Principled Construction of Elicited Imitation Tests

Carl Christensen, Ross Hendrickson, Deryle Lonsdale

Brigham Young University
Provo, UT, USA 84602
cvchristensen@gmail, ross.hendrickson@gmail, lonz@byu.edu

Abstract

In this paper we discuss the methodology behind the construction of elicited imitation (EI) test items. First we examine varying uses for EI tests in research and in testing overall oral proficiency. We also mention criticisms of previous test items. Then we identify the factors that contribute to the difficulty of an EI item as shown in previous studies. Based on this discussion, we describe a way of automating the creation of test items in order to better evaluate language learners' oral proficiency while improving item naturalness. We present a new item construction tool and the process that it implements in order to create test items from a corpus, identifying relevant features needed to compile a database of EI test items. We examine results from administration of a new EI test engineered in this manner, illustrating the effect that standard language resources can have on creating an effective EI test item repository. We also sketch ongoing work on test item generation for other languages and an adaptive test that will use this collection of test items.

1. Background

Researchers have long proposed elicited imitation (EI) as a way to investigate second language acquisition (Naiman, 1974). EI is a testing method that requires subjects to hear a spoken stimulus sentence and then attempt to repeat it back as accurately as possible. The basic premise of EI language testing is that as a stimulus grows in complexity, the performance of the subject should degrade in a corresponding manner. This is because, as a subject is exposed to a given stimulus, they form a representation of that stimulus and then attempt to reproduce a response based on the representation they have stored. Presumably the representation cannot encode linguistic content that exceeds the subject's knowledge of the language in question. However, for short time latencies or simple test items, short-term or working memory may serve to bypass the encoding/decoding steps. It is thus essential that the complexity of the stimulus be controlled. Controlling for complexity allows researchers to make sure that the subject's language proficiency is being investigated rather than their memory capacity.

Recently, studies have also focused on the ability of EI tests to estimate oral proficiency (Vinther, 2002). Many different methods have been proposed to measure oral language proficiency; however, there is considerable debate as to the validity of these measures and of each method's efficacy as both a testing measure and as an instrument of linguistic inquiry (Casad, 1997). Most forms of oral proficiency testing rely on some type of free language production task (such as story-telling or an interview). However, despite common use of free language production tasks in language testing, they are not often used in second language acquisition (SLA) research. The main objection to free language production tasks in SLA research is that they are difficult to rate and do not always provide the desired linguistic phenomena (Erlam, 2006). In a previous study comparing interview, imitation and completion methods the researchers found imitation to be the most valid approach for language testing (Henning, 1983). Elicited imitation (EI) has also been proposed as a valid method for both language testing and linguistic inquiry (Bley-Vroman and Chaudron, 1994).

While EI scholars have identified many of the factors that contribute to language complexity in EI test sentences (or items), skeptics claim the sentences are often contrived and strange and that the process of EI testing less natural than other oral measures (Jessop et al., 2007), which consequently leads to inaccurate results. For example, one test examined the effects of verb-object predictability with the sentence "The spider is playing a drum." (Valian et al., 2006). Similarly, another EI test designed to investigate the effects of lexical density on vocabulary assessment contained the item "The accumulation of poison in the vegetation is appalling." (Graham et al., 2008 in print).

In order to maximize linguistic information gleaned from an EI test about a learner's language ability, test items must be carefully constructed with respect to syntax, morphology, lexical frequency, and sentence length. Table 1 demonstrates EI test items constructed with each of these aspects in mind. These varying constraints have left the methodology underlying construction of EI tests open to interpretation and speculation.

Despite these objections, scholars and researchers have widely acknowledged EI as a quick and inexpensive way to gain some insight into the proficiency of a speaker in a second language. However, many issues still remain that preclude widespread EI testing. Test construction, test administration, and scoring are a few of the uncertainties that remain under investigation (Chaudron et al., 2005). In this paper we focus on test item construction, including the methodology behind test item creation (Jessop et al., 2007) and the need for a collection of annotated EI test items.

2. Test item creation

Our purpose in EI test item construction is to increase the validity of the EI test by creating stronger correlation with current oral proficiency measures, and by creating standard difficulty measures against which EI items can be compared. In order to create difficulty measures, we survey the factors that contribute to the difficulty of an EI test item.

