Predicting Short- and Long-Term Vocabulary Learning via Semantic Features of Partial Word Knowledge

SungJin Nam
School of Information
University of Michigan
Ann Arbor, MI 48109
sjnam@umich.edu

Gwen Frishkoff
Department of Psychology
University of Oregon
Eugene, OR 97403
gfrishkoff@gmail.com

Kevyn Collins-Thompson
School of Information
University of Michigan
Ann Arbor, MI 48109
kevynct@umich.edu

ABSTRACT

We show how the novel use of a semantic representation based on Osgood's semantic differential scales can lead to effective features in predicting short- and long-term learning in students using a vocabulary learning system. Previous studies in students' intermediate knowledge states during vocabulary acquisition did not provide much information on which semantic knowledge students gained during word learning practice. Moreover, these studies relied on human ratings to evaluate the students' responses. To solve this problem, we propose a semantic representation for words based on Osgood's semantic decomposition of vocabulary [16]. To demonstrate our method can effectively represent students' knowledge in vocabulary acquisition, we build models for predicting the student's short-term vocabulary acquisition and long-term retention. We compare the effectiveness of our Osgood-based semantic representation to that provided by Word2Vec neural word embedding [13], and find that prediction models using features based on Osgood scale-based scores (OSG) perform better than the baseline and are comparable in accuracy to those using Word2Vec score-based models (W2V). By using more interpretable Osgood-based scales, our study results can help with better understanding of students' ongoing learning states and designing personalized learning systems that can address an individual's weak points in vocabulary acquisition.

Keywords

Vocabulary learning, semantic similarity, prediction model, intelligent tutoring system

1. INTRODUCTION

Studies of word learning have shown that knowledge of individual words is typically not all-or-nothing. Rather, people acquire varying degrees of knowledge of many words incrementally over time, by exposure to them in context [9]. This is especially true for so-called "academic" words that are less common and more abstract — e.g., *pontificate*, *probity*, or *assiduous* [7]. Binary representations and measures model word knowledge simply as correct or incorrect on a particular item (word), but in reality, a student's knowledge level may reside between these two extremes. Thus, previous studies of vocabulary acquisition have suggested that students' partial knowledge be modeled using a representation that adding an additional label corresponding to an intermediate knowledge state [6] or further, in terms of continuous metrics for semantic similarity [3].

In addition, there are multiple dimensions to a word's meaning [16]. Measuring a student's partial knowledge on a single scale may only provide abstract information about the student's general answer quality and not give enough information to specify *which* dimensions of word knowledge a student already has learned or needs to improve. In order to achieve detailed understanding of a student's learning state, online learning systems should be able to capture a student's "learning trajectory" that tracks their partial knowledge on a particular item over time, over multiple dimensions of meaning in a multidimensional semantic representation.

Hence, multidimensional representations of word knowledge can be an important element for building an effective intelligent tutoring system (ITS) for reading and language. Maintaining a fine-grained semantic representation of a student's degree of word knowledge can be helpful for the ITS to design more engaging instructional content, more helpful personalized feedback, and more sensitive assessments [17, 19]. Selecting semantic representations to model, understand, and predict learning outcomes is important to designing a more effective and efficient ITS.

In this paper, we explore the use of multidimensional semantic word representations for modeling and predicting short- and long-term learning outcomes in a vocabulary tutoring system. Our approach derives predictive features using a novel application of existing methods in cognitive psychology combined with methods from natural language processing (NLP). First, we introduce a new multidimensional representation of a word based on the Osgood semantic differential [16], an empirically based, cognitive framework that uses a small number of scales to represent latent components of word meaning. We compare the effectiveness of model features based on this Osgood-based representation to features based on a different representation, the widely-used Word2Vec word embedding [13]. Second, we evaluate our prediction models using data from a meaning-generation task that was conducted during a computer-based intervention. Our study results demonstrate how similarity-based metrics based on rich

semantic representation can be used to automatically evaluate specific components of word knowledge, track changes in the student's knowledge toward the correct meaning, and compute a rich set of features for use in predicting short- and long-term learning outcomes. Our methods could support advances in real-time, adaptive support for word semantic learning, resulting in more effective personalized learning systems.

2. RELATED WORK

The present study is informed by three areas of research: (1) studies of partial word knowledge; (2) the Osgood framework for multiple dimensions of word meaning, and (3) computational methods for estimating semantic similarity.

Partial Word Knowledge. The concept of partial word knowledge has interested vocabulary researchers for several decades, particularly in the learning and instruction of "Tier 2" words [20]. Tier 2 words are low-frequency and typically have complex (multiple, nuanced) meanings. By nature, they are rarely learned through "one-shot" learning or direct definition. Instead, they are learned partially and gaps are filled in over time.

