What Makes an Item Difficult? A Syntactic, Lexical, and Morphological Study of Elicited Imitation Test Items

Ross Hendrickson, Meghan Aitken, Jeremiah McGhee and Aaron Johnson
Brigham Young University

1. Introduction

Elicited imitation (EI) is an oral language testing method that has been experimented with since the 1960s. Its first uses were focused on assessing first language acquisition in children (Ervin-Tripp, 2001) and diagnosing language abnormalities (Berry, 1976; Lahey, Launer, & Schiff-Myers, 1983). Later, in the 1970s, researchers began to assess how useful the method would be within the field of second language acquisition (Naiman, 1974). Since then, there have been two main focuses of Elicited Imitation studies within second language acquisition (SLA): psycholinguistic research into language competence and SLA processes and research in using the method as an indirect measurement of oral language proficiency (Bley-Vroman & Chaudron, 1994; Erlam, 2006; Vinther, 2002). We focus our research on the second, measuring oral language proficiency with the use of Elicited Imitation.

Elicited imitation is a method of testing where a subject hears an utterance and then is asked to repeat the utterance they have heard. It involves both the subject's short-term and long-term memory. The theory behind using it as measurement of oral proficiency is that a student cannot successfully repeat an utterance if he cannot understand it (Bley-Vroman & Chaudron, 1994). Essentially, a subject will hear an utterance, form a representation of that utterance using his knowledge base (long-term memory) and store that representation in his short-term memory. Then the subject will form and produce a sentence based on the representation he has created. Bley-Vroman and Chaudron state in their research of the method, "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" (Bley-Vroman & Chaudron, 1994, p. 247).

After performing several experiments with EI in an effort to determine its effectiveness in testing oral language proficiency (Graham, Lonsdale, Kennington, Johnson, & McGhee, 2008), we narrowed our focus to the design of individual test items (the sentences subjects are asked to repeat). Since second language learners must depend on their knowledge base to repeat a test item, the test items must be harder for novices than for advanced learners. The difficulty of the test item depends on the complexity of the sentence. For example, a novice learner would have difficulty repeating sentences that involve content words they are unfamiliar with, but an advanced learner who had already acquired those words would be more successful. In this study we examine the effects of different syntactic, lexical, and morphological features present in the items on their difficulty. We began by taking previously administered EI tests and determining the item difficulty based on the results using an Item Response Theory (IRT) analysis. Once this was done, each sentence was tagged for syntactic, lexical, and morphological features that could possibly affect the difficulty of an item. Once the sentences were tagged, Multiple Linear Regression analyses were done to see which factors have the most impact on item difficulty. The purpose of this study is to enlighten us as to the influences of different factors on how students perform on Elicited Imitation test items. The study provides a methodological approach to deconstructing our current test items so that we may construct new items in the future with

© 2010 Ross Hendrickson, Meghan Aitken, Jeremiah McGhee, and Aaron Johnson. Selected Proceedings of the 2008 Second Language Research Forum, ed. Matthew T. Prior et al., 48-56. Somerville, MA: Cascadilla Proceedings Project.

the goal of targeting specific difficulty levels and degrees of proficiency. The test utilized for this study included 60 sentences and each sentence ranged from 3 to 33 syllables in length. The sentences were selected based on criteria as explained by Chaudron et al. (2005) and selected sentences were recorded in a studio with both male and female voices of native speakers. We administered the test to 376 learners of English at the English Learning Center in Provo, Utah whose proficiency levels ranged from Novice to Advanced. The students came from 13 different first language backgrounds, including Spanish, Portuguese, Korean, Japanese, Chinese, and Mongolian and their age ranged from 18 to 53.

