Self-assigned Ranking of L2 Vocabulary

Heidi Brumbaugh*
Simon Fraser University, Burnaby, BC, Canada

Trude Heift*
Simon Fraser University, Burnaby, BC, Canada

Abstract

This article describes a research study that determined the depth of vocabulary knowledge of 28 intermediate ESL learners. The study was carried out with Bricklayer, a vocabulary assessment tool for L2 English which tested the ESL learners on 72 words. Two post-tests collected evidence for concurrent validity. A semantic distance test captured incremental knowledge for 36 words, but Bricklayer's predictive power for this partial knowledge was weak. A standard multiple-choice test of the remaining 36 words showed that Bricklayer predicted 61% of known words and 69% of unknown words; results were better for words which were *strongly* predicted to be known or unknown. These findings provide promise that Bricklayer's assessment paradigm assists in building up models of students' knowledge and behaviour in CALL environments.

Keywords: Computer Assisted Language Learning, vocabulary assessment, vocabulary depth, meta-cognition, self-assessment

Introduction

It may seem intuitive, even obvious, that language learners need to know words of the target language in order to communicate effectively. Nonetheless, Zimmerman (1997) points out that despite vocabulary's central role in language, over the course of the history of language teaching, vocabulary has not been emphasized. A surge of interest in vocabulary over the past decade has shifted this focus. Nation (2013), for instance, states that "over 30 per cent of the research on vocabulary that has appeared in the last 110 years was published in the past eleven years" (p. 5). This new body of research informs strategies for incorporating vocabulary instruction in the language classroom including computer-assisted language learning (CALL) contexts.

At its most basic level, vocabulary knowledge involves connecting the word form (written or spoken) with its associated meaning. Vocabulary researchers, however, have recognized that word knowledge is complex, and thus have tried to articulate a broader structure for vocabulary knowledge (Henriksen, 1999; Nation, 1990, 2001; Richards, 1976). These frameworks capture the idea that word knowledge is multifaceted. In addition to knowledge about a word's meaning, word knowledge also includes such features as associative knowledge, form production and recognition, morphology, collocations, etc.

**Tel: (1) 831 247 1379; Fax: (+1) 866 216-8918; E-mail: heidi@vocabsystems.com; 5733 Hollister Ave. Suite 7, Goleta, CA 93117 USA*
***Tel: (1) 778 782 3369; E-mail: heift@sfu.ca; Robert C. Brown Hall Building, Room 9201, 8888 University Drive, Burnaby, BC V5A 1S6 Canada*

Apart from the multifaceted nature of word knowledge, lexical knowledge is also acquired incrementally. In fact, the idea that a learner does not progress immediately from being unfamiliar with a word to having complete knowledge of all its meanings and usages was observed as far back as the early part of the twentieth century (Dolch, 1927). Durso and Shore (1991) characterize this intermediate level of knowledge as partially known words, or so-called "frontier" words (see also Shore and Kempe, 1999). Durso and Shore's studies show that although learners denied that the word was part of their language knowledge, they nonetheless were able to access some semantic content about the word.

An accurate assessment of the learner's vocabulary knowledge and stage of acquisition is especially critical for the L2 classroom because it informs and drives instructional strategies. CALL, in particular, is well suited for this task. Consider, for instance, that a computer could track and keep a record of whether a particular word was mostly known, mostly unknown, or a frontier word, and construct a model (i.e., a representation) of the learner's vocabulary knowledge accordingly.

Such a model is called a *learner model* or *student model* and is an integral part of a computerized intelligent tutoring system (ITS). A learner model allows an ITS to deliver individualized content for each student by considering each learner's behaviour and performance and tailoring instruction to their individual needs (see Heift & Schulze, 2003). For example, words which are mostly known by the learner would not need to be targeted for direct instruction, whereas mostly unknown words could be targeted for instruction or initial exposure. Unknown or partially known words in the text could be targeted for hyperlink glosses.

By identifying frontier words which are in the process of being assimilated into the mental lexicon, an ITS could target such words for what Nation (2001) calls "rich instruction," which "involves giving elaborate attention to a word, going beyond the immediate demands of a particular context of occurrence" (p. 95).

The following section discusses the most common vocabulary assessment tools in language instruction and evaluates the extent to which current assessment tools can capture multi-faced and incremental word knowledge. We also identify gaps in current vocabulary assessment techniques and introduce the CALL program Bricklayer which presents a new paradigm for L2 vocabulary assessment. We then describe a study which we conducted with 28 ESL learners to validate Bricklayer's performance. After presenting the results of our study, we discuss the merits of different types of vocabulary assessment tools and conclude with improvement suggestions for Bricklayer.

Vocabulary Assessment

Vocabulary assessment tools can generally be classified into two main types: breadth tests and depth tests.

Breadth Tests

The goal of the breadth test is to measure a learner's overall vocabulary size. Two widespread assessment tools of this type are the Vocabulary Levels Test (VLT) (Nation, 1983, 1990; Schmitt et al., 2001) and the Vocabulary Size Test (VST) (Nation & Beglar, 2007). These tests rely on sampling across different frequency bands or ranges in order to generate a comprehensive vocabulary score. The tests provide strong examples of *content validity*, in that a test is typically considered to be a sample of a particular domain (Messick, 1989).

Nonetheless, breadth tests are not designed to assess specific vocabulary items. For example, if the Levels Test indicates that the student knows 400 of the words at the 3,000 frequency band, there is no way of telling *which* 400 words are known and *which* 600 words are unknown. Furthermore, as Milton and Vassiliu (2000) point out, "learners acquire their knowledge from course books and not from frequency lists" (p. 446). The authors researched a small corpus of three first-year EFL course books for Greek students and found that the vocabulary was thematic and idiosyncratic. In addition, vocabulary at the 2,000 word range was underrepresented and vocabulary at the 3,000 word range was overrepresented, challenging the notion that vocabulary is acquired by students in the order suggested by frequency lists. Neither the VLT nor the VST pinpoint specific gaps in vocabulary.

The Checkbox Tests

The checkbox test, also known as the Yes/No test, is a common breadth assessment tool that relies on an examinee's self-assessment of word knowledge. Examinees are presented with a list of words and indicate via a checkmark which words they know. This format is also used as a breadth assessment for the Eurocentres Vocabulary Test (EVT) (Meara, 1990), which samples from the different frequency bands and then estimates total vocabulary size.