Words in this intermediate state, neither novel nor fully known, are sometimes called "frontier words" [5]. Durso and Shore operationalized the frontier word as a word the student had seen previously but was not actively using it [6]. Based on this definition, the student may have had implicit memory of frontier words, such as general information like whether the word indicates a good or bad situation or refers a person or an action. They discovered that students are more familiar with frontier words than other types of words in terms of their sounds and orthographic characteristics [6]. This previous work suggested that the concept of frontier words can be used to represent a student's partial knowledge states in a vocabulary acquisition task [5, 6].

In some studies, partial word knowledge has been represented using simple, categorical labels, e.g., multiple-choice tests that include "partially correct" response options, as well as a single "best" (correct) response. In other studies, the student is presented with a word and is asked to say what it means [1]. The definition is given partial credit if it reflects knowledge that is partial or incomplete. For example, a student may recognize that the word *probity* has a positive connotation, even if she cannot give a complete definition. However, single categorical or score-based indicators may not explain which specific aspects of vocabulary knowledge the student is missing. Moreover, these studies relied on human ratings to evaluate students' responses for unknown words [6]. Although widely used in psychometric and psycholinguistic studies [4, 16], hiring human raters is expensive and may not be done in real time during students' interaction with the tutoring system.

To address these problems, we propose a data-driven method that can automatically extract semantic characteristics of a word based on a set of relatively simple, interpretable scales. The method benefits from existing findings in cognitive psychology and natural language processing. In the following sections, we illustrate more details of related findings and how they can be used in an intelligent tutoring system setting.

Semantic Representation & the Osgood Framework. To quantify the semantic characteristics of a student's intermediate knowledge of vocabulary, this paper uses a "spatial analogue" for capturing semantic characteristics of words. In [16], Osgood investigated how the meaning of a word can be represented by a series of general semantic scales. By using these scales, Osgood suggested that the meanings of any word can be projected and explored in a continuous semantic space.

Osgood asked human raters to evaluate a set of words using a large number of scales (e.g., tall-short, fat-thin, heavy-light) and captured the semantic representation of a word [16]. Respondents gave Likert ratings, which indicated whether they thought that a word meaning was closer to one extreme (-3) or the other (+3), or basically irrelevant (0). A principal components analysis (PCA) was used to represent the latent semantic features that can explain the patterns of response to individual words within this task.

In our study, we suggest a method that can automatically extract similar semantic information that can project a word into a multidimensional semantic space. By using semantic scales selected from [16], we verify if such representation of semantic attributes of words is useful for predicting students' short- and long-term learning.

Semantic Similarity Measures. Studies in NLP have suggested methods to automatically evaluate the semantic association between two words. For example, Markov Estimation of Semantic Association (MESA) [3, 9] can estimate the similarity between words from a random walk model over a synonym network such as WordNet [14]. Other methods like latent semantic analysis (LSA) are based on co-occurrence of the word in a document corpus. In LSA, semantic similarity between words is determined by using a cosine similarity measure, derived from a sparse matrix constructed from unique words and paragraphs containing the words [10].

For this paper, we use Word2Vec [13], a widely used word embedding method, to calculate the semantic similarity between words. Word2Vec's technique [11] transforms the semantic context, such as proximity between words, into a numeric vector space. In this way, linguistic regularities and patterns are encoded into linear translations. For example, using outputs from Word2Vec, relationships between words can be estimated by simple operations on their corresponding vectors, e.g., *Madrid - Spain + France = Paris*, or *Germany + capital = Berlin* [13].

Measures from these computational semantic similarity tools are powerful because they can provide an automated method for evaluation of partial word knowledge. However, they typically produce a single measure (e.g., cosine similarity or Euclidean distance), representing semantic similarity as a one-dimensional construct. With such a measure, it is not possible to determine represent partial semantic knowledge and changes in knowledge of latent semantic features as word knowledge progresses from unknown to frontier to fully known. In following sections, we describe how we address this problem, using novel methods to to estimate the contribution of Osgood semantic features to individual word meanings.

2.1 Overview of the Study

Based on findings from existing studies, this study will suggest an automatized method for evaluating students' partial knowledge of vocabulary that can be used to predict students' short-term vocabulary acquisition and long-term retention. To investigate this problem, we will answer the following research questions with this paper.

The first research question (RQ1): Can semantic similarity scores from Word2Vec be used to predict students' short-term learning and long-term retention? Previous studies in vocabulary tutoring systems tend to focus on how different experimental conditions, such as different spacing between question items [18], difficulty levels [17], and systematic feedback [7], affect students' short-term learning. This study will answer how computationally estimated trial-by-trial scores in a vocabulary tutoring system can be used to predict students' short-term learning and long-term retention.

RQ2: Compared to using regular Word2Vec scores, how does the model using Osgood's semantic scales [16] as features perform for immediate and delayed learning prediction tasks? As described in the previous section, the initial outcome from Word2Vec returns hundreds of semantic dimensions to represent the semantic characteristics of a word. Summary statistics for comparing such high-dimensional vectors, such as cosine similarity or Euclidean distance, only provide the overall similarity between words. If measures from Osgood scales work in a similar level to models using regular Word2Vec scores for predicting students' learning outcomes, we can argue that it can be an effective method for representing students' partial knowledge of vocabulary.

