The Role of Lexical Choice in Elicited Imitation Item Difficulty

C. Ray Graham, Jeremiah McGhee, and Ben Millard
Brigham Young University

1. Introduction

Elicited imitation (EI) has been used for decades as a means of examining the development of oral language skills in various contexts including normal native language development (Ervin-Tripp, 1964; Keller-Cohen, 1981; Menyuk, 1963;) abnormal language development (Berry, 1976; Lahey, Launer, & Schiff-Myers, 1983; Menyuk, 1964) and second language development (Hamayan, Saegert, & Larudee, 1977; Naiman, 1974). In recent years there has been a resurgence of interest in its use in the examination of oral language skills in second language learners (Chaudron, Prior, & Kozok, 2005; Erlam 2006; Jessop, Suzuki, & Tomita, 2007; Vinther 2002). For a fairly comprehensive review of this literature, see Bley-Vroman and Chaudron (1994), Gallimore and Tharp (1981), Lust, Chien, and Flynn (1987), and Vinther (2002).

An examination of the literature on EI suggests that interest in its use centers around two major investigative efforts, (a) psycholinguistic research into the nature of language competence itself (see, for example, Ellis, 2006, and Erlam, 2006), and (b) research into the possibility of finding an indirect and efficient way to estimate the overall oral language proficiency of second language learners (see Chaudron, et al., 2005, and Radloff, 1992). While these two purposes have much in common, it seems to us from the available literature that they may make quite different demands on the design and administration of EI items. If one is to investigate the nature of interlanguage, certain conditions for the elicitation of responses must be met which assure that, on the one hand, the task taps into the implicit linguistic knowledge of the speaker (Ellis, 2006), and, on the other, that the responses are minimally affected by rote memory (Erlam, 2006). It may be critical, for example, to include specific target structures in the EI stimulus sentences. It may be important, as Erlam (2006) claims, to focus the participant's attention on meaning with each stimulus sentence, to delay the imitation of stimuli in order to reduce the chances of rote repetition, and to assure that stimulus sentences are repeated under time pressure to simulate conditions of unplanned speech and reduce the likelihood of the examinee's focusing on the form of the sentence. It may also be crucial to show that the ability of subjects to produce particular forms using the elicitation procedure corresponds to their ability to produce these forms in spontaneous speech.

But none of these conditions have been shown to contribute to the concurrent validity of EI measures of second language proficiency. Perhaps this is because those who have created such measures have not experimented enough with various forms of stimulus presentation. However, up until now, for the creator of EI language proficiency measures, the major strategy appears to have been to try large numbers of items differing in length, morphological complexity, and syntactic complexity, on large numbers of subjects of varying language backgrounds and proficiency levels, and to empirically test which items elicit consistent responses and ultimately discriminate between learners whose proficiency levels in the language have been verified by other measures (Graham, Lonsdale, Kennington, Johnson, & McGhee, 2008).

Thus, concerning our ability to explain how the imitation of utterances provides a reasonable representation of interlanguage, Bley-Vroman and Chaudron (1994) observe, "We regard it as premature to view elicited imitation as a proven method for inferring learner competence, because a considerable amount of research needs to be conducted to understand how performance under

 Selected Proceedings of the 2008 Second Language Research Forum, ed. Matthew T. Prior et al., 57-72. Somerville, MA: Cascadilla Proceedings Project.

imitation conditions compares with other methods and with learners' underlying knowledge" (p. 245). However, with reference to the psychometric use of EI they claim that, "The more you know of a foreign language, the better you can imitate the sentences of the language. Thus, EI is a reasonable measure of global proficiency" (p. 247). In a sense, the use of EI in psycholinguistic research is foundational to establishing the construct validity of its use as a measure of oral language proficiency, but not necessarily of its concurrent validity. It is possible that even without understanding the precise mechanism by which EI works, it may prove to be a reliable measure of oral language proficiency (Vinther, 2002). This seems to be the thrust of the Chaudron et al.'s (2005) claims that "The methodology and results of elicited imitation (EI) tests developed for Vietnamese and Indonesian oral proficiency assessment serve as models for other languages" (p. 1 of handout), since the EI procedure has proven to be a valid and reliable measure of oral proficiency. While we feel that much more work must be done to establish the validity of the EI procedure as a measure of oral language proficiency, we accept the premises of Chaudron et al.'s remarks and will use them as a building block for the current study.

We have identified three major areas of research which we feel need to be conducted regarding the refining of the EI procedure as a measure of language proficiency: (a) test administration, (b) scoring methods, and (c) item design. Little systematic work has been done with test administration to examine the effects of procedures such as those explored by Erlam (2006) on language proficiency measurement. For example, will interjecting a comprehension task between the stimulus sentence and the subject's response improve the degree to which EI results will predict outcomes on established measures of oral proficiency? What about different lengths of pauses between the stimulus and the response? What is the effect of item timing on student responses? What effect does order of presentation of items have on student responses?

Pertaining to scoring methods, little research has been done to date on the differential effects of ways of scoring items on test outcomes. In psycholinguistic research, it is common to score items on the basis of the presence or absence of particular syntactic or morphological features. In language testing literature, generally points are given for correctly produced words or syllables (see review of this issue under section 3.3). How does the number of points given per item affect outcomes? Should longer items be assigned more points? What other weighting factors should be taken into consideration?

The third area, item design, is the focus of this article. Specifically we examine the effects of lexical difficulty on the reproducibility of EI items by language learners. For purposes of this study, we have chosen to examine three aspects of lexical difficulty: lexical frequency, lexical density, and morphological complexity (Gardner, 2007; Nation & Arevart, 1991).

2. The EI process

There are various accounts of the mechanisms and processes underlying EI which attempt to explain how it relates to oral language competence and proficiency. Bley-Vroman and Chaudron's (1994, p. 247) account involves the following:

- The speech comprehension system: The subject hears the input and processes it, forming a representation.
- Representation: The resulting representation includes information at various levels.
- Memory: The representation must be kept in short-term memory.
- The speech production system: The subject formulates a sentence based on the accessed representation. (There may also be monitoring of the phonetic plan, comparing it to the model.)

