2011-06-10

A Study of the Correlation Between Working Memory and Second Language EI Test Scores

Eve Kiyomi Okura
Brigham Young University - Provo

A Study of the Correlation Between Working Memory and Second Language EI Test Scores

Eve K. Okura

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

Master of Arts

Deryle Lonsdale, Chair
C. Ray Graham
Dan Dewey

Department of Linguistics and English Language

Brigham Young University

June 2011

ABSTRACT

A Study of the Correlation Between Working Memory
and Second Language EI Test Scores

Eve K. Okura
Department of Linguistics and English Language, BYU
Master of Arts

A principal argument against the use of elicited imitation (EI) to measure L2 oral proficiency is that performance does not require linguistic knowledge, but requires only rote memorization.

This study addressed the issue by administering two tests to the same group of students studying English as a second language: (1) a working memory test, and (2) an English oral proficiency EI test. Participants came from a range of English language proficiency levels.

A Pearson correlation was performed on the test results for each participant. The hypothesis was that English EI scores and working memory scores would not correlate significantly. This would suggest that the two tests do differ in what they measure, and that the English EI test does measure knowledge of the language to some degree.

The results of the Pearson correlation revealed that there was a small positive correlation between working memory and English EI scores, but that it was not significant. There was also a significantly positive correlation between students' English EI scores and ELC level. These findings suggest that the English EI test fundamentally functions as a language test, and not significantly as a working memory test.

Keywords: working memory, elicited imitation, short-term memory, language testing, linguistic knowledge, explicit linguistic knowledge, implicit linguistic knowledge

ACKNOWLEDGEMENTS

I owe much to the following individuals, without whom this thesis could not have been completed: all of the members of the PSST research group, including Ben Millard, Nate Glenn, Dan Allongo, and in particular Jerry McGhee and Matt LeGare for solving all of my computer problems, figuring out the technical/programming side of the research, and for patiently walking me through the process, and to the graders for getting all of the data graded in miraculous time. I would like to express my appreciation to the Center for Language Studies for making this possible by providing our research group with funding. I am also indebted to Xanthe Farnworth for generously giving of her time and expertise to teach me life-changing lessons on technical writing, and to Dr. Wendy Baker for consultation on the working memory test. Much appreciation to Troy Cox and Brigham Young University's English Language Center for allowing me to carry out this research and for sharing their computer lab for testing, and to the students at the ELC for giving of their time to participate in this study.

I would like to express gratitude to many friends who have been supportive: Zoram Quintana, Catherine Curtis, and Andrea Hardeman for their encouragement; Dr. Cynthia Hallen and Dr. Janis Nuckolls for advising me during the arduous process of determining a thesis topic, and who, in addition to being mentors, have been friends.

I am grateful to my parents, Sanford and Nancy Okura, for their constant support. And to Uncle, who taught me the reasons for getting my education. These continue to be the source of my drive and motivation in school. There are many others, too many to name here, who have helped me along in this project in countless ways.

My thesis committee has been phenomenal. Dr. Dan Dewey helped make this possible by his willingness to work around the shifting timeline, despite travels and other obligations. Without Dr. Ray Graham, this thesis would not exist. His clarifying insights and kind mentorship have guided the framework of this study along from beginning to end. I could not have finished as I did without my graduate committee chair, Dr. Deryle Lonsdale. I have been continually amazed and impressed by his willingness to make time for students. I am deeply grateful for his confidence in me, his invaluable feedback, and his ability to creatively improvise mentoring approaches to meet students' needs. The opportunity to learn from and work with him has been an honor, a privilege, and a pleasure.

I would also like to express appreciation to BYU's Department of Linguistics faculty, who have been examples of brilliance coupled with kindness—many of whom, perhaps unknowingly, have been role models to me of not only academic success, but of the kind of person I would like to be. While extremely capable and engaged in their own research and professional goals, they somehow still manage to be student-oriented and sincerely concerned. I have felt at home in the department, and am grateful to have had the opportunity to learn in such an environment.

Table of Contents

List of Tables

List of Figures

1. Introduction

One of the great questions in linguistics, and language acquisition in particular, is how to measure an individual's linguistic ability. There are countless methods of assessing language competence, ranging from written grammar examinations, to live interviews with a native speaker. Many disagree over the efficacy of various evaluation methods. Some of the discussion revolves around what different types of tests actually measure, and whether or not they are actually assessing linguistic knowledge and ability, or are measuring some other capability. One such testing method is commonly referred to as "elicited imitation" (EI).

EI is a testing method in which linguists/scientists present a series of sentences to a person. The person is asked to imitate each sentence by repeating it as accurately as possible. Researchers record and analyze these imitated responses. Since 1967 EI has been seen as useful in linguistic research, and language acquisition research specifically (Slobin & Welsh 1967). Many have proposed that elicited imitation tests can assist in measuring language ability (Slobin & Welsh 1967; Connell & Myles-Zitzer 1982; Vinther 2002; Hatfield et al. 2007; Erlam 2009; and Jessop et al. 2009). Many others have argued that elicited imitation cannot be used to measure language ability, and that listeners can simply repeat what they hear, without the occurrence of any linguistic processing (Smith 1973; Hamayan et al. 1977). Recently, scholars have readdressed the central problem regarding EI:

> In the 1970s, EI's validity was challenged: the major criticism being the possibility of rote repetition in response to stimuli (i.e., participants may be simply parroting what they hear). (Jessop et al. 2007:216)

Such arguments claim that a person could repeat the utterance without understanding what it means, and without having any ability in that language. Addressing the issue of whether or not

EI test-takers rely solely on “rote repetition” in responses requires a preliminary discussion of memory.

Memory is said to have three main categories: long-term, short-term, and working memory (Cowan 1996). As the name implies, long-term memory stores information for greater lengths of time, somewhat “permanently.” This paper will only touch on long-term memory tangentially, and will not discuss it in-depth. Short-term memory (STM) draws on information stored in long-term memory. The two differ in that information in STM is at a higher level of activation than the rest of the information in long-term memory. This thesis will focus on the third component of memory, working memory (WM). WM contains the information held momentarily, as it is needed to analyze, solve a problem, or perform a task.

Working memory is the part of memory that would determine whether high performance levels on EI tests are a result of mere “rote repetition,” or whether they do require linguistic knowledge and performance competency. An EI test-taker who did not know English would have to hear the stimulus utterance, retain it in WM, and then repeat the utterance exactly to get a perfect score on the item. If the item were beyond their WM capacity, they would be expected to get only part of it correct at best, all of it wrong at worst.

I do not claim that working memory and L2 ability are completely, 100% independent from each other. Some working memory capacity is necessary to even be able to respond in an EI test, just as some WM capacity is necessary in all analytical and linguistic tasks, including spontaneous conversation. However, I show that these two EI tests (the English EI and the WM) are measuring different abilities to some degree. If all EI language tests consisted of mere parroting, then an individual’s scores on the EI WM test and the EI language test should correlate exactly. There should be no variation or discrepancy between the two, because they would essentially be the same type of test, measuring the same ability within the same individual.

In this study ESL students (who are studying English as a second language) take an EI test to rate their English language ability. Each student then takes a working memory test to assess their working memory capacity. Correlations between the two are analyzed.

If there is significant correlation, then it may be that working memory is interfering with elicited imitation tests that are intended to measure language ability. Such a result would support arguments that elicited imitation tests are mere “parroting,” and therefore offer no contribution to measuring language ability. If, however, there is minimal correlation between working memory scores and EI language test scores, then the research will support the idea that EI language tests are not primarily memory tests, thus furthering the argument for the efficacy of elicited imitation as a valid complement to other methods of assessing second language capability. I claim that the latter is true, and that there will be minimal correlation between WM scores and EI scores, showing that what they are measuring is different (i.e. this EI language test is not just a working memory test). It is predicted that some students who have the same level of working memory but differing language ability will score differently on the EI language test. This thesis hypothesizes that EI language tests, when constructed properly[1], primarily measure linguistic ability, and demonstrably do not measure an individual’s working memory to a significant degree.

The next chapter defines the concept of working memory, and summarizes its history. It also reviews the history of elicited imitation use in research, and in particular its relevance to second language acquisition research.

---

[1] What is meant by “properly” here is that EI test items are constructed: (1) to test for knowledge of grammatical features, including third person singular conjugation, plural, tense, etc.; and (2) to specific lengths, in terms of syllables (Christensen et al. 2010).

2. Review of Literature

Límites

Si para todo hay término y hay tasa
y última vez y nunca más y olvido...

Hay, entre todas tus memorias,
una que se ha perdido irreparablemente…

No volverá tu voz a lo que el persa
dijo en su lengua de aves y de rosas,
cuando al ocaso, ante la luz dispersa,
quieras decir inolvidables cosas…

- *Jorge Luis Borges*

Limits

If there is a limit to all things and a measure
And a last time and nothing more and
forgetfulness…

There is among all your memories one
Which has now been lost beyond recall…

You will never recapture what the Persian
Said in his language woven with birds and roses,
When, in the sunset, before the light disperses,
You wish to give words to unforgettable things...

2.1. A Limit to All Things

In his poem, *Límites* Borges observes that, "there is a limit to all things." He alludes to forgetfulness and memories "lost beyond recall." What are the limits of the human mind to remember, retain, and recall information? What are the limits of the scope of the mind in a person's first language, an individual's second language, or an altogether unfamiliar foreign language? He speaks of the "Persian," saying, "You will never recapture" what he said "in his language woven with birds and roses." An unknown language might as well be nonce syllables to the listener attempting to recall specific syllables with accuracy, particularly those that compose lengthy utterances. Borges remarks on the virtual impossibility of memorizing a lengthy utterance in a foreign language one has not studied. Like the Persian's language of "birds and roses," repeating sentences from an unfamiliar foreign language would be similar to repeating nonce syllables.

This chapter discusses the definitions of working memory and elicited imitation, a testing method involving repetition of foreign-language sentences. It also surveys the history of research in these areas, and how the two interact. In particular, it focuses on the average person's working memory capacity and the interaction between working memory and elicited imitation.