In order to quickly assess information about many words at once, a self-assessment tool such as the checkbox test is arguably a good choice. For example, one version of the REAP vocabulary tutor (Rosa & Eskenazi, 2013) uses the checkbox test to create a model of a learner's knowledge, as described in the previous section, in order to individualize instruction. However, there are some concerns with the validity and reliability of the checkbox test. The assumption underlying self-assessment is that examinees know what they know. This idea of self-assessment has been investigated empirically; for the most part, learners can accurately self-assess their knowledge (LeBlanc & Painchaud, 1985). However, if the examinee checks a box for a word, or clicks the "Yes" button in the case of a Yes/No format test, how does the examiner know that this is accurate? The most common approach to verifying test response is to include "pseudowords" mixed in with the target test words (Anderson & Freebody, 1983; as cited in Beeckmans et al., 2001). The general idea here is that pseudowords provide a way to measure the extent to which examinees overestimate their vocabulary knowledge. The tester can count the number of pseudowords incorrectly selected as real words ("false alarms") and then apply a correction formula to modify the examinee's final score based on the number of real words correctly selected ("hits").

Unfortunately, the reliability of this technique, which has been extensively studied, varies widely. Although some researchers have found that the checkbox test correlates highly with other vocabulary measures (Meara & Buxton, 1987; Mochida & Harrington, 2006), others have found conflicting results. Pellicer-Sánchez and Schmitt (2012), for instance, compared different false alarm formulas and found that the accuracy of the corrected score depended on the number of false alarms. Beeckmans et al. (2001) likewise discovered that test scores can change dramatically based on which correction formula is applied.

In addition to these concerns, the checkbox test does not have the ability to capture partial knowledge and thus pedagogical interventions of words on the frontier of acquisition are not possible.

Depth Tests

There are several vocabulary assessments designed to detect the learner's depth of word knowledge and they differ in the types of lexical depth they measure: Webb (2005): different knowledge types; Schmitt (1998): four different kinds of word knowledge; Meara (2009): word association knowledge; Schmitt and Meara (1997): depth of word association as well as depth of knowledge for verbal suffixes; Nagy, et al. (1985) and Collins-Thompson & Callan (2007): precision of semantic meaning; Qian (2002): synonymy, polysemy, and collocational knowledge; Schmitt (1998b) and Crossley et al. (2010): polysemy; and, Laufer and Nation (2001): fluency.

One of the more ambitious assessment instruments is the Vocabulary Knowledge Scale (VKS) (Paribakht & Wesche, 1997; Wesche & Paribakht, 1996), which aims to measure depth of different kinds of word knowledge via varying levels of questions. In the VKS, first students indicate whether they have seen a word, then whether the meaning is known; if known, they first produce the word meaning, and then a sentence. Scoring is based on how much knowledge was indicated.

The advantage to the VKS is that gradations of understanding can be captured. The downside is that it is very time-consuming to administer; furthermore, Laufer and Goldstein (2004) point out that it does not necessarily measure what it purports to measure. Indeed, most of the assessment instruments mentioned above are designed for research purposes. Some are very arduous to administer (Webb's (2005) assessment, for example, requires ten questions for each word), making them impractible for the L2 classroom.

Bricklayer, a Vocabulary Assessment Tool

Bricklayer, the assessment tool developed by the lead author and used in the current research study, combines elements of several existing assessments. Like the VST and other multiple-choice style assessments, it presents learners with quizzes which measure their ability to recognize the correct meaning of a lexical form. Like the checkbox test, Bricklayer relies primarily on learners' self-assessment of their lexical knowledge. Like the VKS, it measures depth of knowledge for individual words. Yet, there are also a number of significant differences between Bricklayer and existing vocabulary tools.

Bricklayer modifies the self-assessment paradigm by addressing the validity issues of the checkbox assessment which were surveyed in the previous section. Most importantly, Bricklayer's goal is to rapidly measure depth of knowledge for a large number of words by providing a task in a game environment which forces the learner to rank a given word list in terms of the learner's semantic knowledge for each word. Because words are ranked along a continuum, Bricklayer is designed to capture mostly known and mostly unknown words as being the words at the edges of the rankings. The words at the middle of the rankings are considered somewhat known. For the purpose of our study, we consider a word mostly known if the primary meaning for a word can be correctly recognized. A word is considered to be somewhat known if it is familiar to the learner, or if the learner can identify the general semantic domain of the word. A word is considered to be mostly unknown if the word is not recognized or familiar.

Bricklayer's quiz presentation is unique in that these quizzes are only presented for a random subset of the words. In this way, the quizzes serve as a means of verifying the accuracy of the learner's self-report. Furthermore, quiz results are weighted differently depending on the rank that the learner assigns the word. That is, if the learner indicates the word is strongly known, and then incorrectly answers a quiz for that word, their score in the game is more strongly penalized than if they rank the word as weakly known. Accordingly, the learner is rewarded with a higher game score if they accurately represent their lexical knowledge.

In Bricklayer, the learner receives feedback and an indication of the kind of knowledge the program is asking for in the form of these random mini-quizzes. When the learner knows that there may be a multiple-choice quiz in which they must associate form with meaning, then it is clear that this is the type of knowledge being elicited. This is in accordance with Eyckmans (2004), who found that different instructions affected the reliability of the checkbox assessment.

Bricklayer also produces mini-quiz results and gamescores which verify that the learner is accurately representing their knowledge. For example, if all of the quiz questions in a game are wrong, it is likely that the word list for that game is too hard and that the rankings are thus not reliable. This addresses the limitations of the use of pseudowords in the checkbox assessment, which is not always a reliable way to verify that knowledge is being accurately represented.

Finally, Bricklayer, unlike the binary checkbox assessment, provides a mechanism for word knowledge to be ranked. Using the Bricklayer results, the examiner can see which words are better known than others, and thus can make inferences about which words may be only partially known.

The following section illustrates the program flow of Bricklayer from a user's point of view.

Bricklayer's Program Flow

Bricklayer begins by presenting the game board, a wall of blank "bricks." To the left of the board is a word bank, that is, a list of vocabulary items. Players are instructed and trained to "strengthen" the wall by dragging words from the word bank onto the bricks. In order to score well on the game players should put the words they know the best on the lowest rows. The game continues until the user fills up the board.