Various studies have identified grammatical and lexical features that contribute to the relative difficulty of an EI test

Syntax Focus	The present director had been writing a new proposal.
Lexical Focus	You are successfully resisting administrative input.
Length Focus: Short	We eat cookies.
Length Focus: Long	When Jim entered the office he was immediately afraid of the uncommunicative boss.

Table 1: Example sentences from EI tests (Graham et al., 2008 in print; Weitze and Lonsdale, 2009 in print)

item. Sentence complexity, sentence length, tense, aspect, lexical density, and verb object predictability are a few of the features that play a significant role in the ability of a speaker to correctly repeat an utterance (Valian et al., 2006; Graham et al., 2008 in print; Weitze and Lonsdale, 2009 in print). As mentioned earlier, the complexity of these features is compounded by the need to control for the effects of working memory. Research has suggested that features that occur at the end of the sentence are the easiest to retain in working memory, and that features at the beginning or middle of the sentence are more difficult (Erlam, 2006). A well-constructed EI test must account for the position of grammatical features in the sentence. The great number of significant features and the complexity of these features mean that EI tests often only focus on a particular subset of features. This in turn affects the generality of the judgments which can be derived from such a test about a language learner's oral proficiency.

The work we report on here involves constructing EI test items having features capable of supporting accurate predictions about a speaker's overall oral proficiency. To do this we utilize various language resources and computational tools which enable us to specify grammatical and lexical features that co-occur in sentences. In this way we optimize the efficiency of test items aimed at providing good correlation between the EI test and other oral proficiency measures. These tools also enable us to fix the position of given features in any test item, which allows us to control for working memory effects.

As suggested by (Jessop et al., 2007), corpus linguistics can resolve some of the issues with EI test item creation. Results have also shown that more natural language increases evaluation performance in EI-like tasks (Luo et al., 2009). For the source of our test items, we selected the English Gigaword corpus (Graff and Cieri, 2003) for a variety of reasons. Most generally, the criticism concerning the contrived nature of EI test items finds its most natural solution in the corpus as it represents a more natural form of language (Biber et al., 1998). The text found in the Gigaword corpus covers a board range of themes, thus creating a more richly diverse collection of sentences to draw from for the various EI test purposes. Finally, the volume presented by the Gigaword corpus increases the amount of sentences annotated and consequently the likelihood of finding sentences with desired features. While we could have mined the already parsed sentences from the Penn Treebank (Marcus et al., 1999), the Gigaword corpus provided a larger number of potential items. This corpus provided us with over 50,000 sentences that met our criteria for acceptable syllable length. These were then annotated and inserted into an EI test item database. The set of features that we chose for annotation encompass several linguistic levels (e.g. vocabulary level, syllable count, morphological complexity, phrasal and clausal syntax). We designed our annotation scheme from our prior studies of features' contributions to statistically significant variance in subject performance.

3. Item creation/generation tool

To aid in constructing and annotating EI test items, we developed an automatic sentence analysis tool (Hendrickson and Lonsdale, 2009). Many freely available resources exist for linguistic processing, but none offer the combination of features necessary for fully specifying sentences in the context of EI test item development. This work is hence innovative since it brings together several disparate resources into a tool that provides information to help in this task. The tool is implemented in the Java programming language to assure maximal portability and compatibility with future tool extensions.

This tool draws its data either from single sentence input by the user, or from a corpus. It then makes use of the Stanford parser (Klein and Manning, 2002), the CELEX database (Baayen et al., 1993), the English version of WordNet (Fellbaum, 1998), and tree regular expressions (Aiken and Murphy, 1991) to parse and annotate relevant features in each sentence. Figure 1 shows a sample sentence from the Gigaword corpus with the results from all stages of analysis including syllable count, average lexical frequency, parse tree complexity, etc.

In order to extract and annotate all the sentences in the Gigaword corpus, we had our system parse out each sentence identified by a sentence disambiguation tool from the raw text files. Each sentence is then run through the Stanford parser. Grammatical structures are then annotated by searching the resulting tree structure with tree-based regular expressions such as those seen in Table 2. When these structures are located we identify the position of the feature in the sentence and mark each feature accordingly.