2.2. Working Memory

Working memory has been called by several names over the course of various studies. "Working memory" has been referred to as "immediate memory span" by Miller (1956), and "short-term memory" (STM) by a few others. Cowan defines the difference between short-term memory and working memory:

> Short-term memory refers to a vivid form of memory that lasts only for a few seconds after one receives a stimulus or thinks of something, and working memory usually refers to short-term memory when it is used to solve a problem or perform a task. (Cowan 1996:2)

Thus working memory is a subcategory of short-term memory, which a person actively applies and uses to do something in a given moment. Any type of problem solving requires maintenance of basic assumptions held in the mind while thinking occurs. The requirement is the same for spontaneous speech. In order for a conversation to occur, both interlocutors must be able to process what is being said by the other, and then maintain that new information in memory while forming a response to ensure that the response is grammatical, relevant, and logical. Short-term memory (STM) is what allows that information to be retained in memory while the response is being crafted. This is what occurs in natural speech. Oral proficiency interviews (OPI) are a generally accepted method of testing L2 abilities (Lehman & Tompkins 1998). In an OPI, a trained rater engages in conversation with the test-taker in the test-taker's L2. One of the benefits of an OPI is that the student is engaged in spontaneous language production. This limits the amount of time they have to think about a response. It also focuses on meaning and being communicative, rather than focusing on form. These two elements—time limit and avoiding meta-awareness of form—are thought to shift the test towards implicit language knowledge instead of explicit (Ellis 2005). However, OPIs require a native speaker to engage in conversation, and several trained graders, making them more expensive to proctor and score. The unpredictable nature of the conversation and differing vocabulary and grammar used in conversations with different test-takers also makes OPIs more difficult to grade consistently. Another problem is that raters have very limited control over what forms the test-taker uses. The test-taker may or may not produce certain grammatical features raters are looking for (Ellis 2005). The current study differs from natural speech in that spontaneous speech does not occur, as it would in a traditional OPI. In an elicited imitation (EI) test, as opposed to natural conversation, there is a response, but it is imitative rather than a spontaneous utterance invented by the individual. Applying Cowan's definition (1996) to this current study, STM is what holds the stimulus information when the response is being planned.

Krashen (1982) distinguishes between "acquisition" and "learning." Acquisition is done intuitively, similar to a child learning a native language.

> The result of "language acquisition," acquired competence, is also subconscious. We are generally not consciously aware of the rules of the languages we have acquired. Instead, we have a "feel" for correctness. Grammatical sentences "sound" right, or "feel" right, and errors feel wrong, even if we do not consciously know what rule was violated.

Krashen associates language acquisition with "implicit learning." It is intuitive. He defines "language learning" as a conscious awareness of grammar, vocabulary, syntax, etc., and refers to this as "explicit learning" (Krashen 1982:10).

Ellis echoes Krashen's use of the terms implicit and explicit. Whereas Krashen applies these terms to the learning process, Ellis extends it to the product of implicit and explicit linguistic knowledge (Ellis 2008:5-6).

> …explicit and implicit knowledge of language are distinct and dissociated, they involve different types of representation, they are substantiated in separate parts of the brain, and yet they can come into mutual influence in processing. (Ellis 2008:7)

Explicit knowledge of a language is form-focused rather than meaning-focused, and, as opposed to implicit linguistic knowledge, involves conscious thought. While the concepts of implicit and explicit linguistic knowledge do surface in the literature, the question of which type or what proportions of linguistic knowledge are accessed in EI tests is outside the scope of this thesis.

Some studies have shown that STM typically lasts about 20 to 30 seconds, but that this may have been an overestimate, due to interference of the articulatory loop (Cowan 1996). Differences in length of short-term memory may be due to experiments that test it with interference as opposed to those that test it without. This study tests working memory without interference.

In previous working memory studies, phonologically similar words (words that start with similar phonemes) were more difficult for people to remember than a list of words that start with dissimilar phonemes. This is likely due to the element of comparison—remembering the words relative to each other, and not remembering the words independently, out of context from the other items in the list (relative vs. absolute memory).

Baddeley and Hitch (1986) experimented to see if storing several numbers in working memory affected the ability to carry out comprehension tasks. It did not. This resulted in the idea that memory is stored in a different "place" from where processing and problem-solving occur, as one did not impede the other.

How does this mechanism work if STM appears to allow for simultaneous listening and comprehension (listening comprehension), and yet STM cannot store more than a number of syllables unless there is comprehension? Comprehension involves combining sounds into packets of meaning (morphemes, lexemes, lexical items, phrases, sentences), that allow STM to group items together, thereby expanding what can be held in STM accordingly. I will discuss this in further detail in the following sections.

2.2.1. Working Memory Limits

Through a number of experiments, the original range of working memory capacity was thought to be 7 ± 2 (Miller 1956), meaning that a person could generally hold up to seven completely unrelated items in mind simultaneously (plus or minus two). Other scholars have suggested a similar "immediate memory span" range as Miller. His 1956 article was a very influential publication in terms of shaping public perceptions of memory capacity. For the past 50 years, the generally accepted view is that a person can remember up to seven different items. Interpretations of Miller's research have developed a myriad of variations. The common perception that the average person can remember up to seven different numbers at once is most likely a derivative of his study.

Interestingly, although Miller concludes that the "magic number" for immediate memory span is 7 ± 2, he references a study by Eriksen and Hake (1955). In their study, researchers placed vibrations of varying intensities on different areas of subjects' chests. Miller notes that "a good observer" was able to discern between four levels of intensity (1956:86). This finding supports Cowan's later research that asserts that the magic number is not seven, but rather four.

In the past forty years, several scholars have advocated the number 4 ± 1 to replace 7 ± 2 as the "magical number" defining the scope of working memory. Studies by Henderson, Sanders, Posner, Scarborough, and Sperling, all agree a more accurate number for average working memory capacity is 4 ± 1 (Cowan 2001). Cowan's (2001) article on short-term memory capacity reiterates the conclusion that the specific number of independent "pieces" of information an average person can retain at once is 4 ± 1.

Miller also introduced the process of "recoding" (Miller 1956) a concept which scholarship years later referred to as "chunking" (Cowan 2001). Recoding is taking multiple separate items and creating patterns that result in fewer separate items to remember. He used an example of a telegrapher using "dits" and "dahs" to send messages. It would be difficult for most people to be able to memorize a long sequence of dits and dahs. However, the telegrapher becomes accustomed to certain frequently used sequences. The individual dits and dahs "recode," or "chunk" together in his mind, so that rather than having to memorize individual dits and dahs, he can remember one set, and another. This speeds up the process of communicating, and it enhances his capacity to remember more. Miller states that recoding is "implicit in nearly all such mnemonic devices," and that it is "an extremely powerful weapon for increasing the amount of information that we can deal with" (1956:95). This will become relevant in the later discussion of the interaction between working memory and language.

2.2.2. Working Memory and Language

Doughty and Long, along with others have associated language-learning aptitudes with memory capacity (Doughty & Long 2003:67). Robinson (2005) noted that phonological working memory capacity in particular is associated with L2 speaking abilities, as it allows "utterances to be maintained in WM long enough for the analytic cognitive comparison…to be made" (2005:51). Likewise, DeKeyser notes the difference between "meaningful" and "nonmeaningful" associations, each in its relation to memory capacity.

> …meaningful form-function mappings required associating an element of the noun phrase with an element of the verb phrase, each element taking a very different concrete form. Associating nonmeaningful co-occurrence of concrete elements logically draws more on memory, whereas establishing meaningful relationships between abstract entities draws more on insight. (DeKeyser 2005:16)

On this basis, one would expect a nonce syllable test would measure memory, and a “meaningful” (i.e. language-based) test would draw less on memory than the nonce, and more on linguistic ability.

Recoding is crucial in linguistic communication, extending the amount of information a person can retain in their working memory by chunking separate phonological components into lexical items, and the morphosyntactic system of the language. This recoding into clusters of information enables interlocutors to both recall and speak many more syllables than 4 ± 1, within the limitations of their working memory capacity. Chunks of information are not separate syllables, but are words, phrases, sentences, and ideas.

> Utterances are constructed as intonation units, and substantive units are fairly strongly constrained to have a modal length of four words in English, a fact indicative of the cognitive limits on how much information can be fully active in the mind at any one time. (Ellis 2002:156)

Previous research has also shown that knowledge of semantics, grammar, syntax, etc., affects verbal short-term memory performance (Morra 2000). This suggests that such linguistic working memory chunking in actuality does occur during the process of verbal repetition.

In addition to recoding, another mechanism of interaction between working memory and language is the articulatory loop. The articulatory loop is the method of either vocal or sub-vocal repetition of a phrase to maintain it in short-term memory (Morra 2000). A brief animation clip on Sesame Street demonstrates a classic example of the articulatory loop. A mother sends her daughter to the grocery store to get a specific list of items. All the way to the store, the girl sings, "A loaf of bread, a container of milk, and a stick of butter…a loaf of bread, a container of milk, and a stick of butter…" over and over to remember what she was told to buy. This is use of the articulatory loop. According to some studies, the items retained in short-term memory using the articulatory loop are limited to that which can be spoken in two seconds. Therefore, the assumption would be that the faster one could speak, the more one could retain sentences in short-term or working memory. However, some studies have shown that the articulatory loop cannot completely account for verbal STM.

> Moreover, verbal STM span is affected by variables that have little or no effect on articulation rate, such as semantic variables (e.g., Poirier & Saint-Aubin, 1995), grammar class (Tehan & Humphreys, 1998), word frequency or familiarity (e.g., Henry & Millar, 1991; Hulme et al., 1997), and order of the stimulus words (Brook & Watkins, 1990). (Morra 2000:192)

In such cases, if verbal short-term memory is not affected by the articulatory loop, but rather by grammar, word frequency, and sequence, then this seems to suggest that linguistic chunking seems to supersede chunking of the articulatory loop type.

2.2.3. Working Memory in This Study

The basic concepts of working memory (including its limits, how it functions with regard to recoding, and the relation of working memory to language) are key components of the primary assumptions upon which this study is founded. Other research regarding efficacy of various types of working memory assessments helped us to decide which type of working memory test to use in this study.