2. Features

The syntactic, lexical and morphological features were chosen based on studies on the order of acquisition for second language learners and elicited imitation literature, as well as the ACTFL Oral Proficiency Interview (OPI) guidelines and handbook (Ellis, 2008; Erlam, 2006; Ortega, 2000). The OPI guidelines as well as the handbook for rating students were reviewed in order to see what features determine a student's level based on a formal OPI (see Appendix for a list of all features chosen). The studies involving order of acquisition and elicited imitation were helpful in that they could give specific features that are acquired generally in a specific order, but studies varied and it was difficult to overcome inconsistencies between experiments. These studies did, however, give us a solid list of features that we evaluated to see if they were affecting the item difficulty for elicited imitation. The OPI guidelines and handbook didn't list explicit features of sentences or words that the raters look for when determining a speaker's proficiency level. The OPI guidelines however do mention that the abilities of speakers that would correlate with specific features. An example of this is the tense and aspect of the verbs in a sentence. One of the distinguishing features between levels the OPI guidelines point out is the ability to speak within certain time frames, beginning with the Novice level where speakers tend to use memorized phrases and present tense to the Advanced High and Superior levels, where students have mastered all time frames. After reviewing second language acquisition studies on the order of acquisition in syntax and morphology and the OPI guidelines, 44 features were chosen to be tagged. This feature set is not an exhaustive set of syntactic, morphological, and lexical features, but is a representative sample, based primarily on syntax and other areas of language. All features were binary with the exception of syllable length, lexical difficulty (calculated by taking the raw frequency count for each word from the British National Corpus and dividing by the total number or words), and T-units. Below we give an example of a tagged sentence. Average score was also recorded along with the results from the IRT analysis.

An example of a tagged sentence:

Are they walking slowly because their feet are sore?

Item ID: 4
Average score: 1.4627
Measure: 54.64
Syllable count: 12
T-units: 2
Lexical: 88593.66
F10: Complex
F13: Interrogative
F17: Copular
F18: Intransitive
F26: Present tense
F30: Imperfect tense
F35: Plural

3. Data collection and annotation

The data for this study was collected at the English Learning Center in Provo, Utah as well as on the Brigham Young University campus. The students were tested at one of two testing centers at each respective campus. The testing environment was a controlled environment wherein the students all used similar computer equipment to take the examination. The audio files recorded for each student (their responses) were then uploaded to a central server and graded by a selection of linguistics students. The scores for each subject were stored in a database. Each test item was first annotated by two linguistics students with a select number of features (such as tense) and then an arbiter compared the two sets of annotation resolving any conflicts. After a selection of features received its official annotation the students moved on to annotate another selection of features. This process was repeated until all the designated features had been searched for and the test items were annotated appropriately.

4. Statistical analysis

4.1. IRT analysis

An IRT (Item Response Theory) based analysis was performed on the 152 test items in order to create a unified statistically based measure of each item's relative difficulty. The IRT was chosen because it allows the comparison of test questions from different tests. This type of analysis requires some of the same questions to appear on each test. Our data comprises three separate tests of 60 questions each where 14 of the 60 questions existed on a previous test. Our population of test takers consisted of 376 students at a local educational institution for ESL learners.

The IRT analysis compared student responses to each question and assigned a numerical difficulty score where higher scored items were generally of a higher difficulty than lower scored items. The IRT assigns difficulty based in part on comparing student performance across the test. The IRT analysis is also partially based upon the premise that students who generally perform well will also be able to answer the more difficult questions. The purpose of performing the IRT was to assign a numerical measure to each test question so that further multiple linear regression analysis could be performed.

The IRT analysis revealed that our tests had reliability scores between 97% and 98%. Also, the IRT provided fairly normally distributed difficulty scores as evidenced in the density plot (see Figure 1). The next step in our analysis was to perform multiple linear regression analysis using all the various sentence-level features tagged and the difficulty score (item measure).