3. METHOD

3.1 Word Learning Study

This study used a vocabulary tutoring system called Dynamic Support of Contextual Vocabulary Acquisition for Reading (DSCoVAR) [8]). DSCoVAR aims to support efficient and effective learning vocabulary in context. All participants accessed DSCoVAR in a classroom-setting environment by using Chromebook devices or the school's computer lab in the presence of other students.

3.1.1 Study Participants

Participants included 280 middle school students (6th to 8th grade) from multiple schools, including children from diverse socio-economic and educational backgrounds. Table 1 provides a summary of student demographics, including location (P1 or P2), age and grade level, sex. Location P1 is a laboratory school affiliated with a large urban university in the northeastern United States. Students from location P1 were generally of high socio-economic status (e.g., children of University faculty and staff). Location P2 includes three public middle schools in a southern metropolitan area of the United States. All students from location P2 qualified for free or reduced lunch. The study included a broad range of students so that the results of this analysis were more likely to generalize to future samples.

3.1.2 Study Materials

DSCoVAR presented students with 60 SAT-level English words (also known as Tier 2 words). These "target words," lesser-known words that the students are going to learn,

Table 1: The number of participants by grade and gender

	6th grade		7th grade		8th grade	
Group	Girl	Boy	Girl	Boy	Girl	Boy
P1	16	28	19	23	18	13
P2	53	51	12	6	21	20

were balanced between different parts of speech, including 20 adjectives, 20 nouns, and 20 verbs. Based on previous works, we expected that students would have varying degrees of familiarity with the words at pre-test, but that most words would be either completely novel ("unknown") or somewhat familiar ("partially known") [8, 15]. This selection of materials ensured that there would be variability in word knowledge across students for each word and across words for each student.

In DSCoVAR, students learned how to infer the meaning of an unknown word in a sentence by using surrounding contextual information. Having more information in a sentence (i.e., a sentence with a high degree of contextual constraint) can decrease the uncertainty of inference. In this study, the degree of sentence constraint was determined using standard cloze testing methods: quantifying the diversity of responses from 30 human judges when the target word is left as a fill-in-the-blank question.

3.1.3 Study Protocol

The word learning study comprised four parts: (1) a pre-test, which was used to estimate baseline knowledge of words, (2) a training session, where learners were exposed to words in meaningful contexts, (3) an immediate post-test, and (4) a delayed post-test, which occurred approximately one week after training.

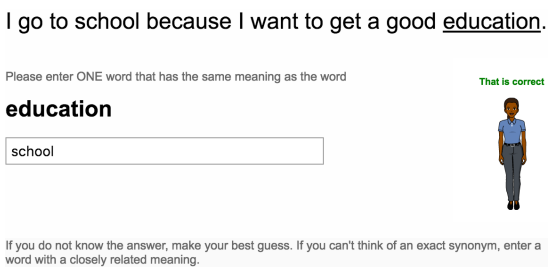
Pre-test. The pre-test session was designed to measure the students' prior knowledge of the target words. For each target word, students were asked to answer two types of questions: familiarity-rating questions and synonym selection questions. In familiarity rating questions, students provided their self-rated familiarity levels (unknown, known, and familiar) for presented target words. In synonym-selection questions, students were asked to select a synonym word for the given target word from five multiple choice options. The outcome from synonym-selection questions provided more objective measures for students' prior domain knowledge of target words.

Training. Approximately one week after the pre-test session, students participated in the training. During training, students learned strategies to infer the meaning of an unknown word in a sentence by using surrounding contextual information.

A training session consisted of two parts: an instruction video and practice questions. In the instruction video, students saw an animated movie clip about how to identify and use contextual information from the sentence to infer the meaning of an unknown word. In the practice question part, students could exercise the skill that they learned from the video. DSCoVAR provided sentences that included a target word with different levels of surrounding contextual information. The amount of contextual information for each sentence was determined by external crowd workers (details described in Section 3.1.2). In the practice question part, each target word was presented four times within

different sentences. Students were asked to type a synonym of the target word, which was presented in the sentence as underlined and bold. Over two weeks, students participated in two training sessions with a week's gap between them. Each training session contained the instruction video and practice questions for 30 target words. An immediate post-test session followed right after each training session.

Figure 1: An example of a training session question. In this example, the target word is "education" with a feedback message for a high-accuracy response.



Students were randomly selected to experience different instruction video conditions (full instruction video vs. restricted instruction video). Additionally, various difficulty level conditions and feedback conditions (e.g., DSCoVAR provides a feedback message to the student based on answer accuracy vs. no feedback) were tested within the same student. However, in this study, we focused on data from students who experienced a full instruction video and repeating difficulty conditions. Repeating difficulty conditions included questions with all high or medium contextual constraint levels. By doing so, we wanted to minimize the impact from various experimental conditions for analyzing post-test outcomes. Moreover, we filtered out response sequences that did not include all four responses for the target word. As a result, we analyzed 818 response sequences from 7,425 items in total.