Our study is based on the premise that due to working memory limitations (e.g. 4 ± 1), correctly repeating longer utterances in an EI test is only possible via linguistic recoding. It follows that working memory of nonce syllables will be different from working memory enhanced by language ability, and that the difference indicates how the EI tests are measuring language. Martin and Ayala (2004:478) researched auditory-verbal short-term memory of individuals with aphasia. They performed both repetition and matching span tasks. One suggestion they gave for future research involving measuring STM was to do non-word span tasks. Their article proposes that if non-word recall is lower than word recall, it would mean that "activation of higher level information" (i.e. linguistic processing) is involved in the process. In such case, an activity such as matching or EI would not just be measuring sound-related short-term memory (i.e. acoustic, phonological, phonetic properties.). If language EI tests were mere memory tests, and nothing else, then an individual's scores on a nonce-syllable working memory test and a second language EI test would both be measuring the same capacity—memory capacity. If such were the case, one would expect results from the two tests to have a perfect—near perfect—correlation of 1, because they would both be testing the individual's working memory, rather than one testing memory and the other testing linguistic ability.

According to studies performed by Lehman and Tompkins (1998), backwards digit recall has little to no correlation to working memory. In backwards digit recall tests, a person is given several numbers in sequence, and is asked to repeat the digits back, but in reverse order. One possible reason backwards digit recall did not have a strong correlation to working memory is perhaps because the sequence was broken. Also, it is very easy for a person to "chunk" with digits. For example, a sequence of seven, six, one, four, eight, can be recoded as 76,148, so it can become a lot easier to remember several digits. In such a test some participants may chunk all the digits, while others may chunk some digits and not others. Still other individuals may not chunk at all. It would be difficult to determine whether or not chunking has occurred, and so the number of digits recalled may not be an accurate measurement of independent pieces of information being held in the mind simultaneously. Through their studies, Lehman and Tompkins also concluded that a simple word recall test also had little correlation to working memory. Due to these views that other attempted memory tests were less effective, we chose to use a nonce syllable working memory test for this study.

2.3 Elicited Imitation

This section will explain what elicited imitation is, how it has been used in language acquisition, including second language, research, and how it relates to this particular study. Elicited imitation (EI) and sentence repetition testing (SRT) are the most common terms used to refer to the same testing method. EI is the "…repetition of a model sentence presented in a context calling for imitation, as opposed to…spontaneous imitation of…utterances" (Slobin & Welsh 1967:2). Bley-Vroman and Chaudron (1994) suggest that EI performance requires both linguistic processing and short-term memory. They outline what they term a “Speech comprehension system,” involved in EI performance, which consists of four steps: (1) the person hears the utterance; (2) the person forms their version of it in their mind (a “representation”); (3) the person stores this “representation” in short-term memory; (4) the person then verbalizes their representation of the initial utterance.

EI can assist in measuring first and second language acquisition and ability. Slobin and Welsh (1967) focused on eliciting imitation of adult speech from a child. They concluded that both recognition and precise imitation of an utterance require linguistic processing, stating that, “We believe that elicited imitation is a useful probe for revealing linguistic competence" and continue on to say that it may be similarly beneficial to study adults imitating elicited utterances.

> …[sentence imitation by adults] suggests a similar model in adult sentence recognition: retrieval of a syntactic structure, lexical items appropriately marked as to syntactic and semantic function, and an attempt to fill in the syntactic structure with whatever of the lexical items from the model sentence are available in short-term memory. (Slobin & Welsh 1967:10)

In general, the child could not perform EI repetitions beyond what she could produce spontaneously. The child studied could "perfectly imitate ungrammatical or anomalous sentences if they [were] short enough for her to hold an auditory image in short-term memory." As this holds true regarding first language acquisition (L1), it would be logical for this principle to also hold true in terms of second language acquisition (L2). Slobin and Welsh note this by observing that length and complexity of a sentence can prevent an adult from using short-term memory to retain and imitate an utterance.

Connell and Myles-Zitzer (1982) compared EI and spontaneous speech of children. They found that although it was not always an accurate predictor of spontaneous speech, it could be reliable in certain cases. They also found that in both spontaneous speech and EI, children uttered major lexical items (nouns, verbs, and adjectives) more accurately than grammatical features.

Radloff uses the term sentence repetition testing (SRT) to refer to EI. She discusses two types of research goals in testing language capability in language acquisition research. The first would be in-depth research on a few select individuals. For this, she prescribes a diagnostic test involving "an in-depth oral proficiency interview" (Radloff 1992:21). The second type of study would aim to cover breadth rather than depth. This allows for a wider sampling of a community, with less detailed information about each person's particular strengths and weaknesses in the language, but would give a better picture of the varying abilities within the community. As their objective was to determine varying language abilities within the community, they chose to use the second test, which was more suited to their needs.

Ellis (2008:236) discussed the role of perception in language learning. He showed that people do not learn grammatical features as well if the features are less phonologically conspicuous in an utterance. He gave the example of the sentence, "*Yesterday* I walk*ed*," in which language learners will often drop the–*ed* past tense suffix. This happens often in English EI tests, suggesting that results of the EI tests are parallel to other studies in language learning. It seems that if the student has not already acquired understanding of that particular grammar principle, they are less likely to successfully process it (especially those that are less salient).

Studies have shown that EI tests do have some limitations, though they cannot always stand alone as a sole measurement of language ability, they do serve as a useful complement to other measures. SRT, another term for EI, and the second language oral proficiency evaluation (SLOPE), based on the oral proficiency interview (OPI) were both administered to bilingual communities in Cameroon. The consensus of the researchers involved in the study was that there is no single test, including SLOPE that will give second language proficiency data that can be used as the only deciding factor in language development needs assessment. The results from an SRT, however, can provide important information on the L2 abilities of members of a speech community (Hatfield et al. 2007:3). The SRT proved to be highly successful at making quick assessments with relatively minimal skill required for administration and with no requirement for the subjects to be literate.

While Jessop, Suzuki, and Tomita are proponents of using EI in SLA research, they delineate the "effect of short-term memory capacity" as one of the challenges in using EI results (Jessop et al. 2007:216). They have noted that test-takers often spontaneously correct grammatical errors from the stimulus in their responses. This stands as strong evidence for an immediate response to an EI item being an indicator of language ability, and that both the hearing and the responding does go through a process of linguistic analysis. Researchers have used EI to gather information about L2 acquisition. They describe EI as being a “powerful research tool,” that researchers can use to gather data with “validity and reliability."

Relevant to this study is the idea that high EI scores assume comprehension of the material. In distinguishing between comprehension and productive ability, Vinther points out that some of the adults tested in the study could have understood the stimulus, but were not able to produce the correct response accurately. On the reverse side of that, he stated that “the opposite situation, that subjects should be able to produce well-formed imitations of sentences they have not understood and are not able to remember, is highly improbable” (Vinther 2002:63).

The literature also provided us with guidelines for our tests. Jessop et al. (2009:347) gave seven points to take into account when using EI for L2 acquisition research. The seventh point they mention is to give clear instructions: "Repeat the sentence exactly as you hear it," and "Repeat the sentence in correct English," are two distinctly different sets of instructions, and could result in a different set of responses. In both our English EI tests and working memory EI tests the instructions were to repeat the utterance (or nonce syllable) exactly as they heard it.

Jessop et al. outlined the advantages and challenges of using EI as an L2 acquisition assessment. They also provided suggestions for using EI effectively. This thesis addresses the central problem they delineate: “In the 1970s, EI's validity was challenged: the major criticism being the possibility of rote repetition in response to stimuli (i.e., participants may be simply parroting what they hear)” (Jessop et al. 2007:216). They assert that, "Clearly, more EI studies and replication studies are necessary for the validation of EI's use as a measurement of L2 performance" (Jessop et al. 2007:217). This research has been carried out in response to these and similar statements.

2.4 The Interaction Between Working Memory and Elicited Imitation

Already we have seen several points at which working memory and elicited imitation intertwine. The limits of working memory, and the enhancement of short-term memory and imitation by linguistic chunking, is evident in language EI tests.

It is hypothesized that an L2 EI test would function more similarly to a working memory test for individuals who do not yet have any ability in the language. This is because without a grasp on the grammar, and without prior knowledge of vocabulary, the individual will not have the ability to linguistically "chunk" the sounds. The syllables uttered would then sound like nonce syllables to the student, rather than coherent words and phrases. One study demonstrated that EI can test language knowledge and is not mere rote repetition (working memory) in that difficulty is not based solely on sentence length (number of words). "Elicited imitation goes beyond rote memory and repetition; rather, sentences are assumed to be 'filtered' through one's grammatical system" (Gass & Mackey 2007:27). They give an example of two sentences that are of equal length in terms of number of words, but of differing grammatical complexity. They state that the sentence of greater grammatical complexity would be more difficult for a language learner to repeat.

The EI test used in this study measures verbal STM performance in the sense that it measures knowledge of these linguistic elements. In other words, these EI tests measure language ability, not just rote memory ability. For those who argue that EI tests are only memory tests, then according to the findings from Morra's article, they are at least memory tests influenced by linguistic ability. If influenced by linguistic ability, then, in a sense language EI tests are at least memory tests that also measure knowledge of the language to some degree. It is apparent that working memory is essential to perform well on a language EI test. However, it is also apparent that working memory is essential to perform spontaneous speech acts. One could not engage in conversation, even in one's first language, without active working memory. The question is how much do language EI tests overlap with working memory tests in what they measure? If the two do not correlate perfectly, then whatever does not overlap implies that the remaining portion, the part of the EI test that does not overlap with WM, is measuring something else. As suggested by previous research, that "something else" is knowledge of grammar, vocabulary, syntax, and other linguistic features.