Given that short-term working memory is limited, the retention of a representation there is, by most accounts, dependent upon the number of units being retained (Cowan, 2001; Miller, 1956) and the structure of associated long-term memory (Ericsson & Kintsch, 1995). On average young adults are able to recall in serial order between five and nine random units, whether they be letters, syllables,

digits, or words. However, as individuals begin to get exposure to predictable sequences of these units, their associated long term memory enables them to process larger numbers of units in succession.

In language acquisition, both native and second language, natural exposure to a language results in the learning of sequences of items called constructions (Clark, 2005; N. Ellis, 2005). These can be sequences of sounds, morphemes, or words. As these sequences are recorded in long-term memory, they form units which are more or less tightly associated with each other. Thus collocations and lexical phrases are formed along with grammatical patterns. These larger units facilitate the processing of longer and more complex sequences.

When sentences are presented for imitation, the learners' are able to process short utterances in a language with which they have little familiarity from rote memory. But as the length of stimulus utterances becomes greater, it necessitates the chunking of information into successively larger units in order that the representation may be retained in working memory until it is repeated. The associations formed in long-term memory of sequences of units established during acquisition are what is believed to facilitate this chunking process (N. Ellis, 2005). The more proficient the speaker is in the language in question, the more efficient are the automatic formulations of representations and the more accurate the reconstruction of utterances. Bley-Vroman and Chaudron (1994) state it this way, "We submit that the connection between subjective length and learner proficiency and the connection between subjective length and memory capacity account for the correlation of imitation accuracy with measures of language proficiency" (p. 251).

2.1. Sentence length in the design of EI instruments

Bley-Vroman and Chaudron (1994) make the following claims about sentence length: (a) "because memory limitations are crucially involved, we expect accuracy when length is short" (p. 252); (b) "As length increases, accuracy will remain good until the limits of memory are approached. Then accuracy should fall quickly and remain…" (p. 252); (c) "Around the limits of memory, there should be a narrow band of sensitivity, where accuracy might be affected by details of the syntactic structure." (p. 252). In a review of the literature on memory span functions, they make a convincing argument for the primacy of sentence length over sentence complexity as the primary measure of L2 proficiency. However, as mentioned above, the choice of structures can make a difference in the ability to imitate within the "narrow band of sensitivity." Perkins, Brutten and Angelis (1986) and Hendrickson, Aitken, McGhee, and Johnson (this volume) give empirical support to the primacy of stimulus length in determining item difficulty.

For adult learners Hudgins and Cullinan (1978) found, working with 40 graduate student native speakers of English, that they had difficulty consistently imitating sentences longer than twenty syllables. Vinther (2002) drew similar conclusions after reviewing the literature on EI with both native speakers and second language learners. Bailey, Eisenstein, and Madden (1976), studying the development of Wh-questions in adult L2 learners and Munnich, Flynn, and Martohardjono (1994) studying adult native Japanese speakers with advanced ESL proficiency both used sentences of 15 syllables in length. On the other hand Perkins et al. (1986) found that some adult ESL speakers could only repeat sentences of seven to eight syllables in length. So the general range of sentence length found in the literature for adult testing is between six syllables and nineteen syllables.

2.2. Sentence complexity in the design of EI instruments

A great deal of the research using EI has focused on eliciting particular structures and comparing the results with other means of elicitation such as sentence completion and interviews. The results of these investigations have been mixed, with some researchers claiming that EI provides an accurate picture of learners' interlanguage (Gallimore & Tharp, 1981; Munnich et al., 1994) and others (Connell & Myles-Zithzer, 1982; Fujiki & Brinton, 1987) finding significant discrepancies between structures produced in spontaneous speech and those elicited through imitation.

With regard to the measurement of second language proficiency, Bley-Vroman and Chaudron (1994) argue that controlling for sentence complexity is of lesser importance. Given the confounding of sentence complexity with sentence length, and given the potential effects of factors such as

vocabulary choice and even issues in pronunciation, they claim that "a broad sampling of stimuli of various lengths and complexities should provide a reasonably good assessment of global proficiency (p. 252).

2.3. Lexical difficulty in the design of EI items

While lexical difficulty has been mentioned along with syntactic complexity as a factor in item difficulty in EI tests (e.g., Bley-Vroman & Chaudron, 1994), we are not aware of any attempt in the literature to isolate the effects of lexical difficulty and to determine how this contributes to item difficulty in an EI measure of language proficiency. In fact, little attempt has been made in the research literature to evaluate individual items in terms of their contribution to the overall effectiveness of EI instruments.

In the literature, lexical difficulty has been defined variously. We will examine three factors which may contribute to the degree to which lexical choice affects the difficulty of items in an EI instrument, lexical frequency, lexical density, and morphological complexity.

2.3.1. Lexical frequency

Lexical difficulty is commonly discussed in the literature on vocabulary learning and teaching, particularly when dealing with reading and listening comprehension. Nation (2001) and Meara and Huw (2001), along with others, have claimed that the frequency of occurrence of lexical items contributes to the likelihood that they will have been learned by a particular individual and thus play a role in overall difficulty in comprehending a giving passage. Nation attempts to account for the "learning burden" of words by grouping them according to word families, that is, by the root forms along with their inflectional and some of their transparent derivational affixes. He then performs frequency counts of these word families as they appear in the British National Corpus (BNC) and lists them in order of frequency. Meara and Huw, on the other hand categorizes words as 0 (function words), 1 (easy words) or 2 (hard words) based on their frequency of occurrence.

In more recent work, Gardner (2007) claims that form–based methods of determining frequency of occurrence may underestimate the real learning burden of certain words. He claims that word meaning must be taken into account in determining frequency. However, his suggested approach to determining frequency is impractical at this point because of the lack of mechanized ways of marking items for meaning. Kilgarriff (1997) has generated a lemmatized frequency list using the BNC which we view as a middle ground between Nation's approach and Gardner's.

By middle ground, we mean that Nation's word families likely underestimate the learning burden of words by combining lemmas whose meanings and uses are not likely transparent to a second language learner. For example using his word count, 'act,' 'action,' 'activity,' and, 'actress,' would all be counted as members of the same word family and their individual frequencies would be aggregated under the word 'act.' In Kilgarriff's list, 'act' the noun would be the 455th most frequent word, 'act' the verb would be the 654th, 'action' would be number 371, 'activity' would be number 440, and 'actress' would be number 5,376.