Figure 1 shows an example of a board that is full of words. Notice that there are more words than there are bricks. For this reason, some words (in this case, *baby*, *ball*, and *born*) are left behind on the bank.

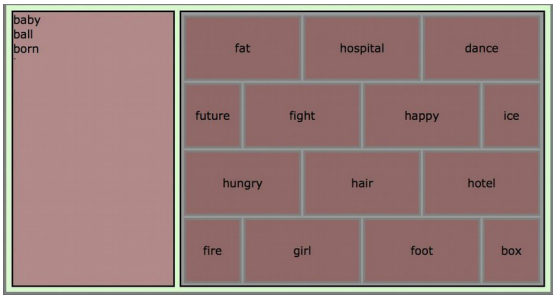


Figure 1. Players Fill the Board by Placing Words on Each Brick

After the board is full, the game goes into mini-quiz mode. At this point, starting from the top, one random brick per row lights up and the player is given a multiple-choice quiz for that word. For instance, Figure 2 displays the quiz for the word *fat*, which the user placed on the top left brick. The user must take the quiz in order to continue. If the player picks the correct definition for a word on a brick, the brick becomes solid. If the incorrect definition is chosen, the brick is destroyed.

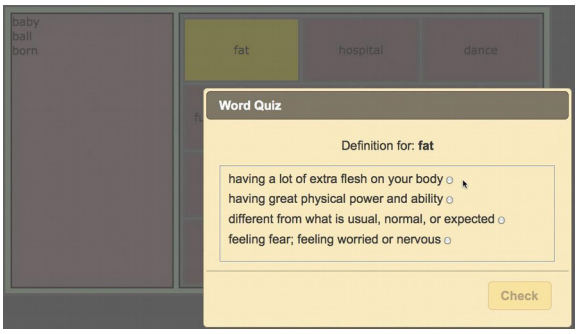


Figure 2. Players Choose the Definition of a Randomly Selected Word

Continuing on, the player is then quizzed on a random word on each of the following rows. The trick to Bricklayer is that if the player picks the wrong definition, not only is the brick that the word was on destroyed, but the bricks above the quizzed brick are also destroyed. The metaphor used in the game is that each brick needs to be supported by the bricks below it. For instance, in Figure 3, the player has incorrectly answered a quiz question for the brick *hotel*. Therefore, the player lost not only that brick but the four bricks above it.

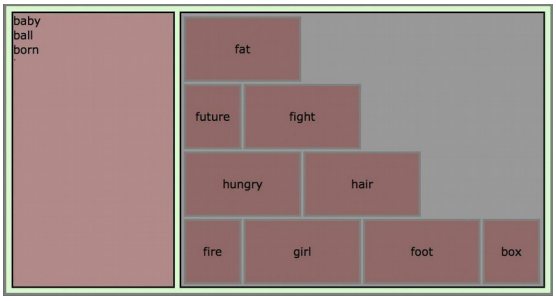


Figure 3. Visual Feedback for an Incorrect Quiz Choice

After the player has taken one quiz per row, the game ends, and the player gets points for all the bricks left on the board. The gamescore is presented as the percentage of all bricks remaining.

Each Bricklayer game presents the learner with a list of words. If this list has a range of words such that some are known, some are unknown, and some are partially known, then this application has the potential to measure not simply whether or not the player knows a word, but how well the player knows the word, at least in comparison to all words contained in the word bank. If a player places a word on the top row and then "loses" the word by an incorrect guess, not much is lost. Therefore, a player may "risk" putting an unknown word on the higher rows. However, maintaining solid bricks on the lowest rows is critical to success, since one wrong guess can knock out many bricks above. Therefore, the strategy for success in the game is for players to put the words they think they know the best at the bottom, words they know pretty well in the middle, and words they are less certain about near the top. For the purposes of the research study described below, participants were explicitly instructed in this strategy.

Research Questions

In order to assess the efficacy of Bricklayer, we conducted a study with 28 ESL learners who were tested on 72 words (Brumbaugh, 2015). For the purpose of this article, we report the results with regards to the following two research questions:

RQ 1: Does the learner's behavior in the Bricklayer game provide a way to accurately predict the learner's knowledge for that word?

RQ 2: Does the strength of this prediction provide a measurement for the learner's depth of knowledge for that word?

Methodology

Participants

28 ESL learners participated in the study which took place at a mid-sized Canadian university.[1] Study participants were recruited from the university's English for Academic Purposes program, which is a remedial ESL program for intermediate-level students seeking admission to the university. According to a background questionnaire, the study participants were about evenly split by gender, ranged from 17 to 21 years old, and had been studying English for an average of 7 years. The participants' English language skills were from lower to upper intermediate according to their self-reported IELTS test scores and student placement in the program. All participants were native speakers of non-Indo-European languages: Chinese (20 participants), Vietnamese (5 participants), and Turkish (1 participant).

Materials

Aside from the ethics release form which was provided as hard copy, all remaining study and assessment materials were presented to the participants sequentially on a web site. In addition to the background information questionnaire, materials included an instructional video, the computer program Bricklayer with the 72 words chosen for the study, and two post-tests.[2]

The two post-tests tested the learners' word knowledge for each of the 72 test items. They were divided equally into one of two post-test categories: a standard multiple-choice test and a semantic distance test. As discussed by Meara (1997) in the context of vocabulary acquisition, "[m]ultiple choice vocabulary tests, of the sort typically used to assess incidental learning, may not be sensitive enough to pick up what is going on [cumulative vocabulary acquisition]" (p. 119). For this reason, the semantic distance test was designed to measure gradations of word knowledge.

All test questions were made up of the correct word definition, and three distractors. Note that the definition for a given distractor was used instead of the distractor itself so that each distractor selection was the definition for an actual word. The definitions were all drawn from the Merriam-Webster Learner's Dictionary.

The multiple-choice test contained three distractors which were not semantically related to the correct answer. Table 1 provides an example of the answer and distractor set for the word *basket*, a word in the multiple-choice test condition.