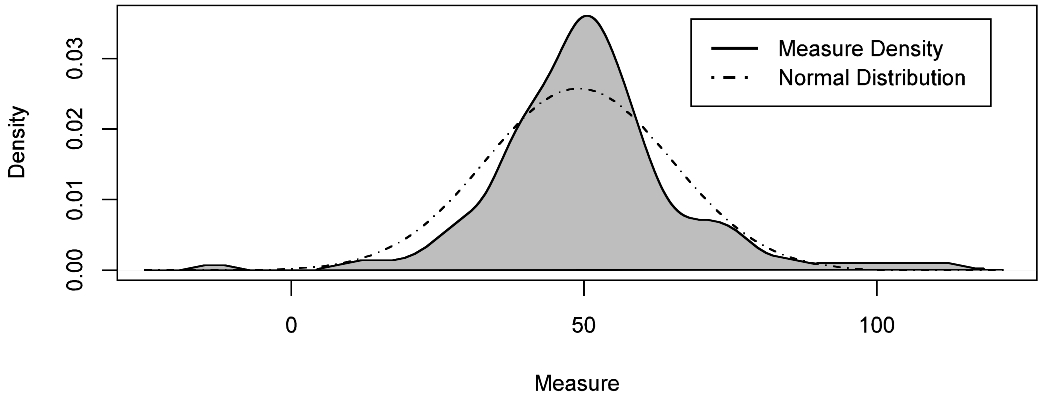

Figure 1. IRT measure density.

Figure 1 illustrates that given the constraints of the IRT the test items did create a near normal distribution of difficulty. The density on the y-axis was calculated using a density function that provides a statistically probable interpolated approximation of the complete population given the data points known. This method was chosen to avoid possible skewing of the results, which may occur if only a histogram is used. The overreaching goal of this work is to determine what makes a particular test item difficult or in other words what features of a sentence discriminate between how well a student performs on a given test item. The IRT analysis provided the first crucial step towards answering this goal, insofar as our 152 test items are concerned. With the IRT provided difficulty scores we performed multiple linear regression analysis of the test items using the IRT measure as the dependent variable and the sentence features as the independent. The analysis yielded a predictable result.

4.2. Step-wise regression

The initial step-wise regression revealed a model that accounted for 67% of the variability in the difficulty measure for a given test item. However, sentence syllable count was the greatest contributor of model accountability with an adjusted *R*-squared value of .65. The next closest feature (third person singular present tense) accounted for only 1% of the variability in the model. The initial model of 44 features was evaluated and it was determined some covariance existed and some features could be condensed. Another factor considered was how often certain features appeared over the whole model. Some features only appeared once or twice out of all the test items so it was decided that if a feature appeared less than 10 times it would be excluded from future analysis. Due to this analysis 14 features were removed based on low frequency of occurrences and four features describing modals were consolidated into a single feature (feature 40), leaving a refined feature set of 27 features out of the original 44 (see Appendix for the original list of features, removed features (Appendix Table 1) and the modified feature set (Appendix Table 2)). Four features describing modals (specifically: 21 (modal (present)), 22 (modal (past)), 40 (modals), 41 (semi-modals)) were condensed into one single feature (Feature 40). A correlation between the existence of contractions and negations was found however, it failed to meet our predetermined threshold of 80% covariance. Another correlation was found between simple and complex sentences but also failed to meet the necessary threshold.

4.3. Variable selection

Using our refined data set we performed a step-wise model selection process using Akaike's Information Criterion (AIC). The resulting model contained 15 sentence-level features and had an adjusted *R*-squared value of .67. This was an important step because it provided a model that accounted for approximately the same amount of variability with just 15 features versus our initial model of 44 features that had an *R*-squared value of .67. The refined model also differed from the initial model in which features were considered statistically significant. Syllable count remained in the top position however, tense and aspect features moved into predominance over the original model's features.

4.4. Further analysis

At this point we decided to perform some exploratory data analysis on the distribution of the number of syllables per sentence item. The 152 test items had a mean syllable count of 10.09 with a minimum three syllables to a maximum of 33. The median syllable count was nine. A simple plot of the 152 syllable counts revealed a few possible outliers however it was only after examining a plot of the density distribution (see Figure 2) it became clear that although from a normal population the density distribution was seriously skewed. The next logical step was to examine the distribution of difficulty scores for each syllable count.