Under Gardner's proposal, the noun 'act' would be divided into at least five separate lemmas including 'an enactment,' 'a deed,' 'a subdivision of an opera,' 'a manifestation of insincerity,' 'a short theatrical performance.' The verb would be divided into as many as ten separate lemmas, and the other members of the Nation's word family would be separate lemmas with their own subdivisions. Not only that, but the word 'act' used in lexical phrases (for example, in phrasal verbs such as 'act up,' 'act out,' and 'act on') and compound words including the word act would be considered separately. Such a system, while certainly being more reflective of the actual linguistic competence of English speakers, would be very impractical in a study such as ours.

Consequently we use Kilgarriff's list as the measure of word frequency in our study. No comparable list exists for the American Corpus which makes the use of the BNC more practical if not as geographically appropriate for our purposes.

2.3.2. Lexical density

While frequency is an indirect measure of individual lexical item difficulty, one can estimate sentence level processing difficulty by a measure called lexical density (Halliday 1985; Schmitt & McCarthy, 1997; Ure, 1971). Lexical density is the ratio of content words to the total words in the sentence. Biber and his colleagues (Biber, Johansson, Leech, Conrad, & Finegan, 1999) have shown through corpus analysis that spoken language has a much lower lexical density index than do newspaper or academic texts. They interpret this to indicate a much less dense packing of information in conversation and a lesser information load.

In measures of readability, lexical density has traditionally been used to predict text difficulty. Typically, the more lexically dense the text, the more difficult the reading of the texts is judged to be. In theory, the more lexically dense a sentence, the greater the demands on memory and thus the greater the difficulty in processing (Halliday, 1985).

2.3.3. Morphological complexity

The inflectional system of English is relatively simple with only a handful of obligatory morphemes. However, the derivational morphological system is considerably more complex. It has been shown that the more complex the word is in terms of morphological components, the more demands are placed on memory for processing the word (Feldman, 1995). Thus, we define morphological complexity as the average number of morphemes per lexical item.

2.4. Research questions

In this study we will examine the degree to which each of these measures of lexical difficulty affects item difficulty in an EI test. Specifically we examine the following three questions:

1. To what extent does lexical frequency as operationalized by frequency in Kilgarriff's lemmatized list affect the difficulty of EI items?
2. To what extent does lexical density, defined as the ratio of content words to total words in a sentence, affect the difficulty of EI items?
3. To what extent does morphological complexity, defined as the average number of morphemes per lexical item, affect the difficulty of EI items?

3. Method

3.1. Instrument

The EI instrument used in this study consisted of 60 sentences, five sets of twelve sentences each, varying in length from four syllables to nineteen syllables (see Table 1). The frequency levels for the lexical items in each sentence were chosen from Kilgarriff's lemmatized list. On this list, the lemmas are organized according to frequency with the most frequent one appearing as the first on the list and the least frequent item appearing as item number 6,318. In each set of twelve sentences, length was kept constant while vocabulary frequency varied from frequent to infrequent. In order to maximize the separation of frequencies between sets of items the frequency ranges were established at: 400-600, 1400-1600, 2400-2600, 3400-3600, 4400-4600 and 5400-5600. Thus all the content words for a given item were chosen from the specified frequency range.

In creating the items, members of the research team generated sentences using lexical items from each frequency range while avoiding late acquired inflectional morphology and syntactic forms. By this we mean that all of the lexical items for each sentence were chosen from the 200 words within that narrow range of frequencies and that all sentences were kept in the simple present or progressive with no third person singular lexical verb conjugations and no negatives, interrogatives, relative clauses, embedded questions, passives and the like. Sentence length was increased by increasing the length of noun and verb phrases, and by including prepositional phrases, verb complements, and adverbs (see Appendix for a complete list of all items).

After creating a pool of items based on sentence length and lexical frequency, we examined the items for morphological complexity and lexical density and selected items that varied along these two variables.

Table 1. *Number of Items Per Frequency Range Per Item Length in Syllables*

Frequency Range	Sentence Length in Syllables				
	4-6	7-9	10-12	13-15	16-19
5400-5600	2	2	2	2	2
4400-4600	2	2	2	2	2
3400-3600	2	2	2	2	2
2400-2600	2	2	2	2	2
1400-1600	2	2	2	2	2
400-600	2	2	2	2	2

Thus, we did not systematically vary the morphological complexity and lexical density of items in the same way we did sentence length and lexical frequency because doing so would have made the process of generating plausible sentences almost impossible. However, average morphological complexity for the selected sentences ranged between 1.00 and 2.00 with a standard deviation of 0.27 and lexical density varied between 0.40 and 1.00 with a standard deviation of 0.124. Thus the process of creating sentences for the EI instrument followed the following criteria in order of importance:

1. Sentences had to be comprehensible and plausible.
2. The criterion of sentence length in syllables had to be followed.
3. The criterion of selecting items of a given lexical frequency had to be followed.
4. Similar inflectional morphology had to be used across all sentences.
5. Morphological complexity of lexical items had to vary.
6. Lexical density of items had to vary.

After creating a pool of items, members of the research team reviewed the items for comprehensibility and naturalness based on their intuitions as native speakers. They selected the most promising items for trials with a cadre of native speakers. Items were selected, eliminated or revised based on the ability of native speakers to understand and repeat them correctly. Once the final selection of items had been made audio recordings of the items were made in a studio by professional male and female voice talent. In the final test administration half of the items were presented by a male voice and half by a female voice. Examples of sentences created with lemmas from the 400-600 frequency range can be seen below.

4-6 Syllables
I understand language.
We are stopping soon.

7-9 Syllables
The students are learning a game.
They are building several churches

10-12 Syllables
The city is developing a central plan.
The patients describe the condition of their health.

13-15 Syllables
The manager is watching the difficult situation.
Half of the teams are sending boys to practice at the field.

16-19 Syllables
The committees usually agree to support raising taxes.
The teachers hope to return with many special experiences.

3.2. Participants and procedures

Participants for this test were 81 learners of English in an intensive English program in the U.S. They were from a variety of L1 backgrounds, including Spanish, Portuguese, Korean, Japanese, and Chinese and a wide range of proficiency levels from novice to advanced levels (see Table 2). These proficiency levels represent locally determined placement levels and were assigned based on the results from a battery of tests including a reading test, a writing test, a listening test, a grammar test, and a oral proficiency interview.