3. Methodology

3.1. Overview

The English Language Center (ELC) at Brigham Young University received 94 new incoming students in January of 2011. All 94 students, from various ELC levels, took the English EI test in the first week of January as a part of an initial assessment at the start of their program. The 94 students were divided into two groups, which I will hereafter refer to as Group A and Group B. The 47 students in Group A were given a computer-administered 60-item English EI test, with half of the 60 items appearing with comprehension tasks (totaling 30), the other half without. Group B (also consisting of 47 students) was given the same 60-item English test, but with the other half of the 60 items with comprehension tasks. The microphones on five of the 94 students' computers did not record their responses, so those five tests were discarded. This left 89 students whose data we could use. Both groups had the section with comprehension tasks before the section without the tasks. The comprehension task involved two illustrations appearing side-by-side on the screen after an EI stimulus sentence was given, but before the test-taker recorded a response. One of the illustrations depicted the meaning of the stimulus sentence that the student had just listened to. The other picture was irrelevant to the stimulus. Test-takers were to click on the picture that was relevant to the meaning of the stimulus sentence. Only then were they allowed to record their audio response. These tasks were intended to divert attention away from form, and shifting the focus to meaning. This would make a test potentially an assessment of implicit linguistic knowledge as opposed to explicit, according to Ellis (2005). The two necessary conditions for an assessment to test implicit language testing are that it be (1) time-constrained, and (2) meaning-focused. However, this particular study is not designed to determine how much of the EI test is accessing implicit knowledge, and how much is accessing explicit knowledge. The comprehension tasks were designed by Dr. Ray Graham of the PSST research group, for future studies. This thesis is concerned primarily with the correlation between the memory test and the EI test scores in general, without an interest in items with the tasks in comparison to items without.

A few months later, in March and April, as many of the original students as could be found were invited to take a 10-item working memory test. This resulted in 67 students, due to some students dropping out of the ELC program in those two months, and others who were unavailable to take the test because of work schedules and other reasons. The first students to take the test took it on a web-based interface. The web version suffered from several technical difficulties, so the remainder of the students took a standalone version temporarily downloaded onto the computers at the ELC computer lab. Due to the technical difficulties of missing audio recordings in the web-based version, we were left with complete data for 44 students.

3.2. Test Creation

Two EI tests were involved in this study: An English EI test and a working memory test. For the first, a pre-existing elicited imitation test was used to measure the students' English language ability. This test was developed by BYU's PSST research group.[2] The EI test used for the language assessment portion of this study, Form D, was in existence previously, and had been in use by the PSST group in previous studies for some time. By measuring working memory and English language on separate EI tests, we will demonstrate a lack of correlation between scores. This in turn will reveal the efficacy of EI tests as an L2 assessment. Jessop et

[2] Pedagogical Software and Speech Technology, http://psst.byu.edu

al. suggested combining EI tests with other types of tasks, including comprehension tasks. In our study, we included comprehension tasks in half of the set of English EI questions (30 out of 60).

While we applied relevant recommendations in the literature, we also took measures to be at least as cautious in our research as previous studies have been, and in some cases even more. For example, the EI test in Radloff's study (1992), which she refers to as SRT, or sentence repetition test, consisted of only 15 sentences. The EI test in this study consisted of 60. In Radloff's study, she constructed the SRT in such a way as to enable non-native speakers of the language to proctor the test. In our study, the tests were proctored on computer. Only native English speakers graded test results. This thesis was not investigating the specific areas of linguistic strength and weakness of particular individuals. Rather, our interest was in the overall English language capabilities among students of varying levels in comparison to working memory levels. Thus, the EI test is a fitting tool to measure language ability for such purposes.

The working memory test involved a few additional production stages. The first version of the working memory test was helpful in providing a prototype for form, but its particular content proved to be problematic. The problems of the first working memory test will be further delineated below. Due to these problems, I developed a second working memory test, based on the format of the first, while learning from the experience and avoiding the pitfalls encountered in the first attempt. One of the principal reasons we created our own working memory test as opposed to using a previously existing one was to make its format and delivery method consistent with the English EI test. By creating our own memory test, we were able to use an EI approach thereby circumventing potential variables that may have arisen had we attempted to use a memory test that required a different type of delivery or response method.

Ebbinghaus (1885) was the first to use nonce syllables to test memory. Since then other have also used nonce syllable tests in memory research (Nevins & Endress 2007). As mentioned previously, DeKeyser (2005) noted that “nonmeaningful” elements drew on memory more than elements with meaning. Based on this previous literature, the format of the first version of the working memory test, and consultation with psycholinguist Dr. Wendy Baker, a nonce syllable test format was chosen. This format orders the questions in increasing length. For example, the first item would contain one syllable, the second item would contain two, and the third item would contain three, until the last item on the test, which contained ten syllables.

Working Memory Test Version #1

A PSST group member developed a working memory test consisting of 10 items. Based on the current literature, a nonce syllable test type was chosen since it circumvents the problems Cowan (2001) discussed regarding nonce word tests. These problems include insufficient internal “chunking” of syllables in multi-syllabic words. As opposed to multi-syllabic words, each syllable stands alone as a separate mental “chunk,” eliminating any need for internal cohesion. The syllables for the nonce syllable test we extracted from various parts of the EI test, but taken out of context and randomized, to lose all meaning and any possibility for logical “chunking.”

While the format and ideas behind this working memory test were based on accepted academic principles, there were multiple problems with its content.

The first problem was that the audio recordings for the test stimuli said "Number one," "number two...," etc. before each item, creating an additional distraction and potential memory load before the actual working memory items were given. The announcement of the item number, rather than just the nonce syllables, occurred on all ten items.

A second, more minor problem was that one of the written items did not exactly match the audio. In particular, the syllable “stu,” from item 5 was displayed as “stu,” in the database, but the audio recording sounded like “stewed.” This was not a critical error, as the entry in the database could easily be changed to “stewed,” to match the audio. However, both “stu” and “stewed” are actual words, rather than nonce syllables. This leads to the third problem, the accidental use of actual English words.

Although each syllable was supposed to be a nonce syllable, some of these isolated “syllables,” ended up being actual English words. The way these syllables were derived was through a program that randomly spliced syllables from words that were in the items of the actual English test. Perhaps the logic behind this was to eliminate ease or difficulty of articulation as a variable in comparing the scores. For example, the syllable “stu” was taken from the word “student.” Other actual words that resulted from the process were “of,” “knee,” “off,” and “teen.” Below is a chart of the items on the working memory test. Actual English words are shaded in gray.

Item	Syllable 1	2	3	4	5	6	7	8	9	10
1	of									
2	die	poe								
3	dren	all	act							
4	fore	evh	lease	span						
5	zert	woe	stu	be	ly					
6	ish	dee	un	cass	tar	hez				
7	sul	tate	less	yes	enz	der	fun			
8	neigh	ing	ter	liss	sock	moo	in	sum		
9	dus	beh	vee	bush	tus	mer	dunce	burs	vay	
10	im	off	knee	ohv	let	teen	chick	gus	jur	buh

Table 1. Working Memory Test Version #1

Closely related to the previous problem, the fourth was extensive inter-syllabic "chunking." For example, item eight, "neigh," is randomly followed by "ing," which makes it easy to chunk the two together to form the word, "neigh + ing," as in "The horse was neighing." The plus sign signifies points of chunking risk. Others include "off + knee" (from item 10 [9009]), all + act (from item 3 [9002]), "less + yes" (from item 7), and "let + teen + chick" (from item 10). This chunking-made-easy lends itself to inaccurate measures of working memory.

The fifth and last concern regarding this working memory test was that the length of time offered available to record an answer (repeat the stimulus) was too short. The amount of time to respond was only five seconds, even for the longer items. It was thought that the response time should grow in length as the items increased in length. Five seconds was an insufficient amount of time for the longer items, so that even if someone were able to memorize the string of syllables, they would not have time to record all of them.

Working Memory Test Version #2

In order to test the influence of working memory on the English EI test, for this thesis I developed a second version of the working memory test that addressed the issues that arose in the first version of the test. The nonce syllables provided in a table in the appendix of a PSST group member's unpublished paper served as the basis for creating the new working memory test (McGhee 2010). As the working memory test required single-syllable words, I used some monosyllabic words directly as they were, such as "kish" (Azuma & Van Orden 1997:500). Many of their pseudo-words were not monosyllabic, and so could not be used directly. These were split into single syllables, with some overlap. In some cases a letter was used twice to create two original nonce syllables. I split the nonce word, "opfu" from the article into "op" and “fu.” Other nonce words were one syllable, but I split them into two nonce words to get the greatest yield of nonce syllables possible. For example, I split “drerm” into “dre” and “erm.”

The 55 nonce syllables were deliberately arranged into ten items of increasing length, displayed in the table below.

Voice (m/f)	Item	Syllable 1	2	3	4	5	6	7	8	9	10
m	1	pav									
f	2	cler	heh								
f	3	ep	oove	ske							
m	4	hig	va	oif	mi						
m	5	op	fri	vlou	ib	pi					
f	6	unk	eathe	tuh	aif	ler	neh				
m	7	duve	brae	een	vli	fu	arl	dri			
f	8	fla	ug	lumn	olp	est	oze	ull	keb		
f	9	swa	erm	fruh	ilt	vuh	dre	voo	im	cuh	
m	10	ud	kish	fe	sli	ek	gra	erd	ra	jom	ep

Table 2. Working Memory Test Version #2

Caution was used to ensure that the syllables were not homophones of actual English words when pronounced either independently or in sequence in their item. Dr. Deryle Lonsdale, Dr. Ray Graham, and psycholinguist Dr. Wendy Baker reviewed the test for possible errors or weak points. All three approved it as a legitimate working memory test.

In addition to the changes in syllable content, other changes were made in the second version of the working memory test. The first such change was the deletion of the vocalizations of the item number (“Number one…,” “Number two…,” etc.) in the process of cutting audio files. Another change was extending the response time for test-takers beyond the five seconds allowed in the original working memory test. I also arranged to alternate the gender of the voice for recording stimulus items. In the original working memory test all the recordings were also done in a male voice. This was not necessarily a problem, but since we had to redo the recordings anyway, we decided to account for gender of voice as well. In the new version of the test, five of the ten items were recorded by a male voice, and the other five were recorded by a male voice, in randomized order. Also, syllables within a single item were separated by pauses of approximately one second in length. I used the audio program Audacity[3] to adjust lengths of pauses where they had been recorded with irregularity. Care was also taken to avoid intonation of syllables that would make them easier to remember in sequence.

3.3. Test Administration

The English Language Center's 94 new incoming students in January of 2011 all took the English EI test in the first week of January. It consisted of a 60-item English EI test as a part of an initial assessment at the start of their program, and was given at the same time as other diagnostic language tests. However, the EI test results were not used in placing students in their ELC levels. All testing occurred in the ELC computer lab on Mac computers. Each student sat at a computer with headphones that contained a microphone piece. Every test began with a test of 30 items of the English EI test without comprehension tasks, concluding with the remaining 30 items of the test, with tasks. Half of the students had items 1-30 of the original Form D without comprehension tasks, and items 31- 60 with tasks. This group of students will be referred to as Group A. The other half of the students, Group B, had items 31-60 of the original form without the tasks, and items 1-30 with the tasks. While the specific items with comprehension tasks and without them differed between the two groups, both groups had the section without the tasks first, and the section with the tasks second.