Immediate and Delayed Post-test. The immediate post-test occurred right after the students finished the training; the delayed post-test was conducted one week later. Data collected during the immediate and delayed post-tests were used to estimate short-and long-term learning, respectively. Test items were identical to those in the pretest session, except for item order, which varied across tests. For analysis of the delayed post-test data, we only used the data from target words for which the student provided a correct answer in the earlier, immediate post-test session. As a result, 449 response sequences were analyzed for predicting the long-term retention.

3.2 Semantic Score-Based Features

We now describe the semantic features tested in our prediction models.

3.2.1 Semantic Scales

For this study, we used semantic scales from Osgood's study [16]. Ten scales were selected by a cognitive psychologist as being considered semantic attributes that can be detected during word learning (Figure 2). Each semantic scale consists of pairs of semantic attributes. For example, the *bad–good* scale can show how the meaning of a word can be projected on a scale with *bad* and *good* located at either

Figure 2: Ten semantic scales used for projecting target words and responses [16].

- bad – good
- complex – simple
- passive – active
- fast – slow
- powerful – helpless
- noisy – quiet
- big – small
- new – old
- helpful – harmful
- healthy – sick

end. The word's relationship with each semantic anchor can be automatically measured from its semantic similarity with these exemplar semantic elements.

3.2.2 Basic Semantic Distance Scores

To extract meaningful semantic information, we have applied the following measures that can be used to explain various characteristics of student responses for different target words. In this study, we used a pre-trained model for Word2Vec,[1] built based on the Google News corpus (100 billion tokens with 3 million unique vocabularies, using a negative sampling algorithm), to measure semantic similarity between words. The output of the pre-trained Word2Vec model contained a numeric vector with 300 hundred dimensions.

First, we calculated the relationship between word pairs (i.e., a single student response and the target word, or a pair of responses) in both the regular Word2Vec (W2V) score and the Osgood semantic scale (OSG) score. In the W2V score, the semantic relationship between words was represented with a cosine similarity between word vectors:

$$D_{w2v}(w_1, w_2) = 1 - |sim(V(w_1), V(w_2))|. \quad (1)$$

In this equation, the function V returned the vectorized representation of the word (w_1 or w_2) from the pre-trained Word2Vec model. By calculating the cosine similarity between two vectors (a cosine similarity function is noted as sim), we could extract a single numeric similarity score between two words. This score was converted into a distance-like score by taking the absolute value of the cosine similarity score and subtracting from one.

For the OSG score, we extracted two different types of scores: a non-normalized score and a normalized score. A non-normalized score showed how a word is similar to a single anchor word (e.g., *bad* or *good*) from the Osgood scale.

$$S^{non}_{osg}(w, a_{i,j}) = sim(V(w), V(a_{i,j})) \quad (2)$$

$$D^{non}_{osg}(w_1, w_2; a_{i,j}) = |S^{non}_{osg}(w_1, a_{i,j})| - |S^{non}_{osg}(w_2, a_{i,j})| \quad (3)$$

In equation 2, $a_{i,j}$ represents a single anchor word (j) in the i-th Osgood scale. The similarity between the anchor word and the evaluating word w was calculated with cosine similarity of Word2Vec outcomes for both words. In a non-normalized setting, the distance between two words given by a particular anchor word was calculated by the difference of absolute cosine similarity scores (equation 3).

The second type of OSG score is a normalized score. By using Word2Vec's ability to do arithmetical calculation of

[1] API and pre-trained model for Word2Vec was downloaded from this URL: https://github.com/3Top/word2vec-api

multiple word vectors, the normalized OSG score provided a relative location of the word from two anchor ends of the Osgood scale.

$$S_{osg}^{nrm}(w, a_i) = sim(V(w), V(a_{i,1}) - V(a_{i,2})) \quad (4)$$

$$D_{osg}^{nrm}(w_1, w_2; a_i) = |S_{osg}^{nrm}(w_1, a_i) - S_{osg}^{nrm}(w_2, a_i)| \quad (5)$$

In equation 4, the output represents the cosine similarity score between the word w and two anchor words ($a_{i,1}$ and $a_{i,2}$). For example, if the cosine similarity score of $S_{osg}^{nrm}(w, a_i)$ is close to -1, it means the word w is close to the first anchor word $a_{i,1}$. If the score is close to 1, it is vice versa. In equation 5, the distance between two words was calculated as the absolute value of the difference between two cosine similarity measures.

3.2.3 Deriving Predictive Features

Based on semantic distance equations explained in the previous section, this section explains examples of predictive features that we used to predict students' short-term learning and long-term retention.

Distance Between the Target Word and the Response. For regular Word2Vec score models and Osgood scale score models, distance measures between the target word and the response (by using equations 1 and 5) were used to estimate the accuracy of the response to a question. This feature represents the trial-by-trial answer accuracy of a student response. Each response sequence for the target word contained four distance scores.