Next, each word is passed through the CELEX database to obtain a count of morphemes. The count is then averaged over the sentence and the average number of morphemes per word is annotated as a feature. We next obtain a count of syllables in each sentence as a measure of length either from the CELEX database or via our own heuristic and annotate the sentence with the count. Then WordNet provides the number of word senses per word, which we average over the sentence as an annotation of a semantic feature. We finally check every word against a frequency dictionary to obtain a lexical density measure for each sentence.

After each sentence has undergone this annotation process, it is entered into the database and the process is repeated for each additional sentence. Thus each item in the database is annotated for lexical, morphological, grammatical, and

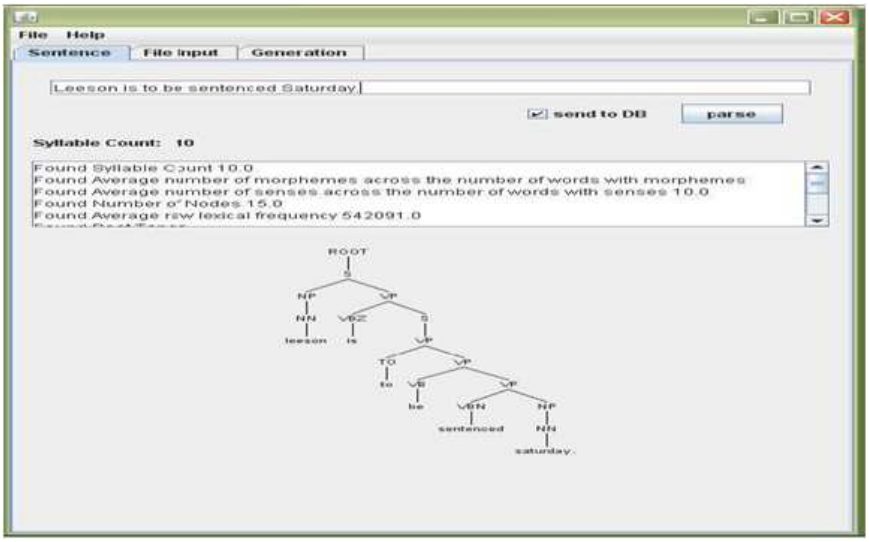


Figure 1: A screen shot of the tool with a corpus sentence that has been parsed and annotated

Copular	((VBZ|VB|VBN < (are|is|am|was|were)) !.. VP)
Intransitive	(VP <‘ VB | <‘ VBD | <‘ VBZ)
Transitive	(VP <‘ NP)
Present perfect progressive	(VP <1 (VBP . (VBN . VBG))) || (VB < (have) . (been))

Table 2: Tree-based regular expressions used in the tool for annotating sentences

length features. This methodology can also be applied to other corpora besides Gigaword newswire.

As further research uncovers more significant features in EI test items, our tool can be extended to include additional features. The tree-based regular expressions make the identification of future syntactic elements as simple as adding the necessary regular expressions to match the syntactic structure.

This tool makes the items accessible through a generation application that enables quick and precise EI test creation. This interface, as shown in Figure 2, enables researchers to specify desired grammatical features, the desired relative position in the sentence, lexical density of the item, the morphological complexity, and sentence length. Test creation thus becomes a question of purpose and helps researchers control for confounding factors. This tool also takes the first steps in opening the possibility to implement an adaptive EI test with a large corpus of EI test items.

This method of item creation provides items that come from real English sentences and thus avoids the contrived nature of many of the hand-made test items previously created while still maintaining the tight constraints required for an EI test to be a valid instrument of oral proficiency. Therefore, the test can be used to evaluate a second language learner's oral proficiency using a more plausible collection of utterances. Table 3 illustrates how sample items created by the tool are more realistic and natural than hand-crafted ones.

4. Results

In order to compare the ability of our test items to predict oral proficiency as opposed to hand-crafted EI test items, we created a new EI test from our database of engineered EI items drawing directly on the features that were annotated. Selecting items with various features, including variable syllable length (between approximetly 6 and 24 syllables for each item) allowed us to better calibrate our test for measuring overall oral proficeiency.