Table 1
Sample Multiple-choice Quiz

Target word	Correct answer	Distractor 1	Distractor 2	Distractor 3
basket	a container usually made by weaving together long thin pieces of material	a covering for the hand that has separate parts for each finger	a strong building or group of buildings where soldiers live	a piece of cloth with a special design that is used as a symbol of a nation or group

The semantic distance test included distractors of varying semantic distance from the target word (Nagy et al., 1985). The choices for a semantic distance test contained the correct answer, for which a full score of 2 is given, two words with a strong semantic relationship to the target (for which a partial score of 1 is given), and two unrelated words, for which a score of 0 is given.

Table 2 provides an example of the answer and distractor set for the word *straw*, a word in the semantic distance test condition. Moreover, Table 2 also shows the word on which the distractor definition was based, although participants only saw the definitions. During Bricklayer gameplay, the same question/answer sets for the words were used as the mini-quizzes.

Table 2
Sample Semantic Distance Quiz

Target word	Correct answer	Distractor 1	Distractor 2	Distractor 3	Distractor 4
straw	the dry stems of wheat and other grain plants	(corn) the seeds of the corn plant eaten as a vegetable	(tractor) a large vehicle that has two large back wheels and two smaller front wheels and that is used to pull farm equipment	(tin) a soft, shiny, bluish-white metal that has many different uses	(sew) to make or repair something (such as a piece of clothing) by using a needle and thread

Data Collection

A video tutorial provided instructions to orient the participants to the Bricklayer game. The video emphasized the strategy for scoring well. Specifically, it showed the player that placing the words they know the best on the lowest row of the game board is the best strategy to minimize the risk of losing all the supported bricks due to a missed quiz question. After two practice games, study participants then played 8 rounds of Bricklayer as part of the research study. There were a total of 72 words, 18 on each board. Given that Bricklayer essentially forces students to rank word knowledge, each word was presented in two different boards because their rankings may depend on which other words are on the board. Finally, the participants took the two post-tests for all 72 words.

Findings

In order to examine whether the learner's placement of each word predicted his or her knowledge for that word (RQ 1), the results were modeled using two Rasch logistic regressions. First the multiple-choice test set was modeled, then, the semantic distance test set.

In the Rasch model, independent variables are referred to as *facets.* In this way, the effect of each individual item is calculated. The facets for the model reported here include all the scores that may have influenced the final prediction for word knowledge. The most important facet is called the *wordscore*; this is based on the word's final position on the board as placed by the learner. Another facet is the *gamescore* which measures how well the learner performed on that individual game. Which board was played (*board*) is included as a facet because the board difficulty may influence the prediction. Finally, *learner* and *word* are included as facets for the model because, in item response theory, word difficulty and learner ability each contribute a measure to the prediction (see Brumbaugh, 2015 for a more detailed analysis of the individual scoring values used in the research study).

The dependent variable is the value of the post-test score. Half of the data, selected randomly, was used as training data to assign weights to the facets. The other half was used for testing purposes to assess the validity of the weights. All results here are from the test set. The resulting prediction for each observation is referred to as the *target score*.

The Rasch model provides goodness-of-fit results for individual facets and so was used to provide an analysis of individual lexical items. These results are presented in the following section.

Rasch Model for Multiple-Choice Post-Test

First, the multiple-choice post-test group was modeled; the results shown in Table 3 indicate that all independent variables (i.e., learner, board, word, wordscore, and gamescore) exert a significant effect on the model except for the board. In the Rasch model, the chi-square values are tests of statistical significance and probability. These statistics are reported for each of the individual facets (degrees of freedom are given in brackets). Accordingly, Table 3 indicates that, for the fixed effects chi-square, the results are significant for learner (χ^2 (25) = 71.3, p < .01), word (χ^2 (35) = 140.2, p < .01), wordscore (χ^2 (6) = 14.7, p = .02), and gamescore (χ^2 (19) = 33.7, p = .02). In contrast, board did not have a significant effect (χ^2 (3) = 4.3, p = .23).

Table 3
Rasch Model Chi-Squared Statistics for Multiple-Choice Test Condition

Variable	Fixed chi-square[df]	Sig.	Random chi-square[df]	Sig.
Learner	71.3[25]	<.01*	19.2[24]	.74**
Board	4.3[3]	.23	1.8[2]	.41**
Word	140.2[35]	<.01*	28.6[34]	.73**
Wordscore	14.7[6]	.02*	4.3[5]	.51**
Gamescore	33.7[19]	.02*	12.0[18]	.85**

* Fixed chi-square is significant <.05 and indicates the probability that items are equal on a rating scale.: ** Random chi-square significance indicates the probability that these items could have been randomly sampled from a normal population.

The random chi-square results identify the probability that the items could have been sampled from a normal population. The highest probability is found with gamescore (χ^2 (18) = 12.0, p = .85), followed by learner (χ^2 (24) = 19.2, p = .74), word (χ^2 (34) = 28.6, p = .73), wordscore (χ^2 (5) = 4.3, p = .51), and board (χ^2 (2) = 1.8, p = .41).

The Rasch model can also be evaluated by means of a confusion matrix, which gives the accuracy of the model's predictions in percentages. Table 4 organizes the observed scores (the multiple choice post-test scores) in rows and the model predictions in columns. Once again, the model was based on observations – each prediction

was for a single instance of a learner/board/word/wordscore/gamescore combination. There were a total of 936 observations (half of the observations were used in the training set and half in the test set).

Table 4
Confusion Matrix for Rasch Results of Multiple-Choice Test Condition

Observation	Predicted 0	Predicted 1	No prediction*	% Correct
0 (unknown)	351**	140	15	69.4%
1 (known)	167	261**	2	60.7%
Total	518	401	17	65.4%

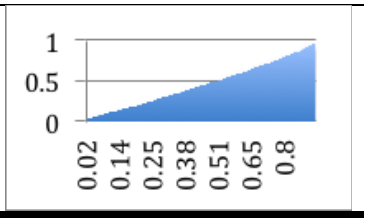
**Note.* There is no prediction for values of .5.
** Accurate predictions.

The first row shows the results for observed scores of 0, that is, cases in which an incorrect answer was given on the post-test. Of these incorrect words, 351 were accurately predicted to be incorrect and 140 were inaccurately predicted to be correct. There were 15 unknown words for which no prediction was made (see below for a discussion), thus the incorrect results were accurately predicted 69.4% of the time. In the next row, the words which were tested to be known are given. Of these, 167 words were inaccurately predicted to be unknown and 261 were accurately predicted to be known. There were 2 words with no prediction. The accuracy rate for known words was 60.7%. Overall, 518 words were predicted to be unknown, 401 were predicted to be known, and 17 words had no prediction. The overall accuracy rate of the model is 65.4%.