Difference Between Responses. Another feature that we used in both types of models was the difference between responses. This feature could capture how a student's current answer is semantically different from the previous response. From each response sequence, we could extract three derivative scores from four responses.

Convex Hull Area of Responses. Alternative to the difference between responses feature, Osgood scale models were also tested with the area size of convex hull that can be generated by responses calculated with non-normalized Osgood scale scores (equation 3). For example, for each Osgood scale, a non-normalized score provided two-dimensional scores that can be used for geometric representation. By putting the target word in an origin position, a sequence of responses can create a polygon that can represent the semantic area that the student explored with responses. Since some response sequences were unable to generate the polygon by including less than three unique responses, we added a small, random noise that uniformly distributed (between -10^{-4} and 10^{-4}) to all response points. Additionally, a value of 10^{-20} was added to all convex hull area output to create a visible lower-bound value.

Unlike the measure of difference between responses, this feature also considers angles that can be created between responses and the target word. This representation can provide more information than just using difference between responses.

3.3 Modeling

To predict students' short-term learning and long-term retention, we used a mixed-effect logistic regression model (MLR). MLR is a general form of logistic regression model that includes random effect factors to capture variations from repeated measures.

3.3.1 Off-line Variables

Off-line variables capture item- or subject-level variances that can be observed repeatedly from the data. In this study, we used multiple off-line variables as random effect factors.

First, results from familiarity-rating and synonym-selection questions from the pre-test session were used to include item- and subject-level variances. Both variables include information on the student's prior domain knowledge level for target words. Second, the question difficulty condition was considered as an item group level factor. In the analysis, sentences for the target word that were presented to the student contained the same difficulty level, either high or medium contextual constraint levels, over four trials. Third, a different experiment group was used as a subject group factor. As described in Section 3.1.1, this study contains data from students in different institutions in separate geographic locations. The inclusion of these participant groups in the model can be used to explain different short-term learning outcomes and long-term retention by demographic groups.

3.3.2 Model Building

In this study, we compared the performance of MLR models with four different feature types. First, the baseline model was set to indicate the MLR model's performance without any fixed effect variables but only with random intercepts. Second, the response time model was built to be compared with semantic score-based models. Many previous studies have used response time as an important predictor of student engagement and learning [2, 12]. In this study, we used two types of response time variables, the latency for initiating the response and finishing typing the response, as predictive features. Both variables were measured in milliseconds over four trials and natural log transformed for the analysis. Third, semantic features from regular Word2Vec scores were used as predictors. This model was built to show how semantic scores from Word2Vec can be useful for predicting students' short- and long-term performance in DSCoVAR. Lastly, Osgood scale-based features were used as predictors. This model was compared with the regular Word2Vec score model to examine the effectiveness of using Osgood scales for evaluating students' performance in DSCoVAR. For these semantic-score based models, we tested out different types of predictive features that were described in Section 3.2.3. All models shared the same random intercept structure that treated each off-line variable as an individual random intercept.

For Osgood scale models, we also derived reduced-scale models. Reduced-scale models were compared with the full-scale model, which uses all ten Osgood scales. In this case, using fewer Osgood scales can provide easier interpretation of semantic analysis for intelligent tutoring system users.

3.3.3 Model Evaluation

To compare performance between different models, this study used various evaluation metrics, including AUC (an area under the curve score from a response operating characteristic (ROC) curve), F_1 (a harmonic mean of precision and recall), and error rate (a ratio of the number of

misclassified items over total items). 95% confidence interval of each evaluation metric was calculated from the outcome of a ten-fold cross-validation process repeated over ten times.

To select the semantic score-based features for models based on regular Word2Vec scores and Osgood scale scores, we used rankings from each evaluation metric. The model with the highest overall rank (i.e., sum the ranks from AUC, F_1, and error rate, and select the model with the lowest rank-sum value) was considered the best-performing model for the score type (i.e., models based on the regular Word2Vec score or Osgood scale score). More details on this process will be illustrated in the next section.

4. RESULTS

4.1 Selecting Models

In this section, we selected the best-performing model based on the models' overall ranks in each evaluation metric. All model parameters were trained in each fold of repeated cross-validation. We calculated 95% confidence intervals for comparison. To calculate the confidence interval of F_1 and error rate measures, the maximum (F_1) and minimum (error rate) scores of each fold were extracted. These maximum and minimum values were derived from applying multiple cutoff points to the mixed-effect regression model.

4.1.1 Predicting Immediate Learning

First, we built models that predict the students' immediate learning from the immediate post-test session. From models based on regular Word2Vec scores (W2V), the model with the distance between the target and responses and the difference between responses (*Dist+Resp*) provided the highest rank from various evaluation metrics (Table 2). From models based on Osgood scales (OSG), the model with the difference between responses (*Resp*) provided the highest rank.

The selected W2V model provided significantly better performance than the baseline model. The selected OSG model also showed significantly better performance than the baseline model, except for the AUC score. The selected W2V model was significantly better than the model using response time features in the AUC score and error rates.