Table 2. *Number of Subjects at Each Proficiency Level*

Proficiency Level	Number of Subjects
Low beginning	7
High beginning	13
Low intermediate	22
High intermediate	22
Advanced low	17
Total	81

Procedures for the administration of the EI instrument were the following. The students entered a computer lab and were oriented to the testing area and computer. They logged on and were presented with a screen which invited them to participate and obtained their informed consent. Following this, audio and video instructions were presented describing the test, telling them that they would hear the sentence only once, and instructing them to repeat items verbatim as nearly as they could. They were then presented on screen with a demonstration of an item with a correct response. Following this, they were presented with an item to which they responded. If they had difficulty performing the task they were asked to raise their hand for assistance. If not, they proceeded on with the test. Items were then presented to the learners one at a time in random order via headphones with high quality sound equipment and they recorded their responses using a microphone attached to the head set. So, for example, for each item they saw on the screen a text that said "Sentence #," they then heard the sentence read by the male or female voice, followed by a beep signaling to begin repeating the sentence. A time bar then appeared on the screen showing the amount of time left to repeat the sentence. The time allotted to repeat sentences varied between six second for the short sentences and 12 seconds for the longest sentences. Thus each participant received the stimulus sentences in a different random order. Once recorded, the files were saved as wave files for later analysis.

3.3. Scoring

There is no standardized method in the literature for scoring EI items (Vinther, 2002). For those interested in determining whether learners control specific morphological or syntactic features of the language, the method of scoring usually takes the form of examining each repeated sentence for the presence or absence of the target features while ignoring other inaccuracies which may have occurred in the repetitions (Erlam 2006; Munnich et al., 1994). For those attempting to develop an indirect method of estimating global language proficiency, items are generally scored on a scale of correctness ranging from a two point scale (Henning, 1983), to a three point scale (Radloff, 1992), to a five point scale (Chaudron, et al., 2005), to a seven point scale (Keller-Cohen, 1981). Lonsdale, Dewey, McGhee, Johnson, and Hendrickson (2009) experimented with a variable scale in which each syllable was awarded one point for being correct or zero points for being incorrect or absent. Differences in correlations with oral proficiency interviews, between tests scored by assigning one point for each

syllable produced correctly and those scored by our four-point scale method were small and inconsistent (Lonsdale et al., 2009). Very little empirical research has been done to determine the effects of scoring method on the concurrent validity of EI instruments. We have chosen to use a modified version of the Chaudron, et al. (2005) method because it provided a relatively objective way of scoring items and it promises to allow for eventual machine scoring of EI items (Graham et al., 2008).

For scoring of items in this study, recorded files from the server for the entire group of participants were placed in a database. Raters accessed randomly selected sentences one at a time from the database and judged the correctness of the sentence. The written form of the stimulus sentence, divided into syllables, was presented on a computer screen while the audio recording of the student's response was presented through earphones. The rater's job was to listen to the sentence and mark with a "1" or "0" whether each syllable was correct. If it was present and intelligible, the rater was to mark the syllable with a "1." If the syllable was absent or unintelligible or if a different word or syllable was pronounced in its place by the participant, the rater was to mark it with a "0." If a participant failed to include a syllable or syllables in the first attempt but corrected himself/herself, the "0" for that syllable was replaced with a "1." Raters were trained to deal with all sorts of cases of omissions, substitutions, mispronunciations, etc. Two raters independently judged each sentence. The raw agreement by syllable between raters was .91 and interrater reliability (Kohen's κ) was .82.

Once each syllable was judged present or not, a score was determined for the entire sentence. The scoring method for each item was a modified version of that proposed in Chaudron et al. (2005). Each sentence was awarded a maximum score of four if every syllable had been judged to be a "1." For each syllable not judged to be correct a point was subtracted from the maximum score until the score for that sentenced reached zero. Based on these individual sentence scores, total scores were awarded for each participant. See Figure 1 for examples of actual scores awarded sentences from a previous test:

1	1	1	1	1	1	1	1	1	
If	she	lis	tens,	she	will	un	der	stand.	Score = 4
1	0	1	1	1	1	1			
Why	had	they	liked	peas	so	much?			Score = 3
1	1	0	1	1	1	1			
Big	ships	will (are)	al	ways	make	noise.			Score = 3
0	1	0	1	0	1	1	0	1	
We (They)	should	have	ea (eat)	ten	break	fast	by (right)	now.	Score = 0

Figure 1. Scoring of sample EI sentences using Chaudron et al.'s (2005) four score method.

3.4. Analyses

Following the scoring of items, all scores were placed in a database for access by statistical programs. The first type of analysis to be performed was an Item Response Theory (IRT) analysis designed to determine overall item quality and item difficulty for each of the 60 sentences. To accomplish this analysis, a one parameter Rasch-model was used. The computer program employed for this purpose was the Winsteps Ministep program (Linacre, 2006). IRT was chosen instead of a classical test theory analysis, even though the number of subjects in this study was small, for two main reasons. First, traditional item difficulty measures are test-dependent, and since we are interested in calibrating item difficulties on this instrument with those of other forms of the test that we are using, we want a statistic which is test independent. Second, traditional statistical indicators of item difficulty are group-dependent, that is the values vary as a function of the group who took the test (Bond & Fox, 2007). We are continuing to administer this and other forms of the EI test to large numbers of subjects and we want to be able to compare items across groups.

In order to determine the effects of lexical frequency, lexical density, and morphological complexity on item difficulty, several analyses were performed. These variables were defined as follows: lexical frequency--the frequency of occurrence of the item on Kilgarriff's (1997) lemmtized frequency list; lexical density--the ratio of content words to the total words in the sentence; morphological complexity--the average number of morphemes per lexical item. Because of the evidence in the literature for the importance of sentence length in determine item difficulty, we included sentence length as a moderator variable. Sentence length was defined as the number of syllables in the sentence.

Following the IRT analysis correlations were performed using item difficulty scores and average item scores as dependent variables. Then a multiple linear regression was performed with average score as the dependent variable. Following this, a step-wise regression analysis was performed, using the same dependent variable to determine the percent of variance accounted for by each of the independent variables.