We then administered this new EI test to 127 adult English as a Second Language (ESL) learners who were students at a university English Language Center. Our test consisted of 60 items and was administered in a fashion directly comparable to similar previous studies (Graham et al., 2008).

Human annotators scored the test syllable-by-syllable using a web-based scoring tool that we developed for this purpose. Figure 3 shows the scoring tool's interface which allows a grader to assign each syllable a binary score after listening to the student's response. Subsequent analysis assigned a 4-score to the entire item where each item receives a maximum possible score out of 4, with a 1-point decrement for each missed syllable (Chaudron et al., 2005). The human-scored items are then inserted into a database.

We compared the results of our test with a speaking language achievement test (SLAT) administered to the students in the same week. This test is administered via another computer application and is designed to measure oral achievement of the ESL students (Graham et al., 2008 in print). We then compared our correlation with the SLAT against previous EI tests' correlation with the SLAT test (Graham et al., 2008).

The syllable-scored results from our test administration showed a 0.75 correlation, significantly better ($p < 8.71e-06$) than the previous EI test's SLAT correlation (which was 0.41). Scoring with the 4-point scale, our form showed a

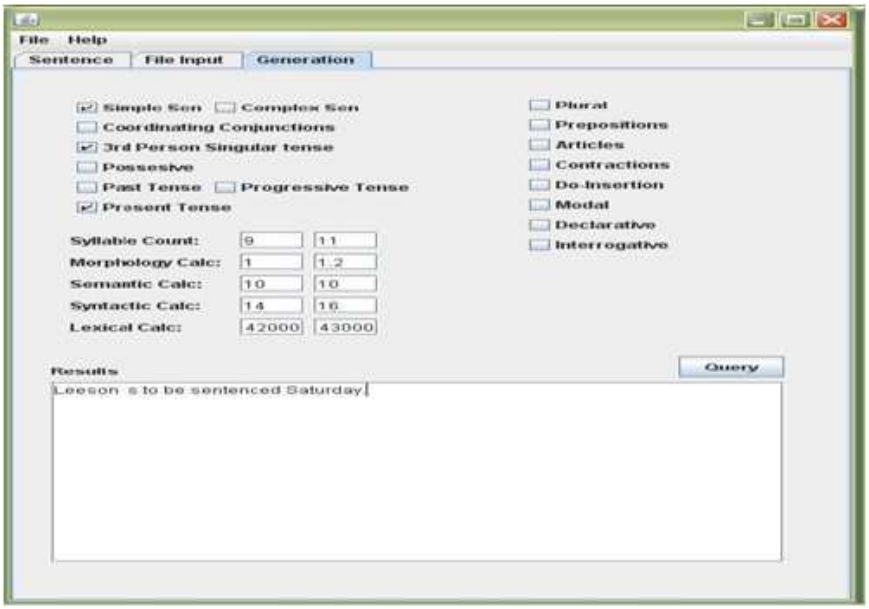


Figure 2: Screen shot of the generation tool showing measurements and features available for specification

Chris has yelled louder than ten sheep.
Naturally dogs are agitated in the presence of lions.
Are they walking slowly because their feet are sore?
During that time we felt a bit helpless.
We accomplished what we set out to do.
I wasn't as nervous and anxious as I thought I'd be.

Table 3: Example EI test items: created manually (above) versus interactively via the tool (below).

0.71 correlation with the SLAT, also significantly better ($p < 0.05$) than the previous test (0.55). Figure 4 shows a scatterplot for the syllable and 4-score correlations.

We attribute the higher correlation to the method of construction of the test items with respect to the linguistic information contained in the EI test items along with the more natural form of language made available by the corpus.

5. Future work

In further work we hope to create various EI tests with specific purposes by applying our methodology to other corpora that may contain even more naturalistic English sentences than newswire materials. Even more interesting would be to use speech corpus transcripts as input, though this would introduce further complexities for item selection and analysis. We also are pursuing matching EI test scores with other methods of oral proficiency testing besides the SLAT addressed in this paper.