The order of items presented in the English language portion of the test was randomized within each set (questions 1-30, and questions 31-60, respectively). This type of randomization

[3] http://audacity.sourceforge.net/

was intended to eliminate the effect of item-order, as well as inhibit students from being influenced by others while taking the test.

Students took a web-based version of the working memory test, through the PSST3 server by accessing the website on ELC lab computers. Because the test server was new, collecting data through the web-based version of the test was somewhat of an experimental endeavor. For this study, a maximum of five to six students took the test simultaneously. Usually there were only one to three students taking it at once.

Web-based version of Working Memory Test

Fortunately, only 38 of the students had taken the test when the problem of data loss delineated above became apparent. According to the table above, the computer generated 41 test-taker IDs, which would imply that 41 people had taken the test that day. Three test-taker ID numbers are completely missing from the database. For some reason, the computer skipped these. This still leaves 39 test-taker ID’s to be accounted for. In analyzing the correlation of Subject_ID’s to Test_ID’s in the database, it was discovered that three students had logged in more than once, so had multiple Test ID’s. When these are taken into account the correct total of different students is 38.

Standalone version of Working Memory Test

To mitigate further data loss, we developed a standalone version of the memory test. All 10 items were exactly the same, and remained in order. The main difference was that the test was downloaded onto each computer and administered locally, instead of being web-based. The instructions were also exactly the same. However, one other difference was that the web-based version proceeded immediately from one item to the next, without much of a pause in between.

The standalone version stopped after each response was recorded, and the test-taker was responsible for clicking the “next” button to continue on to the next item. This may have allowed the test-takers who took the standalone version to clear their thoughts better between each item, resulting in improved performance in comparison to those who took the web-based version and may have felt rushed. Because the web-based version only allowed for set response times, I altered the response times for the standalone test. The formula for response times was:

(number of nonce syllables in the item) x 2 + 3 = response time in seconds

Because the working memory test was in increasing order of length/nonce syllables, “number of nonce syllables” was also equivalent to the item number. For example, item 1 contained one nonce syllable. Item 2 contained two, item 3 had three, and so forth, up to item 10. For example, length allowed for response time for item 1 is derived as follow:

1 x 2 +3 = 5 seconds

Calculation for item 10 was:

10 x 2 + 3 = 23 seconds

Figure # X, *Item response times*, below lists the item lengths in syllables and response times granted for each.

Number of nonce syllables	1	2	3	4	5	6	7	8	9	10
Allotted response time (in seconds)	5	7	9	11	13	15	17	19	21	23

Table 3. Item response times

29 students took the standalone version. 15 of the 38 web version tests retained audio recordings for all 10 items. These two groups combined total 44 individuals' tests with complete results available, roughly half of the original number who took the EI test. The number of students in the final analyses reduced to 40, as a few students did not have English EI scores in the database. However, the data set contained enough usable information to proceed with the analysis.

3.4. Grading

Test results were downloaded into a database, and entered into a program in a user-friendly interface for human grading. The grader consisted of a website which research group members could log into. When a person logs in, an item from the EI test appears on the screen, divided by syllables. An audio recording of a test-taker responding to the stimulus also loads onto the screen. The rater listens to the audio recording, and marks each syllable that is present with a "1," and each syllable in the utterance that the test-taker missed with a "0." Six individuals graded results for the 5,430 English test items (89 test-takers times 60 items per test), and the 890 working memory test items. All graders had been trained in how to score the items on a syllable-by-syllable basis.

Once the items were graded independently, the percentage was taken of syllables pronounced correctly in the response over total number of syllables in the stimulus utterance. In previous tests, this research group has experimented with using various scoring methods. These include on a word-by-word basis (in which one incorrect syllable would make the entire word incorrect), a 4-score method, a holistic rating, and a percentage rating based on the number of correct uttered syllables over the number of total syllables in each item. Many EI researchers have used the 4-score method of grading tests (Chaudron et al. 2005; Christensen et al. 2010). With this method, if a test-taker repeats the utterance correctly, and all syllables are present, the item is given a score of 4. For each error (e.g. missing syllable or transposition), the rater subtracts one point. So, one missing syllable would result in a score of 3. Two missing syllables would be a score of 2. Three missing syllables would be a score of 1, and four or more missing syllables would be a score of 0. While the 4-score method has worked for previous research, it does not account for different-length stimuli sentences. For this study, test items were graded using the syllable-by-syllable percentage method. We felt that taking a percentage of total syllables pronounced over total syllables in the stimuli would give a more accurate assessment, as it allows more degrees of variance.

$$\frac{\text{Number of syllables test-taker says correctly}}{\text{Total number of syllables in stimulus sentence}} = \text{\% Score}$$

The percentage method of scoring was used for both the English EI test as well as the Working Memory EI test.

Together, Form D combined with the working memory test addresses the issues outlined by Jessop et al. (2009). One issue was to ensure clear instructions. They noted that very different responses could result from differing instructions. For example, “Repeat the utterance exactly as you heard it,” versus “Say the correct English sentence,” would have very different results. This was mostly in reference to eliciting spontaneous correction of ungrammatical items. Although we did not use any ungrammatical sentences, to prevent this type of confusion or ambiguity, instructions for both the English test and the working memory test were clear. Test-takers were instructed to repeat each utterance exactly as they heard it.

Practicality of resources and instruments was brought up as an issue to address in administering EI tests. This was not a problem for our research group, as the university's English

Language Center provided the necessary number of new ESL students from varying levels, and the computer lab with headphones and microphones for students to listen to and answer EI tests. The research group also had several students (both graduate and undergraduate) to score the tests, as well as the funding to provide incentive for students to spend several hours on it, to complete scoring more quickly.

Regarding EI testing, it has been suggested that "They [researchers] must ensure that participants' confidentiality and anonymity are maintained" (Tomita 2009:348). Students who participated in this research logged in with ID numbers for their test. The names of participants were never given to scorers or other members of the research group. We maintained anonymity by identifying them solely by nine-digit ID numbers. Also, in reporting the results, none of the identification numbers will even be used, further ensuring protection of privacy and maintenance of anonymity.

4. Results and Discussion

4.1. Results

This chapter presents results from the ELC students' working memory and English EI tests. To examine the authenticity of the working memory test, we performed analyses that will be explained in further detail. This chapter also explains how we obtained these results, the statistical method we chose and the reasons we chose it, the three data sets we compared, and the results from these three analyses. I have also inserted scatter plots to provide visual depictions of the correlations.

To verify the validity of the working memory test (i.e. ensure it was actually testing memory), I analyzed the average scores of all participants for each item of the working memory test. As noted in Chapter 3, items were presented to test-takers in order from shortest to longest. Item 1 consisted of one nonce syllable, Item 2 consisted of two, and Item 10 consisted of ten nonce syllables, separated by one second each. Figure 1 illustrates the results.

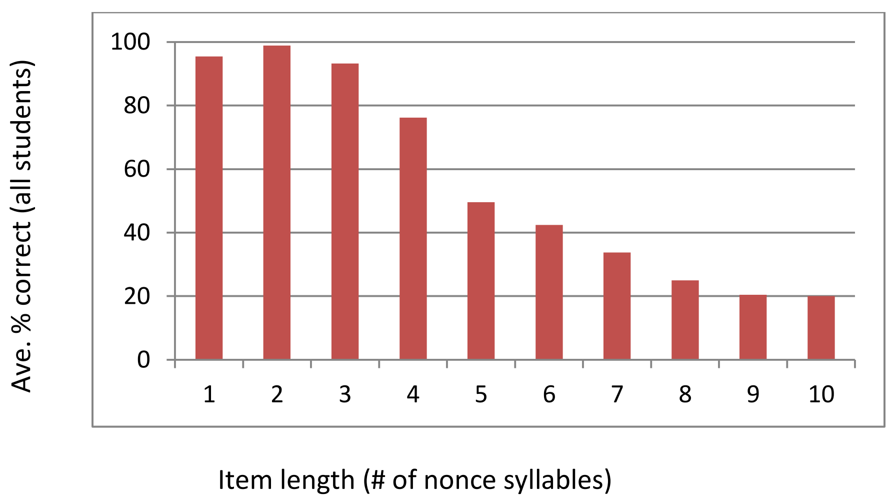

Figure 1. By item analysis of average WM test scores across all students

Average scores of all students on items that consisted of one, two, and three nonce syllables were almost 100% accuracy. At four nonce syllables, the average percentage dropped to just below

80%. After four syllables in length, there was an even greater drop off: at five syllables the score dropped below 50% (49.55%). This coincides with the argument from Cowan (2001) that the "magical number" of working memory capacity for the average person was 4 ± 1. These findings suggest that our WM test was actually measuring memory as intended.

Due to the nature of the hypothesis, we chose to use a Pearson correlation to analyze the data. To reiterate, it was hypothesized that this type of EI language test functions primarily as an assessment of linguistic ability, and not as a rote memory test, on the basis that "nonmeaning" draws more on memory than "meaning" does (DeKeyser 2005). Some positive correlation was expected, as WM is associated with language learning aptitude and language production (Doughty & Long 2003; Robinson 2005). Because the hypothesis revolved around a correlation between English EI scores and working memory scores, we knew we wanted to use a correlation test. I used a Pearson correlation.

We used the program SPSS[4] to run three data sets to analyze Pearson correlations across individuals': (1) English EI item score and working memory test score; (2) working memory test score and ELC level; and (3) ELC level and English EI score. The results of the first set are of the greatest consequence to this study, as they most directly address the hypothesis in question. The English EI test consisted of 60 items, and the working memory test consisted of 10 items. Both tests were scored by a percentage method, i.e. the overall score on the test is a percent of the total number of syllables the student uttered correctly over the total number of syllables in the stimuli of the test. The ELC in essence has 6 grade levels: Level 0, Level 1, Level 2, Level 3, Level 4, and Level 5. Students are placed at these levels based on multiple means of assessing their abilities, including grammar, writing, and listening comprehension tasks.