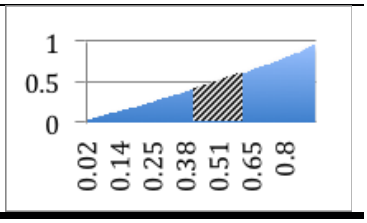
It is important to understand that although these predictions are presented as binary, the Rasch model actually generates an *expected value* which is between 0 and 1. In the case of the multiple-choice data, if the expected value is lower than .5, 0 is predicted. If it is above .5, 1 is predicted. At .5, the model makes no prediction; that is, there is an even probability that the word is known. Because this measurement is probabilistic, expected values close to the midpoint of .5 are less certain than values further from the midpoint (Bond & Fox, 2007). Accordingly, the further away the expected value is from the midpoint, the more accurate the prediction will be. Table 5 provides data to confirm this assumption. It shows a set of four confusion matrices for the Rasch multiple-choice results drawn from various ranges of expected values. In the first matrix, all results are modeled, and the predictions are 65.4% accurate. In the second matrix, data from the mid 20% of predictions are omitted, and the model is 69.1% accurate (although only 78.5% of the data are analyzed). The following two matrices model even less data but the overall predictions are more accurate. In the third matrix, the mid 40% of the predictions are omitted with an accuracy rate of 72.2%, and in the fourth matrix, the mid 60% of the predictions are omitted for an accuracy rate of 75.6%.

Table 5
Confusion Matrix: Various Prediction Levels Modeled

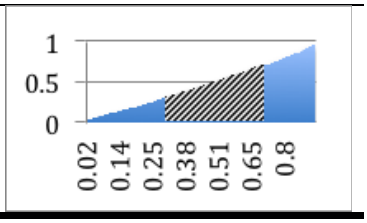
All results

Observed	Pred. 0	Pred. 1	None*	% Correct	% of data included	Range
0	351	140	15	69.4%		
1	167	261	2	60.7%		
		Total		65.4%	100%*	

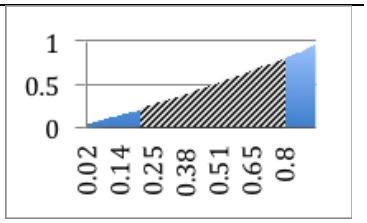
Excluding predictions from .41 to .60

Observation	Predicted 0	Predicted 1	% Correct	% of data included	Range
0	304	96	76.0%		
1	131	204	60.9%		
		Total	69.1%	78.5%	

Excluding predictions from .31 to .70

Observation	Predicted 0	Predicted 1	% Correct	% of data included	Range
0	228	54	80.9%		
1	94	157	62.5%		
		Total	72.2%	56.9%	

Excluding predictions from .21 to .80

Observation	Predicted 0	Predicted 1	% Correct	% of data included	Range
0	158	23	87.3%		
1	62	105	62.9%		
		Total	75.6%	37.2%	

Note. *There is no prediction for values of .5.

Rasch Model for Semantic Distance Post-Test

This section gives the results of the partial credit Rasch logistic regression that was performed on the semantic distance test data to examine whether the target scores predicted the learners' depth of semantic knowledge for words (RQ 2). The facets for the model reported here are the same independent variables (i.e., learner, board, word, wordscore, gamescore) used for the Rasch model of the previous multiple choice test analysis. The dependent variable is the value of the semantic distance post-test score. In this case, the partial credit model

developed by Masters (1982) is used, since the middle scores in the semantic distance post-test correspond to partial knowledge.

As in the previous model, half of the data was used to train the model and the other half was used for testing purposes; all results reported here are from the test set. In order to reduce the level of factoring in the model, the results of the semantic distance test were binned into three groups rather than the original five scores, and then converted to integers (for the purposes of the modeling software).

The chi-square tests of statistical significance and probability are reported in Table 6 for each of the individual facets (degrees of freedom are given in brackets). Accordingly, Table 6 indicates that, for the fixed chi-square, the results are significant for learner (χ^2 (25) = 59.3, p =< .01), word (χ^2 (35) = 161.2, p < .01), and gamescore (χ^2 (18) = 44.9, p < .01). Neither board (χ^2 (3) = 6.3, p = .10) nor wordscore (χ^2 (6) = 11.9, p = .06) had a significant effect on the model. As for the random chi-square, learner (χ^2 (24) = 17.7, p = .82) and gamescore (χ^2 (18) = 12.6, p = .82) have the highest probability of having been sampled from a normal population, followed by word (χ^2 (34) = 29.0, p = .71), wordscore (χ^2 (5) = 3.9, p = .56), and board (χ^2 (2) = 2.0, p = .36).

Table 6
Rasch Model Chi-Squared Statistics for Semantic Distance Condition

Variable	Fixed chi-squared[df]	Sig.	Random chi-square[df]	Sig.
Learner	59.3[25]	<.01*	17.7[24]	.82**
Board	6.3[3]	.10	2.0[2]	.36**
Word	161.2[35]	<.01*	29.0[34]	.71**
Wordscore	11.9[6]	.06	3.9[5]	.56**
Gamescore	44.9[18]	<.01*	12.6[18]	.82**

* Fixed chi-square is significant <.05 and indicates the probability that items are equal on a rating scale.
** Random chi-square significance indicates the probability that these items could have been randomly sampled from a normal population.

Table 7 shows the confusion matrix results for the semantic distance condition. In this case, if the model reports a strong probability that the word is known, the prediction corresponds to full knowledge. A lower probability for word knowledge corresponds to partial knowledge, and a low probability for word knowledge corresponds to no knowledge.