4. Results

Table 3 presents the descriptive statistics and reliability statistic for the 60 items on the EI test as revealed by the IRT analysis. Notice that the internal consistency of the items on the test produced a Cronbach Alpha coefficient of .94.

This high item reliability is typical of reliabilities observed on other forms of the EI tests that we have been using (Graham, 2006) as well as of those reported in the literature (Chaudron et al., 2005). Notice also that the range of raw scores on items goes from a low of 0, meaning that no one got the item correct, to a high of 180, indicating that almost everyone got some credit on the item.

Table 3. *Descriptive Statistics, Reliability for 60 Items and 81 Subjects*

	Raw score	Measure	Model error
Mean	93.40	44.30	1.55
SD	33.50	7.61	1.80
Max.	180.00	60.29	17.44
Min.	0.00	0.57	1.27

Note. Person raw score-to-measure correlation = .92. Cronbach *Alpha* (KR-20) Person raw score reliability = .94

The Person/Item map presented in Figure 2 shows the distribution of subject scores on the left and item difficulty scores on the right. Scores are mapped from easiest on the bottom to most difficult on the top and subjects are arranged from least skilled on the bottom to most skilled on the top. Items are mapped on a scale with a mean of 50 and a standard deviation of 10. Notice that the match between item difficulty and subject ability is pretty good with the mean ability of subjects (represented by the *M* to the left of the axis) being only slightly below the average difficulty of items (represented by the *M* on the right side of the axis). Also, the distribution of scores suggests that there was no ceiling effect on the test.

Table 4. *Pearson Correlation Coefficients for Item Variables (N = 60)*

	ItemDiff	Average	LexFrq	SenLeng	LexDen	MrphCom
ItemDiff	1.00	-0.96	0.25	0.82	0.24	0.36
AveScore	0.96	1.00	-0.25	-0.85	-0.18	-0.33
LexFreq	0.25	-0.25	1.00	-0.04	-0.04	0.03
SenLeng	0.82	-0.85	-0.04	1.00	0.04	0.35
LexDen	0.25	-0.18	-0.04	0.04	1.00	0.20
MrphCom	0.36	-0.33	0.03	0.36	0.20	1.00

Note. ItemDiff = Item difficulty as calculated in the IRT analysis, AveScore = Average raw score for each item, LexFreq=Lexical Frequency scores, SenLeng = Length of sentence in syllables, LexDen = Lexical items/total words in the sentence, MorphCom = average number of morphemes per lexical item in the sentence.

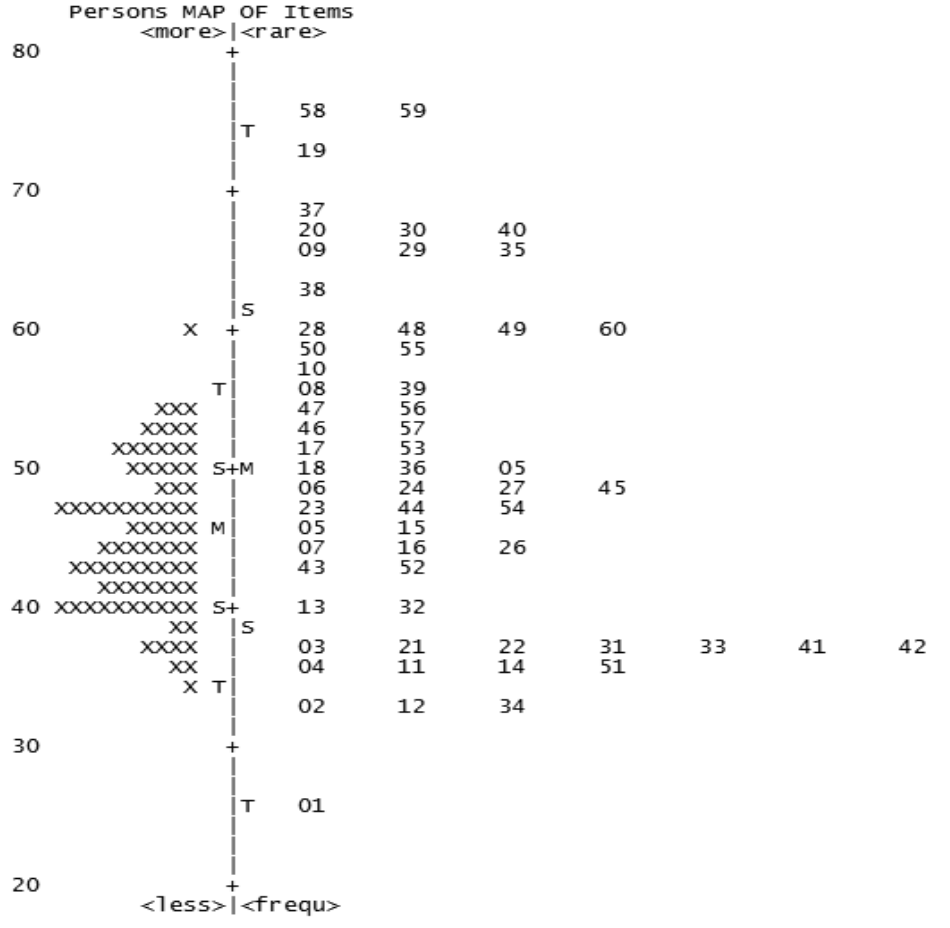

Figure 2. IRT Pearson-item map analysis of EI items.

Following the IRT analysis a multiple regression analysis was performed. Table 4 presents the correlations among variables in the study. It is noteworthy that item difficulty as calculated by the IRT analysis correlates at -.96 with the average scores for all 60 items on the test. This suggests that the average score for items is a fairly good measure of item difficulty. In our regression analyses we will use the average scores as the dependent variable and sentence length, lexical frequency, and lexical density as the independent variables.

Using the enter method, a significant model emerged ($F(4,55) = 68.253$, $p < .0001$). Adjusted R square was .820. Statistically significant variables are shown below:

Predictor Variable	*Beta*	*p*
Sentence Length	-.864	$p < .0001$
Lexical Frequency	-.290	$p < .0001$
Lexical Density	-.160	$p < .007$

(Morphological Complexity was not a significant predictor in this model.)