In Figure 3 it is clear that syllable counts and difficulty scores do not have a linear relationship. For example test items with a syllable count of 10 have difficulty scores ranging from approximately 39 to 62. This is significant because according to our initial and refined models syllable count accounts for the greatest variability in scores and if there is no strict linear relationship then there is room for other features to possibly explain the variability seen in difficulty scores. Therefore, it appears from this graph that there are a range of syllable counts where other features must be responsible for variation in the relative difficulty of the test questions.

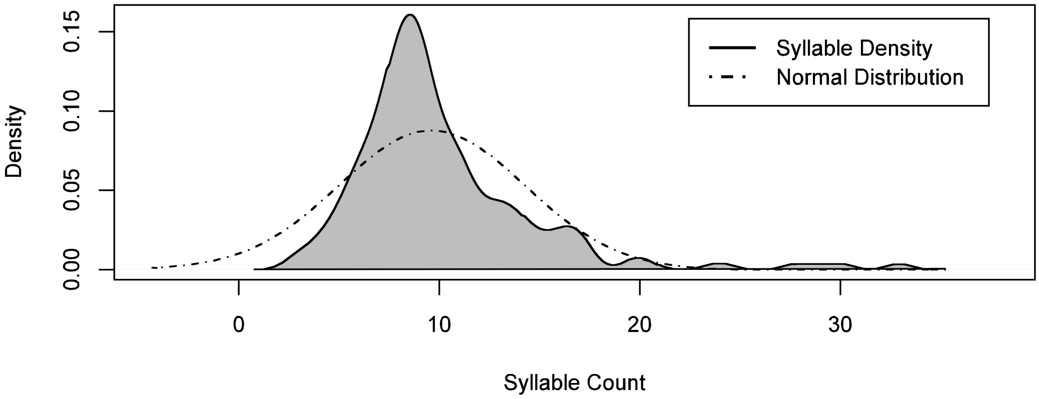

Figure 2. Syllable count density.