Table 7
Confusion Matrix for Rasch Results for Semantic Distance Condition

Observation	Predicted 0	Predicted 1	Predicted 2	% Correct
0 (unknown)	143**	112	44	47%
1 (partial)	135	101**	87	31%
2 (known)	62	91	161**	51%
			Total correct	43%

** Accurate predictions.

As in the case of the previous confusion matrix displayed in Table 4, each row in Table 7 contains the results for an observed score. When the score was observed to be 0 (the participant selected an unrelated distractor), the model accurately predicted an incorrect score 143 times, inaccurately predicted partial knowledge 112 times, and full knowledge 44 times, for an accuracy rate of 47%. Words which were observed to be partially known (the participant selected a semantically similar distractor) were inaccurately predicted to be unknown 135 times, accurately predicted to be partially known 101 times, and inaccurately predicted to be known 87 times for an accuracy rate of 31%. Words which were observed to be known (the participant selected the correct

definition) were inaccurately predicted to be unknown 62 times and partially known 91 times; 161 times they were accurately predicted to be known for an accuracy rate of 51%. The model's overall accuracy rate was 43%.

In summary, applying the partial credit Rasch model to the semantic distance test results weakens the predictive power (which was shown in the multiple choice test set) to approximately chance values (43%), implying that the game does not confidently predict the learners' depth of semantic knowledge for words.[3] Moreover, in the Rasch analysis, which included words and learners as factors by taking into account the difficulty of each word as well as the ability level of each learner, the three independent variables learner, word, and gamescore are significant factors, in contrast to wordscore and board which are *insignificant.*

Discussion

In Messick's (1989) seminal article on test validation, he emphasizes the importance of *test use* in validating an assessment instrument. It is not enough to ask whether a test measures what it purports to measure, but it must also be considered whether the results are appropriate to the particular purpose for which the test was designed. Bricklayer was designed to generate a learner model in the context of an ITS. Thus, it is appropriate to discuss the results of this study in that context.

Unlike a teacher (who may be attuned to the general ability level of his or her students), an ITS could construct a detailed model of a learner's lexical knowledge. In a vocabulary ITS, an assessment tool would be used to seed this model. This learner model should be dynamic, adjusting instruction to the learner's behaviours and knowledge states as they evolve and manifest themselves during system use. Such a model is similar to the one described by Mislevy, et al. (2002), which "refers to a piece of machinery: a set of variables in a probability model, for accumulating evidence about students" (p. 482).

Brumbaugh (2015) compared Bricklayer's results to a standard checkbox assessment, which has also been used in an ITS (Rosa & Eskenazi, 2013), and found that the checkbox assessment fared slightly better overall than the Bricklayer assessment for the multiple-choice word set when words were binned into two (*known* or *unknown*) categories.

However, the Bricklayer prediction model also reports the *probability* that a word is known. Examining the results more deeply, Bricklayer does a better job of modeling the "edge conditions" – words which are strongly predicted to be known or unknown. This is shown by the analysis in Table 5 which models various prediction levels. The two assessments may therefore be better suited for different tasks. The checkbox offers a quick way to make assessments for a lot of words thereby suggesting that the checkbox test would be useful for breadth assessments, or evaluations which require comparisons between students. In contrast, Bricklayer is more accurate at identifying words which are either very likely known or unknown; in an ITS environment the remaining words at the middle ranges might be considered frontier words which merit attention. Even the fact that words in this predictive range are just as likely to be known as unknown might turn out to be indicative of frontier knowledge. A word on the edge of acquisition may be subject to inconsistent test results as the memory trace for the word may be incomplete or not always accessible.

At this point, an ITS could provide additional focused tasks for these words, for example, readings, games, quizzes, concordance exercises, and other activities. Subsequent learner behaviour such as clicking a word to look up the meaning or correctly answering a cloze activity would then present opportunities to update the learner model with more precise information for these words. In other words, Bricklayer's assessment results are not considered definitive, but rather one piece of data in the larger construction of a learner model.

There are some shortcomings of Bricklayer which might be addressed in order to improve its performance, as well as possible limitations in the experimental design of the current study which may have adversely affected the statistical results.

Levels of Word Knowledge

Attempting to map the placement of the words on the bricks onto a continuum of word knowledge was exacerbated by the fact that the game offered seven levels of rankings and yet only three levels of knowledge were actually captured (Brumbaugh, 2015). This is consistent with previous research (Schmitt, 1998) which tried to capture gradations of knowledge but found that slight increases were difficult to measure. Future modifications to this assessment tool should reconsider the number of wordscore categories available, perhaps drawing from Horst and Meara's (1999) matrix model of lexical growth or a modified version of the VKS (Wesche & Paribakht, 1996). It is worth noting, however, that even with only three degrees of resolution, Bricklayer provides a mechanism for predicting knowledge that goes beyond the known/unknown dichotomy currently available in the widely-used version of the checkbox assessment. Such a rubric would furthermore contain structural validity, in that it is consistent with a model of word knowledge as being either known, unknown, or partial.

Computer Adaptation

Neither Bricklayer, nor any game based on this forced-choice ranking model, will ever meet its full potential until it is able to adapt to the word knowledge of the learner. Games with word sets which are either all known or all unknown simply do not do a good job of distinguishing knowledge. Due to the challenges of programming and data analysis, an adaptive study was not feasible for this initial research.

In an ITS context, such computer adaptive testing techniques are used generally to target content to individual learners (Beatty, 2010). Furthermore, a student model that keeps a record of student behavior and performance could ultimately track not only lexical knowledge and ability levels for students, but student "fit" to the model as well.

Better Assessment of Partial Knowledge

Bricklayer is designed to capture partial knowledge. In order to validate this construct, it was necessary to compare Bricklayer's results to a separate measure of partial knowledge. However, there are challenges to this approach. Partial lexical knowledge is a complex construct which is notoriously difficult both to define and to measure (Schmitt, 2014). Thus, the partial-credit Rasch prediction did not predict partial semantic knowledge, but it is impossible to tell from the results of this study whether or not Bricklayer was sensitive to different aspects of partial knowledge, such as collocations, polysemy, or degrees of receptive/productive knowledge. Indeed, Bricklayer produced some interesting results for polysemous items, and yet the limitations of the post-test forced some speculation. An example of such a word is *fare*. The sense used in the post-test is the main dictionary definition entry, *to do something well or badly*. However, the participants, as temporary international students and thus frequent users of public transportation, likely had repeated exposure to the word *fare* in the sense of *the money paid for public transportation*. Indeed, it turns out that on one of the game boards, 15 of the students (58%) put this word on the lower two rows of the board but then provided a wrong answer on the post-test. They thought they knew the word, but fared poorly on the test. A more thorough post-test, using a polysemous testing instrument such as that used by Qian (2002) and Read (1993, 1995), or one-on-one interviews with the participants, would be necessary to better interpret these results. Such an improved post-test measure might well improve the partial-credit Rasch predictions.

