Japanese, Part II*. Provo, UT: Bunkyoasha.
- Wilkinson, S. (1998). Study abroad from the participants' perspective: A challenge to common beliefs. *Foreign Language Annals*, 31, 23-39.
- Yoshioka, K. (1991). Elicited Imitation tests for oral production assessment. *Working papers*, 2, 127-140. Retrieved from <http://ci.nii.ac.jp/>

Appendix I: Consent Form

INFORMED CONSENT

Japanese Test Development Research Project

Introduction

This research is being conducted by professor Dr. Dan Dewey and Hitokazu Matsushita of Brigham Young University for the purpose of developing better methods of assessing Japanese language proficiency.

Procedures

As a participant, you will be asked to complete two tests of Japanese: the Japanese Elicited Imitation and the Japanese Simulated Oral Proficiency (SOPI) Test. The first test involves hearing a series of sentences and repeating them aloud. The second test involves reading the task instructions and talking about the topics in Japanese according to the instructions. Each test takes about 15 minutes, so altogether these will take approximately 30 minutes. Also, we would like to have access to your test scores from any of the following you may take this semester: the Japanese Diagnostic Test, the Japanese Bypass Exam, and the Oral Proficiency Test (OPI). The OPI is given only to select students for program evaluation purposes. The other two exams (Bypass and Diagnostic) are given for course credit (Bypass) and program/student evaluation purposes (Diagnostic). Your instructors will inform you if you will be taking these exams. For this research project, you are simply giving us permission to receive your test scores from those instructors if you complete those tests.

Risks and Discomforts

There are no known risks associated with participating in this research project. Some students feel a bit anxious when taking the exams. Your participation in this research project is voluntary and you are free to withdraw from participating in this project at any time.

Benefits

No direct benefits to participants are associated with this project. Your participation will help us produce assessments that will help us to better measure Japanese language proficiency, benefitting both BYU and the Japanese language education community. You will receive \$10.00 in cash as a thank-you for your participation once you complete the Japanese Elicited Imitation and Japanese SOPI tests.

Confidentiality

Your performance in this research project will be kept completely confidential, and your name will never be mentioned or used in any publications or presentations of the results of this project. Your scores will be kept in a database that includes only an ID number (not your name).

Questions about the Research Study

You may contact the researchers, Dr. Dan Dewey (ddewey@byu.edu, 801-422-6005) and Hitokazu Matsushita (hitokazu@byu.edu) with any questions.

Questions about your Rights as Research Participants

You may also contact the BYU IRB Administrator, A-285 ASB, Brigham Young University, Provo, UT 84602, 801-422-1461, irb@byu.edu if you have any questions or concerns.

I have read and understand the above, and I voluntarily agree to participate in this research project. I understand that I may keep a copy of this form.

Signature of Participant

Date

I agree to allow the researchers to collect scores from tests administered by the Japanese Section of Asian and Near Eastern Languages, as described above.

_____ Yes

_____ No [Please initial]

Appendix II: Translations of Discriminant EI sentences (Romanized)

	items	English Translations
1	明日行きませんか？ Asita ikimassen ka?	‘Won’t (you) go tomorrow?’
2	これは手紙じゃありません。 Kore wa tegami zya arimassen.	‘This is not a letter.’
12	児玉と申しますが、またあとでご連絡いたします。 Kodama to moosimasu ga, mata ato de gorenraku itasimasu.	‘My name is Kodama (polite), could (you) please call (me) back later?’
15	明日はいい天気になるね。 Asita wa ii tenki ni naru ne.	‘Tomorrow will be good weather.’
32	この翻訳は全部あなたがさせられたんだそうですね。 Kono honyaku wa zenbu anata ga saserareta n da soo desu ne.	‘(You) were made to translate all of this, right?’
33	調べたいことがあるときはどうしたらいい？ Sirabetai koto ga aru toki wa doo sitara ii?	‘When (I) have something I want to research, what should I do?’
34	難しく考えることはありません。 Muzukasiku kangaeru koto wa arimassen.	‘(You) shouldn’t think too hard.’
35	実は仕事をやめさせられちゃったんです。 Zitu wa sigoto o yamesaseraretyatta n desu.	‘Actually, (I) was fired from my work.’
36	冷めないうちに召し上がってください。 Samenai uti ni mesiatte kudasai.	‘Please eat before it gets cold.’
42	これを見るとおもしろいことに気がつくはずですよ。 Kore o miru to omoshiroi koto ni ki ga tuku hazu desu.	‘If (you) see this, (you) will notice something interesting.’
45	いつの間にかよく話しかけられるようになりました。 Itu no ma ni ka yoku hanasikakerareru yoo ni narimasita.	‘Before (I) realized (it), people started to talk to (me) more frequently.’
47	お金にめぐまれていたら幸せに決まっているでしょう。 Okane ni megumarete itara siawase ni kimatte iru desyoo.	‘If (you) are blessed with money, (you) must be happy.’
55	日本がどんな国なのか知りたがっている人がたくさんいます。 Nihon ga donna kuni na no ka siritagatteiru hito ga takusan imasu.	‘There are many people who want to know what kind of country Japan is.’
56	学生にとって十分に寝る時間を取るのは大変でしょう。 Gakusee ni totte zyuubun ni neru zikan o toru no wa taihen desyoo.	‘For students, it is probably difficult to have enough time to sleep.’
59	知られてはいるんですが、見た人はあまりいないようです。 Sirarete wa iru n desu ga, mita hito wa amari inai yoo desu.	‘It is well known, but not too many people have seen it.’