The selected W2V model showed significantly better performance than the OSG model only with the AUC score. Figure 3 shows that the W2V model has a slightly larger area under the ROC curve than the OSG model. In the precision and recall curve, the selected W2V model provides more balanced trade-offs between precision and recall measures. The selected OSG model outperforms the W2V model in precision only in a very low recall measure range.

4.1.2 Predicting Long-Term Retention

We also built prediction models to predict the students' long-term retention in the delayed post-test session. In this analysis, a student response was included only when the student provided correct answers to the immediate post-test session questions. Among W2V score-based models, the best-performing model contained the same feature types as the immediate post-test results (Table 3). By using the distance between the target and responses and difference between responses (*Dist+Resp*), the model achieved significantly better performance than the baseline model, except for the AUC score.

For OSG models, the model with a convex hull area of responses (*Chull*) provided the highest overall rank from evaluation metrics (Table 3). The results were significantly better than the baseline model, and marginally better than the W2V model. Both selected W2V and OSG models were marginally better than the response time model, except the error rate of the OSG model was significantly better.

In Figure 3, the selected W2V model slightly outperforms the OSG model in mid-range true positive rates, while the OSG model performed slightly better in a higher true positive area. Precision and recall curves show similar patterns to those we observed from the immediate post-test prediction models. The OSG model only outperforms the W2V model in a very low recall value area.

4.1.3 Comparing Models

Compared to the selected W2V model in the immediate post-test condition, the selected W2V model in the delayed post-test retention condition showed a significantly lower AUC score, marginally higher F_1 score, and marginally higher error rate. In terms of OSG models, the selected OSG model for delayed post-test retention showed a significantly better F_1 score and error rates than the selected OSG model in the immediate post-test condition. Based on these results, we can argue that Osgood scale scores can be more useful for predicting student retention in the delayed post-test session than predicting the outcome from the immediate post-test.

In terms of selected feature types, the best-performing OSG models used features based on the difference between responses (*Resp*) or the convex hull area (*Chull*) that was created from the relative location of the responses. On the other hand, selected W2V models used both the distance between the target word and responses and difference between responses (*Dist+Resp*). When we compared both W2V and OSG models using the difference between responses feature, we found that performance is similar in the immediate post-test data. However, the OSG model was significantly better in the delayed post-test data. These results show that Osgood scale scores can be more useful for representing the relationship among response sequences.

4.2 Comparing the Osgood Scales

To identify which Osgood scales are more helpful than others for predicting students' performance, we conducted a scale-wise importance analysis. The results from this section reveal which Osgood scales are more important than others, and how the performance of prediction models with a reduced number of scales is comparable with the full-scale model.

4.2.1 Identifying More Important Osgood Scales

In this section, based on the selected Osgood score model from Section 4.1, we identified the level of contribution for features based on each Osgood scale. For example, the selected OSG model for predicting the immediate post-test data uses the difference between responses in ten Osgood scales as features. In order to diagnose the importance level of the first scale (*bad–good*), we can retrain the model with features based on the nine other scales and compare the

Table 2: Ranks of predictive feature sets for regular Word2Vec models (W2V) and Osgood score models (OSG) in the immediate post-test data. All models are significantly better than the baseline model. (Bold: the selected model with highest overall rank.)

	W2V models			OSG models		
Features	AUC	F_1	Err	AUC	F_1	Err
baseline	0.68 [0.67, 0.69] (5)	0.74 [0.73, 0.74] (5)	0.33 [0.33, 0.34] (5)	0.68 [0.67, 0.69] (5)	0.74 [0.73, 0.74] (5)	0.33 [0.33, 0.34] (7)
RT	0.69 [0.68, 0.70] (4)	0.75 [0.75, 0.76] (3)	0.31 [0.31, 0.32] (4)	0.69 [0.68, 0.70] (2)	0.75 [0.74, 0.76] (2)	0.31 [0.31, 0.32] (2)
Dist	0.72 [0.71, 0.74] (1)	0.76 [0.75, 0.76] (2)	0.29 [0.28, 0.30] (2)	0.67 [0.66, 0.68] (7)	0.73 [0.73, 0.74] (7)	0.33 [0.32, 0.34] (6)
Resp	0.70 [0.69, 0.71] (3)	0.75 [0.74, 0.76] (4)	0.31 [0.30, 0.32] (3)	**0.69 [0.68, 0.70] (1)**	**0.75 [0.75, 0.76] (1)**	**0.31 [0.30, 0.32] (1)**
Chull	NA	NA	NA	0.69 [0.68, 0.70] (3)	0.74 [0.73, 0.75] (4)	0.32 [0.31, 0.33] (4)
Dist+Resp	**0.72 [0.71, 0.73] (2)**	**0.76 [0.75, 0.77] (1)**	**0.29 [0.28, 0.30] (1)**	0.68 [0.67, 0.69] (4)	0.74 [0.73, 0.75] (3)	0.31 [0.31, 0.32] (3)
Dist+Chull	NA	NA	NA	0.67 [0.66, 0.68] (6)	0.74 [0.73, 0.74] (6)	0.33 [0.32, 0.34] (5)

Table 3: Ranks of predictive feature sets for W2V and OSG models in the delayed post-test data. All models are significantly better than the baseline model. (Bold: the selected model with highest overall rank.)