Conclusion

This research study introduced Bricklayer, an assessment tool which can identify *strongly* known and unknown words, and which can suggest which words might be on the frontier of acquisition. An analysis of the results also ascertained ways in which the tool's performance might be improved by fine-tuning the scoring rubric and by using computer adaptive testing techniques to customize game boards for each learner.

Bricklayer, which presents a new paradigm for L2 vocabulary assessment, connects with research on vocabulary acquisition by providing a mechanism to capture partial word knowledge. While Bricklayer was the

primary focus of the empirical investigation, the original contributions to the vocabulary assessment field are not about Bricklayer *per se*, but rather about some fundamental characteristics unique to Bricklayer. From this perspective, Bricklayer is a working exemplar of a novel self-assessment paradigm.

Bricklayer essentially presents learners with the meta-cognitive task to rank a list of words according to how well they know them. This differs in a qualitative way from typical self-assessments which force a binary choice. Learners must consider not just whether they know a word, but how well they know it. It is possible that this leads to a deeper level of cognitive reflection. In the Bricklayer study, participants only spent about a minute in total on the three screens of checkbox items (n=24 words); in contrast, they spent on average 2 ½ minutes on each game (n=18 words per game). This may indicate that they were giving more focused attention to the game task.

There are certain drawbacks to ranking data. Primarily, if two words are equally known or equally unknown, the ranking data are not useful. This could be mitigated in several ways in task design. For example, participants might be presented with two or three words and then instructed to rank them in terms of knowledge. In a computer interface, this could be achieved by dragging the words into an ordered list. Words could be repeated in different contexts and then results subjected to an item response analysis such as Rasch. Alternately, the participant could simply report, for example, by pressing a button on a screen, that both words are known or both are unknown.

From a quantitative point of view, measurements derived from rankings provide a mechanism for sensitivity to partial lexical knowledge. Implementing such a modification to the standard self-assessment tools might result in more robust results with a higher level of structural validity.

Currently, vocabulary assessment falls into two broad categories: traditional tests in which the learner must select or give the correct answer, and checkbox self-assessments in which the test administrator must either rely on the learner's response or depend on pseudowords to gauge the learner's accuracy. The assessment paradigm on which Bricklayer is based offers a third option: random spot-checks of learners' self-assessments. The mini-quizzes in the game serve three important functions. Firstly, they give a way to validate the learner's responses. Secondly, they provide accountability to the learners – since they know the test may be coming, they have a reason not to misrepresent their knowledge. Finally, they provide a mechanism for clarifying the expectation about what type of word knowledge is being tested.

There are typically three uses for assessment: evaluation, instruction, and research. In a context in which a student is being evaluated for aptitude for a given program or in which learning gains for a course are being assessed, Bricklayer's probabilistic results might be too subtle to accomplish the test purpose. However, in instructional contexts, such as a classroom or ITS environment, Bricklayer's paradigm might be well-suited to identify frontier words which would benefit from further, direct instruction.

Endnotes

1. Two of the participants were excluded from the final data analysis due to incorrect usage of the software which may have corrupted the results.
2. It should be noted that these are post-tests in the sense that they are taken *after* the main part of the study for the purposes of collecting data for concurrent validity; this study did not use a pre-test/post-test design.
3. Interestingly, although the results could not *predict* partial knowledge, deeper analysis of the data showed that Bricklayer was *sensitive* to this knowledge (Brumbaugh, 2015).

References

Anderson, R. C., & Freebody, P. (1983). Reading comprehension and the assessment and acquisition of word kowledge. In B. Huston (Ed.), *Advances in Reading/Language Research* (Vol. 2, pp. 231–256). Greenwich, CT: JAI Press.

Beatty, K. (2010). *Teaching and researching computer-assisted language learning* (2nd ed.). Harlow, England ; New York: Longman.

Beeckmans, R., Eyckmans, J., Janssens, V., Dufranne, M., & Van de Velde, H. (2001). Examining the Yes/No Vocabulary Test: Some Methodological Issues in Theory and Practice. *Language Testing*, *18*(3), 235–274.

Bond, T. G., & Fox, C. M. (2007). *Applying the Rasch Model: Fundamental Measurement in the Human Sciences, Second Edition* (2nd ed.). Routledge.

Brumbaugh, H. (2015). *Self-assigned ranking of L2 vocabulary: using the Bricklayer computer game to assess depth of word knowledge* (Doctoral dissertation, Arts & Social Sciences:). Retrieved from http://summit.sfu.ca/item/15287.

Collins-Thompson, K., & Callan, J. (2007). Automatic and Human Scoring of Word Definition Responses. In C. L. Sidner, T. Schultz, M. Stone, & C. Zhai (Eds.), *HLT-NAACL* (pp. 476–483). The Association for Computational Linguistics.

Crossley, S., Salsbury, T., & McNamara, D. (2010). The Development of Polysemy and Frequency Use in English Second Language Speakers. *Language Learning*, *60*(3), 573–605.

Dolch, E. W. (1927). *Reading and word meanings,*. Ginn and company.

Durso, F. T., & Shore, W. J. (1991). Partial knowledge of word meanings. *Journal of Experimental Psychology: General*, *120*(2), 190–202. http://doi.org/10.1037/0096-3445.120.2.190

Eyckmans, J. (2004). *Measuring receptive vocabulary size: reliability and validity of the yes/no vocabulary test for French-speaking learners of Dutch*. Utrecht: LOT.

Heift, T., & Schulze, M. (2003). Student Modeling and ab initio Language Learning. *System*, *31*(4), 519–535.

Henriksen, B. (1999). Three Dimensions of Vocabulary Development. *Studies in Second Language Acquisition*, *21*(2), 303–317.

Horst, M., & Meara, P. (1999). Test of a model for predicting second language lexical growth through reading. *Canadian Modern Language Review/La Revue Canadienne Des Langues Vivantes*, *56*(2), 308–328.

Laufer, B., & Goldstein, Z. (2004). Testing Vocabulary Knowledge: Size, Strength, and Computer Adaptiveness. *Language Learning*, *54*(3), 399–436.