Table 4 below contains the results of both correlation analyses. The column labeled "English EI" contains the correlation across the percentage-based scores—individual students'

---

[4] http://spss.en.softonic.com/

English EI test scores and WM test scores. The column labeled "ELC Level" contains the correlation across individual students, between their percentage score on the WM test and their ELC level. The results are as follows:

		English EI	ELC Level
Working memory	Pearson Correlation	.249	.130
	Significance (2-tailed)	.121	.430

Table 4. Working memory to English EI and ELC Levels - Pearson correlations

Individuals' overall scores on the working memory test and the English EI test do have a slight positive correlation. However, the Pearson correlation is not significant (r = .249), with any reasonable value of *p* (*r* = .121, *p* > .05).

The correlation of working memory test scores and ELC level is similar to that of Working Memory and English EI scores. While N = 40 for the WM to EI correlation, with WM to ELC level, N = 39. This is because ELC level information was unavailable for one of the 40 students. The last column of Table 4 above reveals that the correlation is also not significant (r = .130). Again, there is a slight positive correlation, yet not significant (*r* = .430, *p* > .05). As even the ELC has a slight positive correlation with WM test scores, then it is not unreasonable that the English EI test also has a slight positive correlation with regard to the WM test. The similarity of these results to those of the English EI-Working memory correlation suggests that the slight positive correlation is due to the part of language ability, as measured by the ELC diagnostics, that requires working memory, and not that the EI test is a memory test.

The third set of data analyzed with a Pearson correlation was students' ELC level to their English EI test score. This analysis gives a comparison between the English EI test and the overall holistic grade level determined by multiple diagnostics, as explained above.

		ELC Level
English EI	Pearson Correlation	.786
	Significance (2-tailed)	.000

Table 5. English EI scores and ELC Level – Pearson correlation

The correlation between English EI score and ELC Grade Level is significant (r = .786, p = .000, $p < .01$). This finding suggests that the English EI test does correlate to language rather than just memory. The EI test does seem to measure linguistic oral competency, at least as defined by the ELC.

The scatter plot graphs present a visual depiction of the relationship in each of the three comparisons above. In Figure 1, which plots English EI scores and working memory scores, each plot-point represents an individual student who took both tests. Clearly several students have average working memory scores, but have very divergent scores on the English EI test.

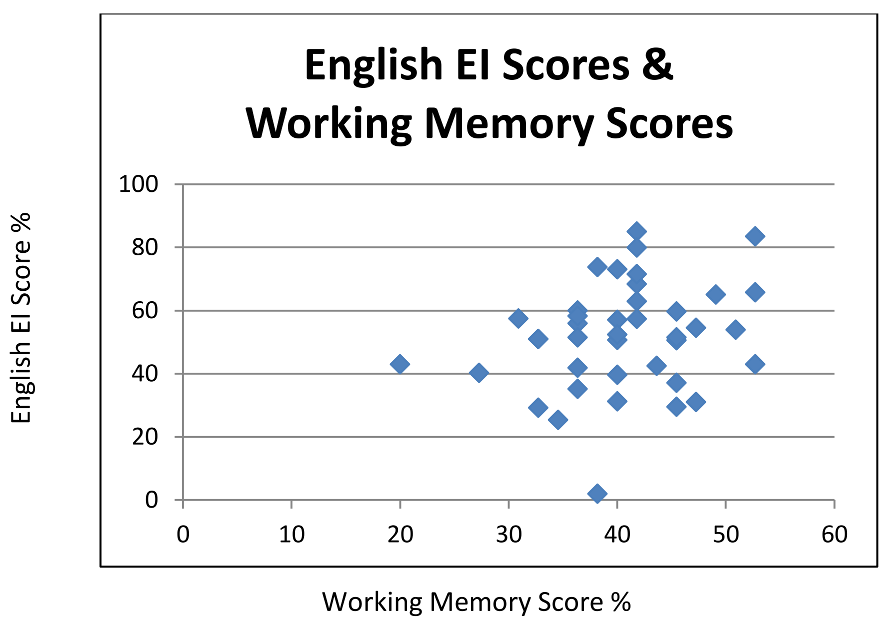

Figure 2. Scatter diagram of English EI scores and working memory scores

The mode tends to cluster around 40% on the working memory test. Yet of these, there is almost a vertical line ranging almost the entire spectrum of English language ability. The highest is 84.97% and the lowest is 1.9%, with everything from scores in the 20s, 30s, 40s, 50s, 60s, and 70 percentile. We also see students with lower than average working memory scores, but average English scores. On the other hand, there are students who did not do as well on the English EI test, but have above average working memory.

The next two figures depict the relation between ELC level and working memory scores, and ELC level and the English EI test. If the relationship between students' ELC level and working memory scores were to correlate significantly, the points would generally align on a diagonal slope. This would mean that students with high WM scores were in higher ELC levels, and students with the lowest WM scores would be in the lowest ELC grade levels. However, as is demonstrated in Figure 2 below, the points do not align neatly. For example, the student at the highest ELC level, Level 5, had a WM score of 41.82%. A student at the lowest ELC grade, Level 0, had a WM score of 47.27%. The mean of all WM scores was 40.50%. In other words, a student at the lowest ELC level had a WM score higher than a student at the highest ELC level. There is thus little to no correlation between the two.

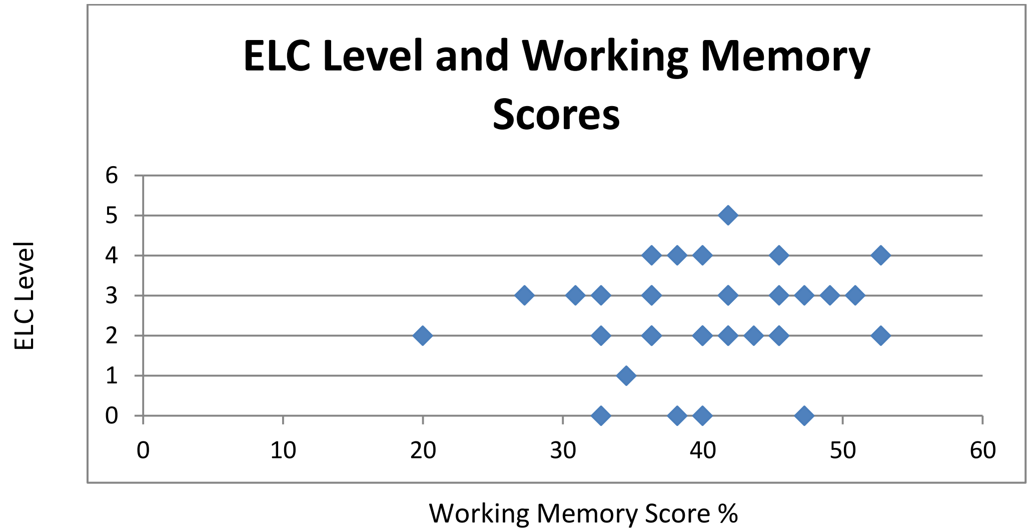

Figure 3. Scatter diagram of ELC levels and working memory scores

Figure 3 plots students' ELC level to their English EI scores. This graph reveals a much clearer correlation. One could draw a diagonal line from the origin to the upper-right corner and find most of the points on or near that line.

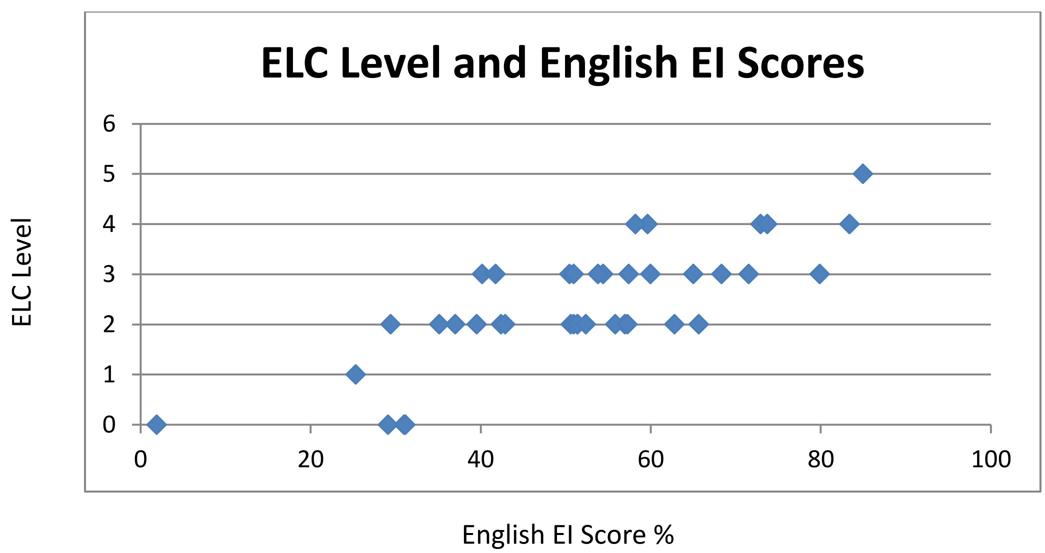

Figure 4. Scatter diagram of ELC levels and English EI scores

Of the three students in Level 0 at the ELC (the ELC's lowest beginning level), one of them scored the very lowest on the English EI test, being able to correctly repeat only 1.9% of the syllables on the test. This student's working memory score was perfectly fine, at 38.18%. The other two Level 0 students scored 29.11% and 31.01% respectively on the English EI test. Further, note that the lowest beginning level students could not break past the 30-40% barrier on the English EI test. Also, recall that a widely accepted value for the working memory magical number is 4 ± 1. While we were unable to test the specific number of chunks a person could retain, and so do not have a number to correlate to Cowan's magical number, the principle remains the same: there is a limit to the average human working memory capacity, and it generally varies very little. Note again that the mode of WM scores hovered around 40%, and all the scores ranged from 27.27% to 52.73%, with one outlier at 20%. The mean of all WM scores was 40.50%. Disregarding the single outlier, the range is roughly 30% to 50%. While the direct relationship between the magical number and the percentage scores on our WM test is beyond the scope of this thesis, the numbers are uncannily reminiscent of Cowan's 4 ± 1 (an average of 4, with a general range from 3-5). Having little to no knowledge of the English language, it appears that the lowest-scoring students were unable to repeat anything beyond their working memory capacity. So it seems that below scores of roughly 30%, working memory tends to be the primary contributor to the test-takers' English EI results. Above 30%, knowledge of the language seems to have a greater influence, and working memory proportionately less so. This particular result follows our prediction—that the English EI test would primarily be an assessment of an individual's linguistic ability, and not an assessment of their working memory capacity, at least not significantly.