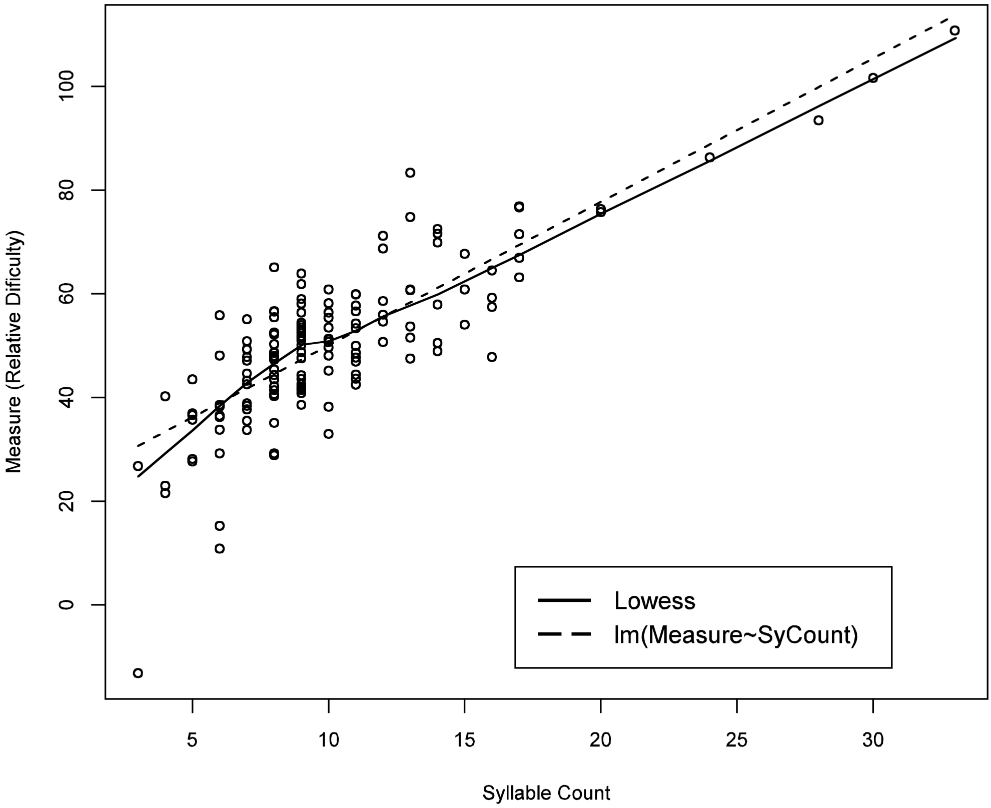

Figure 3. Syllable count vs. IRT measure.

4.5. Syllable bands

Figure 3 prompted further investigations into specific syllable count bands. Test questions with a syllable count of 7, 8 and 9 were isolated into three bands and then tailored linear models were created using Akaike's Information Criterion. The resulting linear models had surprising and informative results. Figures 4 and 5 illustrates how each syllable band contains overlapping difficulty regions and that difficulty is spread across all three bands contrary to what was expected based upon syllable counts overwhelming influence in the initial step-wise regression model.

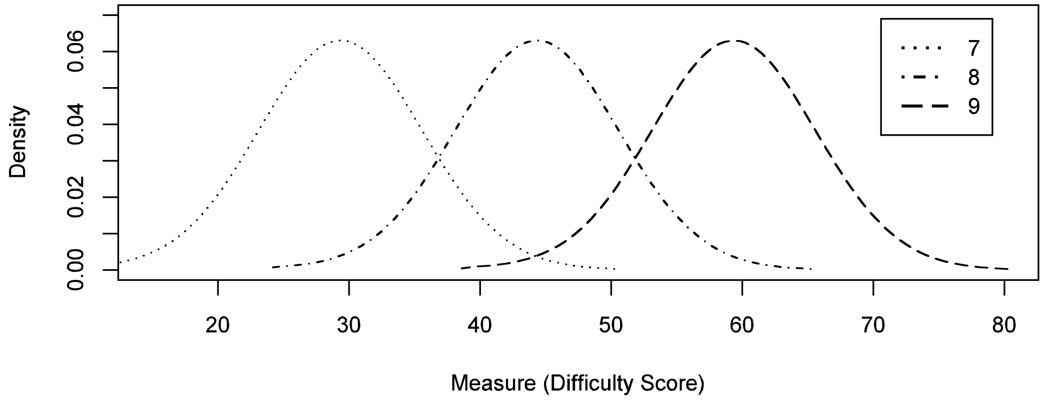

Figure 4. Seven, eight, nine syllable count's expected measure density.

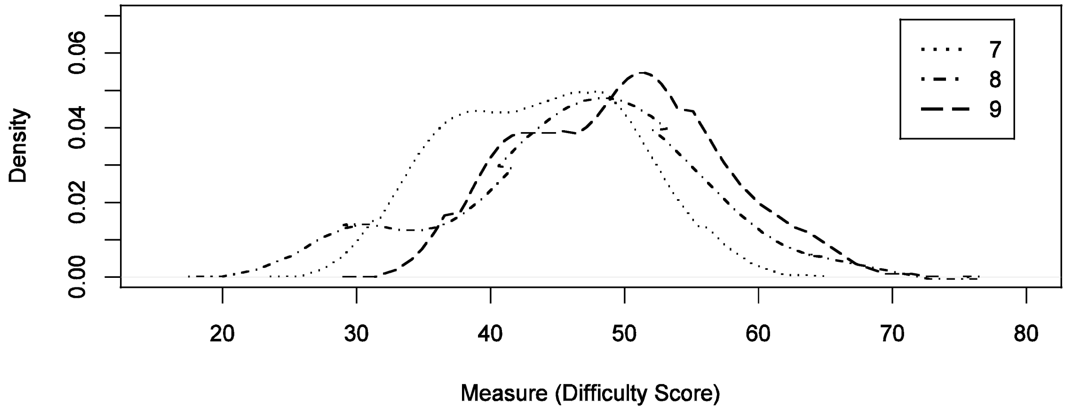

Figure 5. Seven, eight, nine syllable count's actual measure density.