Laufer, B., & Nation, I. S. P. (2001). Passive Vocabulary Size and Speed of Meaning Recognition: Are They Related? *EUROSLA Yearbook*, *1*, 7–28.

LeBlanc, R., & Painchaud, G. (1985). Self-Assessment as a Second Language Placement Instrument. *Tesol Quarterly*, *19*(4), 673-687.

Masters, G. N. (1982). A Rasch Model for Partial Credit Scoring. *Psychometrika*, *47*(2), 149–74.

Meara, P. (1990). Some notes on the Eurocentres Vocabulary Tests. In J. Tommola (Ed.), *Vieraan kielen ymmärtäminen ja tuottaminen (Foreign Language Comprehension and Production)* (pp. 103–113). Turku: Suomen Soveltavan Kielitieteen Yhdistys AFinLA.

Meara, P. (1997). Towards a new approach to modelling vocabulary acquisition. In N. Schmitt & M. McCarthy (Eds.), *Vocabulary: Description, Acquisition and Pedagogy* (pp. 109–121). Cambridge, UK: Cambridge University Press.

Meara, P. (2009). *Connected Words: Word Associations and Second Language Vocabulary Acquisition*. Amsterdam: John Benjamins Pub. Co.

Meara, P., & Buxton, B. (1987). An Alternative to Multiple Choice Vocabulary Tests. *Language Testing*, *4*(2), 142–154.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational Measurement* (3rd ed, pp. 13–103). Washington, D.C.: American Council on Education.

Milton, J., & Vassiliu, P. (2000). Frequency and the lexis of low-level EFL texts. In *Proceedings of the 13th Symposium in Theoretical and Applied Linguistics, Aristotle University of Thessaloniki* (pp. 444–55).

Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (2002). Design and Analysis in Task-Based Language Assessment. *Language Testing*, *19*(4), 477–496.

Mochida, A., & Harrington, M. (2006). The Yes/No test as a measure of receptive vocabulary knowledge. *Language Testing*, *23*(1), 73–98. http://doi.org/10.1191/0265532206lt321oa

Nagy, W. E., Herman, P. A., & Anderson, R. C. (1985). Learning Words from Context. *Reading Research Quarterly*, *20*(2), 233–253.

Nation, I. S. P. (2013). *Learning Vocabulary in Another Language* (Second). Cambridge: Cambridge University Press.

Nation, I. S. P. (1983). Testing and teaching vocabulary. *Guidelines*, *5*(1), 12–25.

Nation, I. S. P. (1990). *Teaching and Learning Vocabulary*. New York: Newbury House Publishers.

Nation, I. S. P. (2001). *Learning Vocabulary in Another Language*. Cambridge: Cambridge University Press.

Nation, I. S. P., & Beglar, D. (2007). A vocabulary size test. *The Language Teacher*, *31*(7), 9–13.

Paribakht, T. S., & Wesche, M. (1997). Vocabulary enhancement activities and reading for meaning in second language vocabulary acquisition. In J. Coady & T. N. Huckin (Eds.), *Second Language Vocabulary Acquisition: A Rationale for Pedagogy* (pp. 175–200). Cambridge, U.K: Cambridge University Press.

Pellicer-Sánchez, A., & Schmitt, N. (2012). Scoring Yes–No vocabulary tests: Reaction time vs. nonword approaches. *Language Testing*, *29*(4), 489–509. http://doi.org/10.1177/0265532212438053

Qian, D. D. (2002). Investigating the Relationship between Vocabulary Knowledge and Academic Reading Performance: An Assessment Perspective. *Language Learning*, *52*(3), 513–536.

Read, J. (1993). The Development of a New Measure of L2 Vocabulary Knowledge. *Language Testing*, *10*(3), 355–371.

Read, J. (1995). Validating the word associates format as a measure of depth of vocabulary knowledge. In *17th language testing research colloquium, Long Beach, CA*.

Rosa, K. D., & Eskenazi, M. (2013). Self-Assessment in the REAP Tutor: Knowledge, Interest, Motivation, & Learning. *International Journal of Artificial Intelligence in Education (IOS Press)*, *21*(4), 237–253.

Richards, J. C. (1976). The Role of Vocabulary Teaching. *TESOL Quarterly*, *10*(1), 77–89.

Schmitt, N. (1998). Tracking the Incremental Acquisition of Second Language Vocabulary: A Longitudinal Study. *Language Learning*, *48*(2), 281–317.

Schmitt, N. (2014), Size and Depth of Vocabulary Knowledge: What the Research Shows. *Language Learning*, 64: 913–951.

Schmitt, N., & Meara, P. (1997). Researching Vocabulary through a Word Knowledge Framework: Word Associations and Verbal Suffixes. *Studies in Second Language Acquisition*, *19*(1), 17–36.

Schmitt, N., Schmitt, D., & Clapham, C. (2001). Developing and Exploring the Behaviour of Two New Versions of the Vocabulary Levels Test. *Language Testing*, *18*(1), 55–88.

Shore, W. J., & Kempe, V. (1999). The Role of Sentence Context in Accessing Partial Knowledge of Word Meanings. *Journal of Psycholinguistic Research*, *28*(2), 145–163. http://doi.org/10.1023/A:1023258224980

Webb, S. (2005). Receptive and Productive Vocabulary Learning: The Effects of Reading and Writing on Word Knowledge. *Studies in Second Language Acquisition*, *27*(1), 33–52.

Wesche, M., & Paribakht, T. S. (1996). Assessing Second Language Vocabulary Knowledge: Depth versus Breadth. *The Canadian Modern Language Review/La Revue Canadienne Des Langues Vivantes*, *53*(1), 13–40.

Zimmerman, C. B. (1997). Historical trends in second language vocabulary instruction. In J. Coady & T. N. Huckin (Eds.), *Second Language Vocabulary Acquisition: A Rationale for Pedagogy* (pp. 5–19). Cambridge, U.K: Cambridge University Press.

About The Authors

After receiving her PhD from Simon Fraser University in 2015, *Heidi Brumbaugh* founded Vocabulary Systems, Inc., where she is currently developing a suite of vocabulary learning games students can play on their smartphones.

Trude Heift is Professor of Linguistics in the Department of Linguistics at Simon Fraser University, Canada. Her research focuses on the design and evaluation of CALL systems with a particular interest in learner-computer interactions and learner language. She is co-editor of *Language Learning & Technology*.