Next, a step-wise regression with average score as the dependent variable and sentence length, lexical frequency, and lexical density as the independent variables was performed. Table 5 presents the results of this analysis. Notice that morphological complexity was omitted from this analysis because its contribution to item difficulty was negligible. Whatever variance it may have accounted for must be overlapping with that of the other variables. You will also recall that an attempt was made to control for the effects of grammatical complexity by designing the sentences to include only certain forms.

Table 5. *Summary of Stepwise Regression*

Step	Variable	Partial R Square	R Square	$C(p)$	F
1	Sentence length	.728	.728	33.288	155.07
2	Frequency	.080	.807	9.064	23.70
3	Lexical density	.051	.859	3.097	8.10

Notice that sentence length accounts for almost 73% of the variance in the scores on the EI items in our test. This is in line with expectations given our earlier discussion. Lexical frequency accounted for an additional 8% of the variance in item scores and lexical density added only another 2%. So the model accounts for a little more than 83% of the total variance in item difficulty. Table 6 gives a statistical summary for the model with sentence length, frequency and lexical density as variables.

Table 6. *Full Model Summary of Stepwise Regression*

Variable	*df*	Parameter Estimate	Standard Error	*t*-value	*p*r < \t\
Intercept	1	5.893	0.392	15.03	< .0001
Frequency	1	-0.021	0.004	-5.28	< .0001
Sentence length	1	-0.257	0.016	-15.63	< .0001
Lexical density	1	-1.577	0.554	-2.85	.0062

Note. Overall model: $F(3, 56) = 92.46$, $p < .0001$

In order to examine the relative effects of sentence length and lexical frequency on the difficulty of items, we have created Figure 3, which presents the average score for each of the sixty items. Figure 3 is organized by sentence length in number of syllables and by lexical frequency levels in 100s (Remember that the most frequent lexical items have the lower numbers.). The horizontal axis represents sentence length in number of syllables from shortest to longest. On the vertical axis are the lexical frequency rankings with items with the most frequently occurring vocabulary at the bottom and the least frequent vocabulary at the top. Cells are shaded according to the average item score with darker shades representing easier items and successively lighter shades representing more difficult items. The maximum score possible on a given item as explained earlier is 4.00.

The item with the highest average score (that is, the easiest item) is represented in the bottom left hand corner of the table with a score of 3.75. Notice that this item is from 4-6 syllables long with lexical items chosen from the 400 to 600 frequency range. If both sentence length and lexical frequency are having an effect, one would predict that the easiest items would be the shortest sentences with the most frequent vocabulary. This turns out to be the case. Likewise the most difficult items should appear in the upper right hand corner, that is, they should be the longest sentences with the least

frequent vocabulary. This pattern is evident up to sentence lengths of 15 syllables. In fact there is a pretty even distribution of difficulty scores from the darkly shaded items in the bottom left to the white items in the upper right corner with only a few items, which fall outside this pattern. This suggests that both sentence length and lexical frequency are having a fairly consistent effect on item difficulty at least until item length reaches 16 syllables. For items of sixteen syllables and longer, lexical frequency seems not to be having a consistent effect.

Item Length in Syllables

Frequency Level	4-6	4-6	7-9	7-9	10-12	10-12	13-15	13-15	16>	16>
5400-5600	3.28	2.52	1.00	1.54	0.69	0.30	0.84	0.02	0.02	0.23
4400-4600	3.20	3.11	2.46	1.68	1.40	0.86	0.65	0.26	0.25	0.32
3400-3600	3.10	2.81	3.46	3.19	1.20	0.10	0.06	0.15	0.52	0.09
2400-2600	3.12	3.12	1.63	1.41	2.25	1.28	1.43	0.25	0.10	0.07
1400-1600	3.53	3.26	3.28	2.73	2.00	2.20	1.16	0.91	0.04	0.07
400-600	3.75	3.49	3.14	3.35	1.80	1.36	2.21	0.54	0.11	0.44

Figure 3. Average scores of items by sentence length and lexical frequency.

5. Discussion and Conclusions

The purpose of this study was to examine the effects of lexical difficulty on the design of items in an elicited imitation test. Lexical difficulty was operationalized as a combination of three variables, lexical frequency as determined by frequency of occurrence on Kilgarriff's (1997) lemmtized frequency list, lexical density (the ratio of content words to the total words in the sentence), and morphological complexity (the average number of morphemes per lexical item). Sentence length, defined as the number of syllables in the sentence, was also included as a moderating variable in the study because of the effect it has had on item difficulty in previous studies (Bley-Vroman & Chaudron, 1994; Perkins et al., 1986). Finally, syntactic complexity, instead of being calculated for each item, was controlled for by designing sentences with similar inflectional morphology and similar syntactic features.

Results showed that lexical frequency did make a significant contribution to item difficulty, accounting for approximately eight percent of the variance in difficulty. This effect seemed rather constant across all levels of sentence length except for those exceeding 15 syllables (see Figure 3). That is, even for sentences of four to six syllables in length, item difficulty was increased by including less frequent vocabulary. However, as sentence length reached 16 syllables, lexical frequency appeared to have no consistent influence on item difficulty. This may be a result of the fact that the most proficient speakers in our sample had proficiencies approximately equivalent to that of "advanced" on the ACTFL Oral Proficiency Interview (OPI) scale.

It is clear from the Person-Item map in Figure 2 that the most difficult items on the test exceeded the level of ability of the persons taking the test. Perhaps if we had students at the superior level of oral proficiency, differences in difficulty among the longer sentences could have been detected. This is

consistent with the theory of the effects of sentence length on item difficulty, which will be discussed later.

Results also showed that lexical density made a significant contribution to item difficulty, accounting for approximately 2% of the variance. While this finding is consistent with the theory of information load and language processing (O'Loughlin, 1995), the effects of lexical density may be mitigated by the effects of increased ease of processing of memorized chunks in language use (N. Ellis, 2005). As mentioned earlier, normal language acquisition entails the repeated processing of linguistic elements which form units themselves. These formulaic constructions facilitate processing and lessen the cognitive burden occasioned by the greater lexical density. This may account for the small amount of variance accounted for by this variable.