4.2. Discussion

The non-significant results of ESL students' working memory scores' correlation to English EI scores suggest that the PSST group's English EI test functions primarily as a measurement of English language aptitude, and not as a rote memory test. While not significant, the slight positive correlation between the two scores can be explained by the very nature of how language and working memory interact. Working memory is inherent in all linguistic communication to some degree (Doughty & Long 2003; Robinson 2005). This may account for the majority of the overlap.

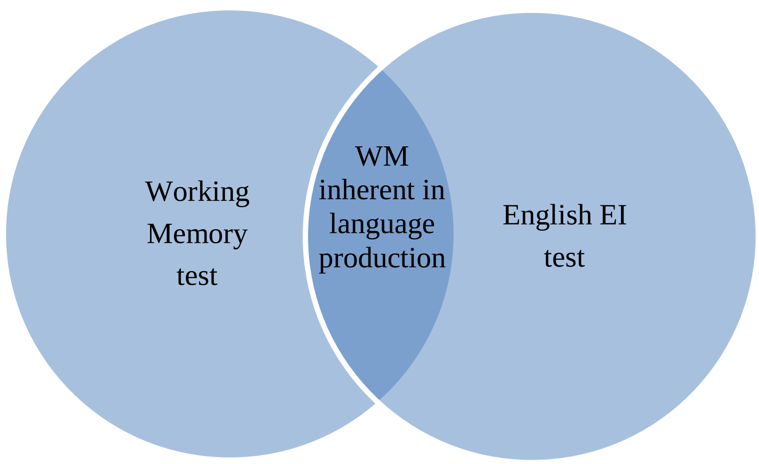

Figure 5. Sample overlap between working memory and L2 EI tests

An understanding of current literature on implicit and explicit language knowledge assists in understanding what role working memory plays in the EI test, and possible explanations as to why we received the results we did. Some scholars agree that "It is implicit rather than explicit language knowledge that linguistic competence is comprised of" (Erlam 2006:465). However, most types of second language capacity tests assess explicit knowledge. One study demonstrated that language tests can be constructed in a way to distinguish explicit language knowledge from implicit (Ellis 2005). In that study, Ellis stated that the requirements for a test that could diagnose implicit knowledge would be (1) the test must be time-constrained, and (2) the test must get at meaning, and, as much as possible, avoid focusing on form.

In this study, our English EI test met with Ellis' first constraint to test implicit knowledge, and may have met with the second. Our test was time-constrained. Test-takers were only allowed a few seconds to respond, and the response time began automatically after the stimulus was given. This did not give students a chance to analyze sentence structure or grammar, or indeed rehearse the item before repeating it. Although we were not measuring actual language chunking in this test (Cowan 2001), if chunking by meaning is the mechanism allowing higher-level ELC students to perform better on the English EI test, then it may be that this EI test (or at least some of its items) also meet Ellis' second requirement, which is to focus on meaning rather than form.

For shorter stimuli that allow test-takers to operate within the constraints of working memory, the focus is on separate sounds, like a nonce syllable working memory test. This is especially true of students at Level 0, who have little knowledge of the language. For test-takers of slightly higher English language abilities, recalling the string of syllables of which actual words and entire utterances are composed may become form-focused. The EI test may be testing both explicit and implicit knowledge of the language. But lengthier sentences are too long for working memory capacity, and perhaps form-focused chunking does not sufficiently recode so many syllables into few enough chunks to retain in working memory. It is possible that the longest sentences on the test can be remembered and uttered correctly more easily by those who can recode it into its meaning, which accesses their implicit knowledge of the language.

EI test items may measure working memory, explicit knowledge, and implicit knowledge depending on length and/or linguistic features. As discussed previously, working memory operates by recoding information, to enable to retain more at once. Language—or rather knowledge of a language—allows a person to recode the information they receive, enabling him/her to remember and reproduce longer utterances than a person who has no knowledge of the language. It is possible that as either the length or the complexity of a sentence increases, what the test-taker is most reliant on to repeat it precisely changes from working memory alone, to working memory drawing on linguistic knowledge, once the stimulus surpasses the limits of working memory capacity.

In the interaction between language knowledge and working memory, the mechanism of recoding takes sheer sound (syllables), and recodes it into a structure or form. For the advanced student, a likely possibility is that the original sounds are recoded even further into meaning. The meaning would may allow the test-taker to "chunk" the separate "bits" (Miller 1956) of information together so tightly that they would be able to recall a complex, lengthy utterance perfectly, or at least nearly perfectly, when a novice would be able to recall only a few, if any, syllables from it.

Presumably, the length of the item (in terms of syllables), in combination with deliberately chosen linguistic features will determine what that particular item is testing for. For example, one of the items on our English EI test, "Joe writes poetry," is short enough that working memory may have a greater correlation to scores on this particular item than the test overall. While due to its length it could be considered a working memory test item, its carefully designed grammatical features make it a test of at least some explicit linguistic knowledge. While spontaneous correction of ungrammatical sentences is telling and gets at higher levels of implicit knowledge under time pressure, spontaneous altering of perfectly grammatical sentences into ungrammatical utterances is equally telling. The table below contains a list of common errors students made on the English EI test, with specific examples.

Error Type	Original Stimulus Sentence	Student's Response
Loss of plural	Big ships always make noise.	"Big ship will make noise"
Loss of conjugation	Joe writes poetry.	"Joe write poetry"
Incorrect article, loss of articles	"Did you buy that at the store?"	"Did you buy that at a store?"
Tense change	He had to have been falling in love. I have never met the girl in the green shirt.	"He have to be falling in love" "I have never meet the girl with the green shirt"
Meaningless utterances	Do good children run from their parents?	"Do good childrens dance in the Paris?"[5]
Vocabulary replacement	They should have eaten breakfast by now.	"They shouldn't eat breakfast right now"

Table 6. English EI test-taker error types

Complex, lengthy items such as "When Jim entered the office he was immediately afraid of the uncommunicative boss," and "Hesitating before she spoke her next line the actress reached the pinnacle of her nervousness," likely contain too many separate features for a person wholly dependent on explicit knowledge to maintain all the separate elements in their working memory. A translation of the utterance into meaning (which would require implicit knowledge) would allow the test- taker to respond correctly. If such is the case, then specific items could be used to distinguish between working memory, explicit knowledge, and implicit knowledge of the language.

In summary, the lack of significant correlations between working memory and English EI scores and between working memory and ELC levels, and the significant correlation between English EI scores and ELC levels suggest that there is more to performance on EI tests than working memory capacity. Ellis proposed that EI tests could be developed to test for different types of knowledge, namely, explicit linguistic knowledge and implicit linguistic knowledge. He stated that time-limitations and a focus on meaning were necessary to construct an EI test that measured implicit linguistic knowledge. The test-takers' errors indicate that using items with specific grammatical features can distinguish between levels of ability. Beginning level test-

[5] This particular response contains errors with pluralizing, incorrect articles, and vocabulary replacement as well.

takers had difficulty with even very short items, if they contained specific grammatical structures beyond their linguistic ability. Long sentences were also difficult or impossible for the beginning-level students, but doable for advanced speakers. These findings produce further research questions. Do utterances that are too long to repeat using average working memory capacity require chunking by meaning to repeat correctly? If EI performance on longer sentences does access a translation into meaning, then according to Ellis' framework, these types of long utterances also test implicit knowledge. A further question still is, do test-takers even recall hearing those morphemes they were unable to accurately repeat, such as the word-final 3rd person singular "s" (e.g. "run*s*")?

5. Conclusion

The findings of this thesis support the hypothesis that language EI tests primarily target linguistic ability, and do not reflect working memory capacity to a significant degree. This is an agreement with DeKeyser's findings (2005) that elements without meaning (in this case, nonce syllables) drew on memory more than elements of meaning. Some working memory is involved in language learning and production (Robinson 2005), and so working memory does correlate positively to EI scores as expected, but not to a significant degree. This suggests that the English EI test is testing linguistic knowledge. Working memory also correlates positively to ELC level, although, like the correlation with EI scores, not significantly.

There are several implications and potential benefits of these findings. The correlation of EI scores to students' ELC level was so strong that it is reasonable and natural to consider the possibility of using EI tests in the form of diagnostic or placement exams. This would be extremely useful for several reasons. For large institutions with mass amounts of potential/entering students, the EI test can be administered to any number of students at any time, without the need of additional administrative personnel. For any type of institution, the EI is a much cheaper and more efficient way to assess people. It could be used in courts to determine the need for an interpreter, in language learning programs to determine individuals' starting levels, in government language programs to determine entering level, progress, and overall linguistic ability. It is also much easier to grade this type of assessment than unscripted, spontaneous speech. In addition to being easier to grade, the evaluations would also be more consistent and objective, with less chance of arbitrary, subjective, and/or idiosyncratic influences involved in administering and grading language assessments.

While the hypothesis proved to be correct, and the study was useful, there were a couple of potential limitations in this study. The first is the relatively small sample size of test-takers

(40). As was described in Chapter 3, we originally began with 94 participants. However, due to students moving away, dropping out of the ELC program, and unavailability during testing, that number dropped significantly. Another contributing factor to such a small sample size was data loss through unforeseeable computer program problems. However, we were still able to attain complete data sets for 40 individuals.

Another possible limitation involves the content of the working memory test. The goal was to use only nonce syllables that would have no meaning to the test-takers. The limitation lies in the construction of the syllables to have no meaning from an English language perspective. However, there is a possibility some of the "nonce" syllables were words or morphemes in other languages, perhaps even in languages spoken by the test-takers as among them there was a wide range of native languages.