4.5.1. Seven syllable count

There were 14 test items that had a syllable count of seven. An example of a 7-syllable test question is "Did you buy that at the store." Performing the AIC analysis revealed a linear model with an adjusted *R*-squared value of .95 and a *p*-value of < .01. The AIC analysis selected 8 features as the independent variables for the model. The feature determined to be most statistically significant was the presence of a preposition. Prepositions were closely followed by the presence of a modal and if the sentence was declarative. Other features found to have statistical significance were the presence of past and present tensed verbs as well as the presence of plurals. The lexical complexity of the item was also found to be statistically significant.

4.5.2. Eight syllable count

There were 25 test items that had a syllable count of eight. An example of an 8-syllable test question is "Is John going to the party?" Performing the AIC revealed a linear model with an adjusted *R*-squared value of .75 and a *p*-value of .01. The final model contained 16 independent variables which caused the adjusted *R*-squared value to go from a .92 to only .75. This increase in variables was mainly due to a few variables that approached significance and were therefore included in the model. The three most statistically significant features for this syllable band were the presence of a modal, the present tense, and contractions. The perfect tense and the presence of intransitive verbs were also found to be statistically significant.

4.5.3. Nine syllable count

The 9-syllable count band of test items returned to a lower number of independent variables. The linear model the AIC calculated only had eight sentence features. An example of a 9-syllable sentence from our test would be “If she listens she will understand.” This test question also contains the third person singular which was determined to be the most statistically significant sentence feature for this model. The model also suffered a severe adjustment from an R-squared of .70 to an adjusted figure of .50 with the largest p-value yet of .01. Other sentence features that were found to have a statistically significant impact on the measure score were the imperfect tense, the presence of articles, whether the sentence was a simple or complex sentence, transitive verbs, the perfect tense, interrogatives, and contractions.

4.6. Conclusions

An interesting pattern emerged when all the models' most statistically significant contributing features were compared. Some features would be statistically significant for one syllable band and then as the number of syllables increased they would slowly lose significance. The presence of prepositions seemed to follow this pattern, starting as the most statistically significant for our 7-syllable test items to less statistically significant in 8-syllable items to disappearing in the 9-syllable count only to re-emerge as having some statistical significance in our initial model over all items. Other features would also fluctuate in statistical significance across syllable bands. In fact five of the most statistically significant features in the 8-syllable band were from the 7-syllable band. However none of the 7-syllable band features were statistically significant in the 9-syllable band's model.

5. General conclusion

While these results are encouraging and future work will use them to help guide future test question creation, it is important to acknowledge these results are not conclusive or exhaustive in nature. Correlation does not imply causation and it will take further testing and analysis of new test items to fully understand how Elicited Imitation performs as a method of psycholinguistic inquiry as well as a proficiency measure. However it must be stressed that these results are promising and raise questions regarding the importance of other item features (morphological features or syntactic features) that can influence subject performance beyond syllable count.

6. Future work

Further investigation of how sentence features influence difficulty will be aided by using this research to create new tests that focus in on specific features at specific syllable count lengths. Also we plan in the future to explore the use of Natural Language Processing techniques to create a program to assist with the automatic detection of features in sentences. It is desirable that a large bank of test items with their features tag should be created and an automatic method of feature tagging would assist in this work immeasurably. If a reasonable relationship between different features and the outcome of a student’s performance can be found then there will be a basis for the creation of an adaptive language proficiency test using the Elicited Imitation methodology. An adaptive test could also then be integrated with our previous research into the use of Automatic Speech Recognition to grade student responses.