The third variable, morphological complexity was not a significant predictor of item difficulty in this study. On its face, this finding seems counter intuitive. While we have not found any studies which examine the effects of increased derivational morphology in lexical items on processing difficulty, it seems likely that, all other things being equal, greater morphological complexity should lead to greater difficulty in interpreting and encoding lexical items and thus to EI sentences. However, as Table 4, shows, morphological complexity has a rather high correlation with sentence length. This colinearity may have accounted for the fact that this variable did not contribute independently to the prediction of item difficulty.

Finally, the moderating variable, sentence length, turned out to be highly significant as we expected, accounting for nearly 73% of the variance in item difficulty. This finding provides support for the model proposed by Bley-Vroman and Chaudron (1994). According to their account of the EI process, "we expect accuracy when length is short. As length increases, accuracy will fall rather quickly and remain low. Around the limits of memory, there should be a narrow band of sensitivity, where accuracy might be affected by details of the syntactic structure" (p. 252).

This is precisely the pattern observed in this study. For learners of low proficiency, imitating sentences of moderate length, no matter the lexical difficulty of the items in the sentence, was beyond their capabilities. For learners of greater proficiency, production of sentences of greater length became possible and lexical difficulty became a factor in their productions. At sentence lengths of 16 syllables and above, none of the participants in the study were proficient enough for lexical difficulty to make a consistent difference. Given that native speakers could produce all of the sentences in the study with great accuracy in the design phase of the study, it is probable that the lack of effect of lexical difficulty in sentences greater than 16 syllables was a result of the lower proficiency of the learners. The most proficient subjects in the study were at about "advanced low" to "advanced mid" proficiency on an ACTFL OPI scale.

In spite of the overwhelming effects of sentence length on item difficulty, this study has shown that lexical difficulty needs to be taken into account when creating sentences for EI instruments. The effects of item frequency can be felt by subjects as sentence length challenges their working memory limits. Admittedly it is a bit artificial to process sentences in which all of the lexical items are from a given frequency range, creating such sentences was a way for controlling for word frequency. The next step in the process of research on this topic might be to take sentences which are already being used in an EI test, assign each a vector representing its frequency and use those values in an analysis to examine the effects of lexical frequency under more natural conditions.

Appendix

400-500 Frequency Band
I understand language.
We are stopping soon.
The students are learning a game.
They are building several churches.
The city is developing a central plan.
The patients describe the condition of their health.
The manager is watching the difficult situation.
Half of the teams are sending boys to practice at the field.
The committees usually agree to support raising taxes.
The teachers hope to return with many special experiences.

1400-1500 Frequency Band
The victim was famous.
The meal is lovely.
The visitors warn the farmers.
They are delivering dinner.
The author is a master of description.
The appearance of the coffee was excellent.
The official examination was extremely detailed.
The regulations contribute to unemployment.
The housing crisis is attracting concern from the administration.
The judges are launching a broad internal investigation.

2400-2500 Frequency Band
Researchers compete.
The coaches qualify.
The landlords negotiate the rent.
The advisors fire the consultants.
Producers organize genuine drama.
We are celebrating numerous birthdays.
We are retiring the electronic certificate.
The consultants are detecting widespread disaster.
You are successfully resisting administrative input.
Perceptions differ greatly about radical demonstrations.

3400-3500 Frequency Band
The rumor is painful.
The satellites float.
We march to the colony.
The separation is convenient.
The ruling rebels overlook the invasion.
They are debating the unfair proposition.
The uncertain trustees initiate the merger.
They are reluctant to hire such a diplomatic fellow.
They are translating the Bible in a magnificent cathedral.
They are aggressively recruiting reporters for the tournament.

4400-4500 Frequency Band
They price the stamps.
I grip the bow.
Prejudice is an embarrassment.
They are inspecting the cave fossils.
His girlfriend was delighted when he proposed.
They enquire about the missing brass blades.
The hunting expedition is practically ruined.
Corruption is an immense obstacle to tourism.
The neglected saints long for inspiration and revelation.
The psychologist is campaigning for racial diversity.

5400-5500 Frequency Band
We recycle balloons.
You film in the suburbs.
Donors group the outfits in bins.
Furs shrink with one washing.
The warriors pledge to invade in the daylight.
The bored waiter is straightening the stool.
The vigorous tigers are roaring in the jungle.
Their unusually innovative thinking is constructive.
Unreasonable deadlines inflict suffering on newcomers.
The accumulation of poison in the vegetation is appalling.

References

Bailey, Nathalie, Eisenstein, Miriam, & Madden, Carolyn (1976). The development of wh-questions in adult second-language learners. *On TESOL*, *76*, 350–362.

Berry, P. B. (1976). Elicited imitation of language: Some ESNS population characteristics. *Language and Speech*, *1*, 350–362.

Biber, Douglas, Johansson, Stig, Leech, Geoffrey, Conrad, Susan, & Finegan, Edward (1999). *Longman grammar of spoken and written English*. Harlow, Essex: Pearson Education.

Bley-Vroman, Robert & Chaudron, Craig (1994). Elicited imitation as a measure of second-language competence. *Research methodology in second language acquisition*, *7*, 245–261.

Bond, Trevor G. & Fox, Christine M. (2007). *Applying the Rasch model: Fundamental measurement in the human sciences* (2nd ed.). Mahway, NJ: Lawrence Erlbaum.

Chaudron, Craig, Prior, Matthew, & Kozok, Uli (2005). *Elicited imitation as an oral proficiency measure*. Paper presented at the 14th World Congress of Applied Linguistics, Madison, WI.

Clark, Eve Vivienne (2003). *First language acquisition*. Cambridge, UK: Cambridge University Press.

Connell, Phil & 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 (2001). The magical number 4 in short-term memory: A reconsideration of mental storage capacity. *Behavioral and Brain Sciences*, *24*, 87–185.

Ellis, Nick (2005). Constructions, chunking, and connectionism: The emergence of second language structure, in Catherine J. Doughty & Michael H. Long (Eds.), *The handbook of second language acquisition* (pp. 63–103). Malden, MA: Blackwell.

Ellis, Rod (2005). Measuring implicit and explicit knowledge of a second language: a psychometric study. *Studies in Second Language Acquisition*, *27*, 141–172.

Ellis, Rod (2006). Modeling learning difficulty and second language proficiency: The differential contributions of implicit and explicit knowledge. *Applied Linguistics*, *27*(3), 431–463.