	W2V models			OSG models		
Features	AUC	F_1	Err	AUC	F_1	Err
baseline	0.65 [0.64, 0.67] (5)	0.75 [0.74, 0.76] (5)	0.33 [0.32, 0.34] (5)	0.65 [0.64, 0.67] (5)	0.75 [0.74, 0.76] (7)	0.33 [0.32, 0.34] (7)
RT	0.67 [0.65, 0.68] (3)	0.76 [0.76, 0.77] (4)	0.31 [0.30, 0.32] (3)	0.67 [0.65, 0.68] (3)	0.76 [0.76, 0.77] (5)	0.31 [0.30, 0.32] (5)
Dist	0.66 [0.64, 0.68] (4)	0.77 [0.76, 0.78] (3)	0.31 [0.30, 0.32] (4)	0.66 [0.64, 0.68] (4)	0.78 [0.77, 0.79] (3)	0.30 [0.29, 0.31] (3)
Resp	0.69 [0.67, 0.71] (1)	0.77 [0.76, 0.78] (2)	0.30 [0.29, 0.31] (2)	0.63 [0.61, 0.65] (7)	0.76 [0.75, 0.77] (6)	0.32 [0.31, 0.33] (6)
Chull	NA	NA	NA	**0.69 [0.68, 0.71] (1)**	**0.78 [0.77, 0.79] (2)**	**0.28 [0.27, 0.29] (1)**
Dist+Resp	**0.68 [0.66, 0.70] (2)**	**0.78 [0.77, 0.79] (1)**	**0.30 [0.29, 0.31] (1)**	0.64 [0.62, 0.66] (6)	0.77 [0.76, 0.78] (4)	0.31 [0.29, 0.32] (4)
Dist+Chull	NA	NA	NA	0.69 [0.67, 0.71] (2)	0.78 [0.78, 0.79] (1)	0.29 [0.27, 0.30] (2)

performance of the newly trained model with the existing full-scale model.

In Table 4, we picked the top five scales that were important in individual prediction tasks. We found that *big-small*, *helpful-harmful*, *complex-simple*, and *fast-slow* were commonly important Osgood scales for predicting students' performance in immediate post-test and delayed post-test sessions. Scales like *bad-good* and *passive-active* were only important scales in the immediate post-test prediction. Likewise, *new-old* was an important scale only in the delayed post-test prediction.

Table 4: Scale-wise importance of each Osgood scale. Scales were selected based on the sum of each evaluation metric's rank. (Bold: Osgood scales that were commonly important in both prediction tasks; *: top five scales in each prediction task including tied ranks)

	Imm. post-test				Del. post-test			
Scales	AUC	F_1	Err	All	AUC	F_1	Err	All
bad-good	1	1	1	1*	4	10	4	6
passive-active	2	4	3	2*	8	6	6	7
powerful-helpless	7	9	6	7.5	10	8	10	10
big-small	3	3	4	3*	1	3	2	2*
helpful-harmful	4	6	5	5.5*	2	1	1	1*
complex-simple	8	5	2	5.5*	3	5	7	4.5*
fast-slow	5	2	7	4*	6	4	3	3*
noisy-quiet	6	8	8	7.5	7	9	9	9
new-old	9	7	9	9	5	2	8	4.5*
healthy-sick	10	10	10	10	9	7	5	8

4.2.2 Performance of Reduced Models

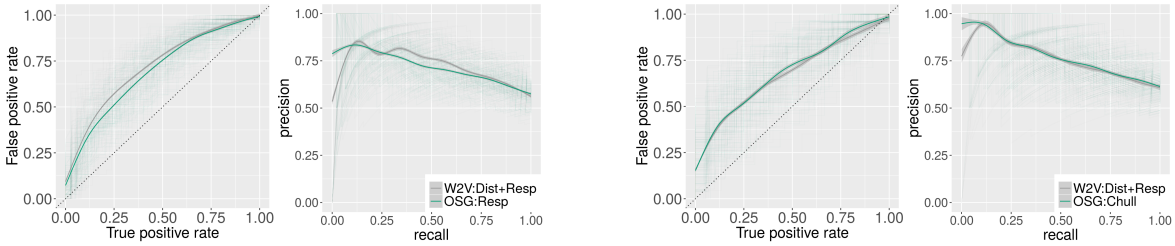
Based on the scale-wise importance analysis results, we built reduced-scale models that only contain features with more important Osgood scales. The prediction performance of reduced-scale models was similar or marginally better than full-scale OSG models. For example, the OSG model for predicting the immediate post-test outcome with the top two scales (*bad–good* and *passive–active*) were marginally better than the full-scale model (AUC: 0.71 [0.70, 0.72], F_1: 0.76 [0.75, 0.77], error rate: 0.30 [0.29, 0.30]). Similar results were observed for predicting retention in the delayed post-test (selected scales: *helpful–harmful*, *big–small*) (AUC: 0.71 [0.69, 0.72], F_1: 0.79 [0.78, 0.80], error rate: 0.28 [0.27, 0.29]). Although differences were small, the results indicate that using a small number of Osgood scales can be similarly effective to the full-scale model.