7. Appendix

Original List of Features:

F1 Subordinate clause
F2 Non-Finite clause
F3 Embedded clause
F4 Relative clause (obj)
F5 Relative clause (subj)
F6 Wh- questions
F7 Passive
F8 Simple Sentence
F9 Compound
F10 Complex
F11 Compound-complex
F12 Declarative
F13 Interrogative
F14 Exclamatory
F15 Imperative
F16 Zero place
F17 Copular
F18 Intransitive
F19 Transitive
F20 Ditransitive
F21 Modal (present)
F22 Modal (past)
F23 Negative
F24 Do insertion
F25 Past tense
F26 Present tense
F27 Future tense
F28 Simple tense
F29 Perfect tense
F30 Imperfect tense
F31 Perfect Progressive
F32 Perfect Passive
F33 Progressive Passive
F34 Articles
F35 Plural
F36 Preposition
F37 Contraction
F38 Possessive
F39 3rd P. singular present tense
F40 Modals
F41 Semi-modals
F42 Number of T-Units
F43 Syllable Length
F44 Average Lexical Frequency

Table 1. *Removed Features*

Feature	Explanation	Count
F14	Exclamatory	0
F3	Embedded clause	0
F32	Perfect Passive	0
F33	Progressive Passive	0
F6	Wh- ?	0
F15	Imperative	1
F4	Relative clause (obj)	1
F5	Relative clause (subj)	1
F7	Passive	1
F11	Compound-complex	2
F31	Perfect Progressive	3
F10	Complex	4
F20	Ditransitive	7
F16	Zero place	8
F41	Semi-modals	9
F22	Modal (past)	12
F21	Modal (present)	24

Table 2. *Modified Feature Set*

Feature	Explanation	Count
F1	Subordinate clause	25
F2	Non-Finite clause	11
F8	Simple Sentence	123
F9	Compound Sentence	23
F12	Declarative	81
F13	Interrogative	69
F17	Copular verb	16
F18	Intransitive	61
F19	Transitive	82
F23	Negative	23
F24	Do insertion	37
F25	Past Tense	53
F26	Present Tense	91
F27	Future Tense	14
F28	Simple Tense	105
F29	Perfect Tense	31
F30	Imperfect Tense	22
F34	Articles	50
F35	Plurals	44
F36	Prepositions	52
F37	Contractions	22
F38	Possessive	17
F39	3rd Person singular present tense	13
F40	Modals	37
F42	T-unit	-
F43	Syllable Length	-
F44	Lexical Measure	-

References

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

Bley-Vroman, Robert & Chaudron, Craig (1994). Elicited imitation as a measure of second-language competence. In Elaine E. Tarone, Susan M. Gass & Andrew D. Cohen (Eds.), *Research methods in second-language acquisition*. (pp. 245–261). Hillsdale, NJ: Lawrence Erlbaum Associates.

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

Ellis, Rod (2008). Investigating grammatical difficulty in second language learning: Implications for second language acquisition research and language testing. *International Journal of Applied Linguistics*, *18*, 4–22.

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

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

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.

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

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

Ortega, Lourdes (2000). *Understanding syntactic complexity: The measurement of change in the syntax of instructed L2 Spanish learners* (Unpublished Doctoral dissertation). University of Hawaii, Honolulu, HI.

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

Copyright information

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:

Hendrickson, Ross, Meghan Aitken, Jeremiah McGhee, and Aaron Johnson. 2010. What Makes an Item Difficult? A Syntactic, Lexical, and Morphological Study of Elicited Imitation Test Items. In *Selected Proceedings of the 2008 Second Language Research Forum*, ed. Matthew T. Prior et al., 48-56. Somerville, MA: Cascadilla Proceedings Project. www.lingref.com, document #2384.