Ericsson, K. Anders & Walter, Kintsch (1995). Long-term working memory, *Psychological Review*, *102*(2), 211–245.

Erlam, Rosemary (2006). Elicited imitation as a measure of l2 implicit knowledge: An empirical validation study. *Applied Linguistics*, *27*, 464–491.

Ervin-Tripp, Susan M. (2001). Imitation and structural change in children's language. *New directions in the study of language*, *16*, 163–189.

Feldman, Laurie, B. (1995). Morphological aspects of language processing, Hillsdale, NJ: Lawrence Erlbaum Associates.

Fujiki, Martin & Brinton, Bonnie (1987). Elicited imitation revisited: A comparison with spontaneous language production. *Language, Speech and Hearing Services in the Schools*, *18*, 301–311.

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

Gardner, Dee (2007). Validating the construct of word in applied corpus-based vocabulary research: A critical survey. *Applied Linguistics*, *28*, 241–265.

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

Graham, C. Ray, Lonsdale, Deryle, Kennington, Casey, Johnson, Aaron, & McGhee, Jeremiah (2008). Elicited imitation as an oral proficiency measure with ASR scoring. *Proceedings of LREC 2008*, 1604–1610.

Halliday, Michael A. K. (1985). *Spoken and written language*. New York: Oxford University Press.

Hamayan, Else, Saegert, Joel, & Larudee, Paul (1977). Elicited imitation in second language learners. *Language and Speech, 20*, 86–97.

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

Hudgins, Jo C. & Cullinan, Walter (1978). Effects of sentence structure on sentence elicited imitation responses. *Journal of Speech and Hearing Research*, *21*, 809–819.

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

Kilgarriff, Adam (1997). Putting frequencies in the dictionary. *International Journal of Lexicography*, *10*, 135–155.

Jessop, Lorena, Suzuki, Wataru, & Tomita, Yasuyo (2007). Elicited imitation in second language acquisition research, *The Canadian Modern Language Review*, *64*(1), 215–220.

Lahey, Margaret, Launer, Patricia, & Schiff-Myers, Naomi (1983). Prediction of production: elicited imitation and spontaneous speech productions of language disorded children. *Applied Psycholinguistics*, *14*, 317–343.

Linacre, John, M. (2006). A user's guide to winsteps: Ministep rasch-model computer programs. Retrieved from http://www.winsteps.com/winman/index.htm?copyright.htm

Lonsdale, Deryle, Dewey, Dan P., McGhee, Jeremiah. Johnson, Aaron, & Hendrickson, Ross (2009, March). *Methods of scoring elicited imitation items: An empirical study*. Paper presented at the annual conference of the American Association for Applied Linguistics, Denver, CO.

Lust, Barbara, Yu-Chin, Chien, & Suzanne, Flynn (1987). What children know: Methods for the study of first language acquisition. *Studies in the Acquisition of Anaphora*, *2*, 271–356.

Meara, Paul & Huw, Bell (2001). P lex: A simple and effective way of describing the lexical characteristics of short L2 texts. *Prospect*, *16*, 5–19.

Menyuk, Paula (1963). A preliminary evaluation of grammatical capacity in children. *Journal of Verbal Learning and Verbal Behavior*, *2*, 429–439.

Menyuk, Paula (1964). Comparison of grammar of children with functionally deviant and normal speech. *Journal of Speech and Hearing Research*, *7*, 109–121.

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

Munnich, Edward, Suzanne, Flynn, & Gita, Martohardjono (1994). Elicited imitation and grammaticality judgment tasks; what they measure and how they relate to each other. In Elaine E. Tarone, Susan M. Gass, & Andrew D. Cohen (Eds.), *Research methodology in second-language acquisition* (pp. 263–86). Hillsdale, NJ: Lawrence Erlbaum.

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

Nation, Paul, I. S. (2001). *Learning vocabulary in another language*. Cambridge: Cambridge University Press.

Nation, Paul, I. S., & Supot, Arevart (1991). Fluency improvement in a second language. *RELC*, *22*, 84–94.

O'Loughlin, Kieran (1995). Lexical density in candidate output on direct and semi-direct versions of an oral proficiency test. *Language Testing, 12*, 217–237.

Perkins, Kyle, Brutten, Sheila R., & Angelis, Paul J. (1986). Derivational complexity and item difficulty in a sentence repetition task. *Language Learning*, *36*, 125–141.

Radloff, Carla, F. (1992). Sentence repetition testing for studies of community bilingualism: an introduction. *Notes on Linguistics*, *56*, 19–25.

Schmitt, Norbert & McCarthy, Michael (Eds.). (1997). *Vocabulary: Description, acquisition and pedagogy*. Cambridge: Cambridge University Press.

Ure, Jean (1971). Lexical density and register differentiation. In George Perren & John L. M. Trim (Eds.), *Applications of linguistics* (pp. 443–452). London: Cambridge University Press.

Vinther, Thora (2002). Elicited imitation: a brief review. *International Journal of Applied Linguistics*, *12*, 54–73.

Selected Proceedings of the 2008 Second Language Research Forum: Exploring SLA Perspectives, Positions, and Practices

edited by Matthew T. Prior, Yukiko Watanabe, and Sang-Ki Lee

Cascadilla Proceedings Project Somerville, MA 2010

ISBN 978-1-57473-439-3 library binding

Ordering information

Orders for the library binding edition are handled by Cascadilla Press.
To place an order, go to www.lingref.com or contact:

Cascadilla Press, P.O. Box 440355, Somerville, MA 02144, USA
phone: 1-617-776-2370, fax: 1-617-776-2271, e-mail: sales@cascadilla.com

Web access and citation information

This entire proceedings can also be viewed on the web at www.lingref.com. Each paper has a unique document # which can be added to citations to facilitate access. The document # should not replace the full citation.

This paper can be cited as:

Graham, C. Ray, Jeremiah McGhee, and Ben Millard. 2010. The Role of Lexical Choice in Elicited Imitation Item Difficulty. In *Selected Proceedings of the 2008 Second Language Research Forum*, ed. Matthew T. Prior et al., 57-72. Somerville, MA: Cascadilla Proceedings Project. www.lingref.com, document #2385.