Several aspects of the research that could be improved if someone were to carry out a similar study. Because of performance issues with the web administration of the EI tests, some data was lost, as noted earlier. These issues should be addressed before future follow-up work is done. Directly related to the issue of lost data is the number of students involved in the study. It would be preferable to have a greater number of students take the test. This would not have been as much of a problem for this study if the web test had functioned properly. In addition to correlating working memory to EI scores and ELC levels, another valuable research question may be to investigate the correlation between OPI scores and working memory.

In the future, it would be interesting to look at students of each grade level, and specific items that served as "breaking points" for each level, in terms of item length and complexity. This may be useful in determining which specific items are beyond the reaches of basic working memory, truly testing implicit knowledge of the language. For such purposes, it may be useful to involve an equal number of students from each ELC level (or equivalent ranking, if at a different institution). Then one could see the effects of working memory on each specific EI item, rather than just the test as a whole.

While we must always operate within the finite limits of working memory, language seems to be able to enhance the amount of information we can contain within those limits by recoding individual syllables into larger "chunks" of information. Implicit knowledge of a linguistic principle that originates from explicit knowledge, experience, and deepens through exposure, can then allow a person to remember more than was originally possible. Comprehension of meaning seems to be one of the most powerful forms of recoding, allowing more to be retained in working memory than even knowledge of grammatical principles.

If there is a limit to all things and a measure
And a last time and nothing more and forgetfulness…

There is among all your memories one
Which has now been lost beyond recall…

You will never recapture what the Persian
Said in his language woven with birds and roses,
When, in the sunset, before the light disperses,
You wish to give words to unforgettable things...

That brief moment in which a spark of information can live in working memory, "In the sunset, before the light disperses," while nearly all else is "lost beyond recall," we have an opportunity to give meaning to things, thereby overcoming, or at least extending the scope of limits. The words we utter are perhaps merely the reflection of meaning, one of those few "unforgettable things."

Bibliography

Azuma, Tamiko and Guy C. Van Orden. 1997. Why Safe is Better Than Fast: The Relatedness of a Word's meanings affects lexical decision times. Journal of Memory and Language 36.484-504.

Baddeley, Alan D. and Graham Hitch. 1986. Working Memory. Oxford: Oxford University Press.

Bley-Vroman, Robert and Craig Chaudron. 1994. Elicited Imitation as a Measure of Second Language Competence. Research methodology in second-language acquisition, Hillsdale: Lawrence Erlbaum. 245-261.

Chaudron, M. Prior, and Uli Kozok. 2005. Elicited imitation as an oral proficiency measure. Paper presented at the 14th World Congress of Applied Linguistics, Madison Wisconsin.

Christensen, Carl, Ross Hendrickson, and Deryle Lonsdale. 2010. Principled Construction of Elicited Imitation Tests. In (N. Calzolari, K. Choukri, B. Maegaard, J. Mariani, J. Odijk, S. Piperidis, M. Rosner, and D. Tapias, Eds.) Proceedings of the 7th Conference on International Language Resources and Evaluation (LREC '10). European Language Resources Association (ELRA): Valetta, Malta 233-238.

Connell, Phil J. and Catherine Myles-Zitzer. 1982. An analysis of elicited imitation as a language evaluation procedure. Journal of Speech and Hearing Disorders 47.390-396.

Cowan, Nelson. 1996. Short-Term Memory, Working Memory, and Their Importance in Language Processing. Topics in Language Disorders 17.1-18.

Cowan, Nelson. 2001. The Magical Number 4 in Short-Term Memory: A Reconsideration of Mental Storage Capacity. Behavioral and Brain Sciences 24.87-185.

DeKeyser, Robert M. 2005. What Makes Learning Second-Language Grammar Difficult? A Review of Issues. Language Learning 55.1-25.

Doughty, Catherine J. and Michael H. Long. 2003. Optimal Pscyholinguistic Environments for Distance Foreign Language Learning. Language Learning & Technology 7.50-80.

Ebbinghaus, Hermann. 1885. On Memory (H.A. Ruger & C.E. Bussenius, Trans.) New York: Dover, 1964.

Ellis, Nick C. 2002. Frequency Effects in Language Processing. Studies in Second Language Acquisition 24.143-188.

Ellis, Nick C. 2008. The Dynamics of Second Language Emergence: Cycles of Language Use, Language Change, and Language Acquisition. The Modern Language Journal 92.232-49.

Ellis, Rod. 2005. Measuring Implicit and Explicit Knowledge of a Second Language: A Psychometric Study. Studies in Second Language Acquisition. 27:141-172.

Eriksen, Charles W. and Harold W. Hake. 1955. Absolute judgments as a function of the stimulus range and the number of stimulus and response categories. Journal of Experimental Psychology 49.323-332.

Erlam, Rosemary. 2006. Elicited Imitation as a Measure of L2 Implicit Knowledge: An Empirical Validation Study. Applied Linguistics 27.464-491.

Erlam, Rosemary. 2009. The Elicited Imitation Test as a Measure of Implicit Knowledge. Implicit and explicit knowledge in second language learning, testing and teaching. Ed: Ellis,R. Bristol, UK: Multilingual Matters. 65-92.

Gass, Susan M. and Alison Mackey. 2007. Data elicitation for second and foreign language research. Mahwah, New Jersey: Lawrence Erlbaum Associates.

Hatfield, Deborah H., Marie C. South, and Stuart D. Showalter. 2007. The sentence repetition test (SRT) revisited: evaluation of the SRT against the Second Language Oral Proficiency Evaluation (SLOPE) as an assessment tool for community bilingual proficiency. SIL Electronic Working Papers 2007-2008.13.

Jessop, Lorena, Wataru Suzuki, and Yasuyo Tomita. 2007. Elicited Imitation in Second Language Acquisition Research. The Canadian Modern Language Review 64.215-238.

Jessop, Lorena, Wataru Suzuki, and Yasuyo Tomita. 2009. Duff, Patricia A. (ed.) Elicited Imitation: Toward Valid Procedures to Measure Implicit Second Language Grammatical Knowledge. TESOL Quarterly 43.345-50.

Lehman, Margaret T. and Connie A. Tompkins. 1998. Reliability and validity of an auditory working memory measure: data from elderly and right-hemisphere damaged adults. Aphasiology 12.771-785.

Martin, Nadine and Jennifer Ayala. 2004. Measurement of auditory-verbal STM span in aphasia: Effects of item, task, and lexical impairment. Brain and Language 89.464-483.

McGhee, Jerry. 2010. Lexical Decision Reaction Times: The effects of ambiguity and bilingualism. Unpublished paper.

Miller, George A. 1956. The Magical Number Seven, Plus or Minus Two: Some Limits on Our Capacity for Processing Information. Psychological Review 63.343-355

Morra, Sergio. 2000. A New Model of Verbal Short-Term Memory. Journal of Experimental Child Psychology 75.191-227.

Nevins, Andrew and Ansgar Endress. 2007. The edge of order: Analytic bias in ludlings. Harvard Working Papers in Linguistics 12: 43-53.

Radloff, Carla F. 1992. Sentence Repetition Testing for Studies of Community Bilingualism: An Introduction. Notes on Linguistics 56.19-25.

Robinson, Peter. 2005. Aptitude and Second Language Acquisition. Annual Review of Applied Linguistics 25.46-73.

Slobin, Dani and Charles A. Welsh. 1967. Elicited imitation as a research tool in developmental psycholinguistics. Working Paper 10. Reprinted in C.A. Ferguson & D.I. Slobin (1973). Studies of child language development. New York: Holt, Rinehart and Winston Inc. 485-489.

Vinther, T. 2002. Elicited Imitation: a brief overview. International Journal of Applied Linguistics 12.54-73.

Test-taker ID	# syllables recorded	Item 1	Item 2	Item 3	Item 4	Item 5	Item 6	Item 7	Item 8	Item 9	Item 10
1507	2	x	0	0	x	0	0	0	0	0	0
1508	9	x	0	x	x	x	x	x	x	x	x
1509	8	x	x	x	x	x	0	x	x	x	0
1510	10	x	x	x	x	x	x	x	x	x	x
1511	10	x	x	x	x	x	x	x	x	x	x
1512	6	x	x	x	0	0	0	x	0	x	x
1513	10	x	x	x	x	x	x	x	x	x	x
1514	10	x	x	x	x	x	x	x	x	x	x
1515	10	x	x	x	x	x	x	x	x	x	x
1516	8	x	x	x	x	x	x	x	0	x	0
1517	10	x	x	x	x	x	x	x	x	x	x
1518	9	0	x	x	x	x	x	x	x	x	x
*1519	missing										
1520	1	x	0	0	0	0	0	0	0	0	0
*1521	missing										
1522	4	x	x	0	x	0	0	0	x	0	0
1523	8	x	x	0	x	0	x	x	x	x	x
1524	10	x	x	x	x	x	x	x	x	x	x
1525	10	x	x	x	x	x	x	x	x	x	x
1526	8	x	x	x	x	x	x	x	0	0	x
1527	1	0	0	0	0	0	0	0	x	0	0
1528	10	x	x	x	x	x	x	x	x	x	x
1529	10	x	x	x	x	x	x	x	x	x	x
1530	9	x	0	x	x	x	x	x	x	x	x
1531	3	0	x	x	0	0	0	x	0	0	0
1532	10	x	x	x	x	x	x	x	x	x	x
1533	10	x	x	x	x	x	x	x	x	x	x
1534	9	x	x	x	x	x	x	0	x	x	x
1535	6	0	0	0	x	x	x	0	x	x	x
1536	2	0	0	0	0	0	x	0	0	0	x
1537	1	0	0	0	0	0	0	0	0	0	x
1538	4	x	0	0	0	x	0	0	x	x	0
1539	9	x	x	x	x	0	x	x	x	x	x
1540	10	x	x	x	x	x	x	x	x	x	x
1541	9	x	x	x	x	x	x	0	x	x	x
*1542	missing										
1543	9	x	x	x	x	x	x	x	0	x	x
1544	7	x	x	x	x	0	x	0	x	x	0
1545	10	x	x	x	x	x	x	x	x	x	x
1546	9	0	x	x	x	x	x	x	x	x	x
1547	10	x	x	x	x	x	x	x	x	x	x
1548	9	x	x	x	x	x	x	x	0	x	x

Appendix: Analysis of missing items from web test