Another goal of our work is to develop an on-line adaptive EI test. This type of testing tool would combine our corpus of EI items, ongoing research into ASR scoring methods (Graham et al., 2008), and EI administration procedures in real time to better calibrate test items with the learner's level of proficiency.

We are also in the process of applying our methodology to other languages and creating EI test item databases for those languages, given the availability of relevant language resources analogous to the ones we used for English. Perhaps eventially it will be possible to generalize our EI test item difficultly measures across languages to provide a multilingual standard EI test item scale.

6. Acknowledgements

We would like to thank members of the BYU PSST research group (see http://psst.byu.edu), the BYU English Language Center, the BYU Center for Language Studies, and the BYU ORCA MEG program.

7. References

A. Aiken and B.R. Murphy. 1991. Implementing regular tree expressions. In *Proceedings of the 5th ACM conference on functional programming languages and computer architecture*, pages 427–447, Cambridge, MA.

R. H. Baayen, R. Piepenbrock, and H. van Rijn. 1993. The CELEX lexical database (CD-ROM).

D. Biber, S. Conard, and R. Reppen. 1998. *Corpus linguistics: Investigating language structure and use*. Cambridge University Press, Cambridge.

R. Bley-Vroman and C. Chaudron. 1994. Elicited imitation as a measure of second-language competence. In E.E. Tarone, S. Gass, and A.D. Cohen, editors, *Research methodology in second language acquisition*, pages 245–261. Lawrence Erlbaum, Hilldale.

F. H. Casad. 1997. Language assessment tools: Uses and limitations. In Martin Puetz, editor, *Language Choices:*

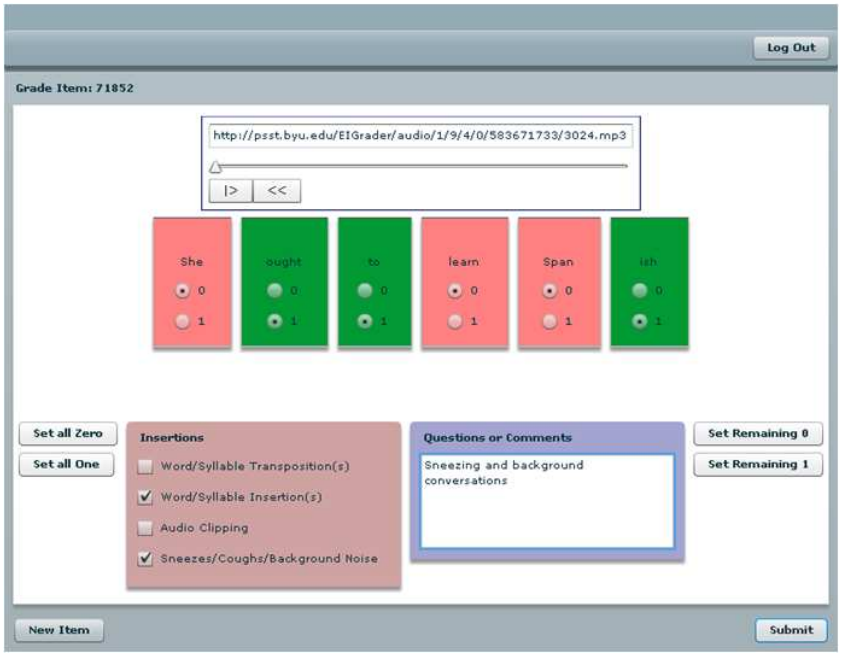
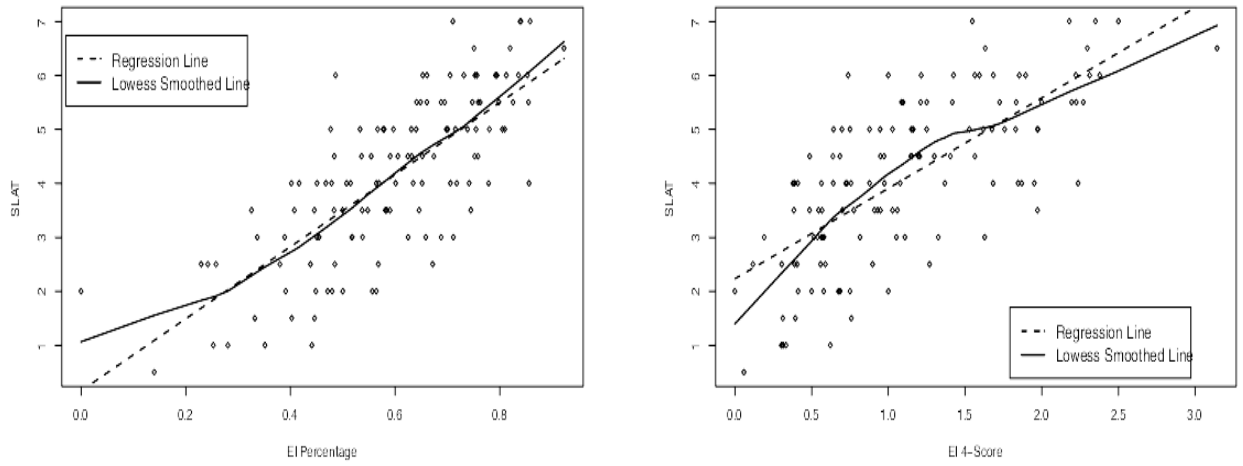


Figure 3: A screen shot of the scoring tool with an EI item



(a) **(b)**

Figure 4: Correlations between EI and SLAT scores: percentage 4(a) and 4-score 4(b)

Conditions, Constraints, and Consequences, pages 253–273. Benjamins, Amsterdam. Impact: Studies in Language and Society, 1.

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

R. Erlam. 2006. Elicited imitation as a measure of L2 implicit knowledge: An empirical validation study. *Applied Linguistics*, 27(3):464–491.

C. Fellbaum. 1998. *WordNet: An Electronic Lexical Database*. MIT Press, Cambridge, Massachussets.

D. Graff and C. Cieri. 2003. English Gigaword.

C.R. Graham, D. Lonsdale, C. Kennington, A. Johnson, and J. McGhee. 2008. Elicited imitation as an oral proficiency measure with ASR scoring. In N. Calzolari (Conference Chair), K. Choukri, B. Maegaard, J. Mariani, J. Odjik, S. Piperidis, and D. Tapias, editors, *Proceedings of the 6th International Language Resources and Evaluation Conference (LREC'08)*, Marrakech, Morocco. European Language Resources Association (ELRA).

R. Graham, J. McGhee, and B. Millard. 2008, in print. The role of lexical choice in elicited imitation item difficulty. In *Proceedings of Second Language Research Forum (SLRF) 2008*.

R. Hendrickson and D. Lonsdale. 2009. The use of NLP

technologies to engineer oral proficiency test items. Paper presented at CALICO Pre-Conference Workshop on Automatic Analysis of Learner Language (AALL '09).

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

L. Jessop, W. Suzuki, and Y. Tomita. 2007. Elicited imitation in second language acquisition research. *The Canadian Modern Language Review/La Revue canadienne des langues vivantes*, 64:215–238.

D. Klein and C.D. Manning. 2002. Fast exact inference with a factored model for natural language parsing. *Advances in Neural Information Processing Systems (NIPS 2002)*, pages 3–10.

D. Luo, Y. Qiao, N. Minematsu, Y. Yamauchi, and K. Hirose. 2009. Analysis and utilization of speaker adaptation techniques for shadowing and read-speech pronuncation evaluation (in Japanese). Technical report, IEICE. 109(99), pp 51-56.

M. Marcus, B. Santorini, M.A. Marcinkiewicz, and A. Taylor. 1999. Treebank-3.

N. Naiman. 1974. The use of elicited imitation in second language acquisition research. *Working Papers on Bilingualism*, 2:137–53.

V. Valian, S. Prasada, and J. Scarpa. 2006. Direct object predictability: Effects on young children's imitation of sentences. *Journal of Child Language*, 33:247–269.

T. Vinther. 2002. Elicited imitation: a brief overview. *International Journal of Applied Linguistics*, 12(1):54–73.

M. Weitze and D. Lonsdale. 2009, in print. The effect of syntax on English language learning. *LACUS Forum XXXVI*.