5. DISCUSSION AND CONCLUSIONS

In this paper, we introduced a novel semantic similarity scoring method that uses predefined semantic scales to represent the relationship between words. By combining Osgood's semantic scales [16] and Word2Vec [13], we could automatically extract the semantic relationship between two words in a more interpretable manner. To show this method can effectively represent students' knowledge in vocabulary acquisition, we built prediction models that can be used to predict the student's immediate learning and long-term retention. We found that our models performed significantly better than the baseline and the response-time-based models. In the future, we believe results from using an Osgood scale-based student model could be used to provide a more personalized learning experience, such as generating questions that can improve an individual student's understanding for specific semantic attributes.

Based on our findings, we have identified the following points for further discussion. First, in Section 4.1, we found that models using Osgood scale scores perform similarly with models using regular Word2Vec scores for predicting students' long-term retention of acquired vocabulary. However, we think our models can be further improved by incorporating additional features. For example, non-semantic score-based features like response time and orthographic similarity among responses can be useful features for capturing different patterns of false predictions of current models. Moreover, some general measures to capture a student's meta-cognitive or linguistic skills could be helpful to explain different retention results found even if students provided the same response sequences. Similarly, in Section 4.1.3, we found that Osgood scores can be a better metric to characterize the relationship between responses in terms of predicting students' retention. A composite model that uses both regular Word2Vec score-based feature (target-response distance) and Osgood scale score-based feature (response-response distance) may also provide better

Figure 3: ROC curves and precision and recall curves for selected immediate post-test prediction models (left) and delayed post-test prediction models (right). Curves are smoothed out with a local polynomial regression method based on repeated cross-validation results.



prediction performance.

Second, we found that models with a reduced number of Osgood scales performed marginally better than the full-scale model. However, differences were very small. Since this study only used some of the semantic scales from Osgood's study [16], further investigation would be required to examine the validity of these scales, including other scales not used for this study, for capturing the semantic attributes of student responses during vocabulary learning.

Also, there were some limitations in the current study and areas for future work. First, expanding the scope of analysis to the full set of experimental conditions used in the study may reveal more complex interactions between these conditions and students' short- and long-term learning. Second, this study used a fixed threshold of 0.5 for investigating false prediction results. However, an optimal threshold for each participant group or prediction model could be selected, especially if there are different false positive or negative patterns observed for different groups of students. Lastly, this study collected data from a single vocabulary tutoring system that was used in a classroom setting. Applying the proposed method to data that was collected from a non-classroom setting or other vocabulary learning system would be useful to show the generalization of our suggested method.

6. ACKNOWLEDGMENTS

The research reported here was supported by the Institute of Education Sciences, U.S. Department of Education, through Grant R305A140647 to the University of Michigan. The opinions expressed are those of the authors and do not represent views of the Institute or the U.S. Department of Education. We thank Dr. Charles Perfetti and his lab team at the University of Pittsburgh, particularly Adeetee Bhide and Kim Muth, and the helpful personnel at all of our partner schools.

7. REFERENCES

[1] S. Adlof, G. Frishkoff, J. Dandy, and C. Perfetti. Effects of induced orthographic and semantic knowledge on subsequent learning: A test of the partial knowledge hypothesis. *Reading and Writing*, 29(3):475–500, 2016.

[2] J. E. Beck. Engagement tracing: Using response times to model student disengagement. *Artificial Intelligence in Education: Supporting Learning Through Intelligent and Socially Informed Technology*, 125:88, 2005.

[3] K. Collins-Thompson and J. Callan. Automatic and human scoring of word definition responses. In *HLT-NAACL*, pages 476–483, 2007.

[4] M. Coltheart. The MRC psycholinguistic database. *The Quarterly Journal of Experimental Psychology*, 33(4):497–505, 1981.

[5] E. Dale. Vocabulary measurement: Techniques and major findings. *Elementary English*, 42(8):895–948, 1965.

[6] F. T. Durso and W. J. Shore. Partial knowledge of word meanings. *Journal of Experimental Psychology: General*, 120(2):190, 1991.

[7] G. A. Frishkoff, K. Collins-Thompson, L. Hodges, and S. Crossley. Accuracy feedback improves word learning from context: Evidence from a meaning-generation task. *Reading and Writing*, 29(4):609–632, 2016.

[8] G. A. Frishkoff, K. Collins-Thompson, S. Nam, L. Hodges, and S. A. Crossley. Dynamic support of contextual vocabulary acquisition for reading (DSCoVAR): An intelligent tutoring system for contextual word learning. *Handbook on Educational Technologies for Literacy*, 2016.

[9] G. A. Frishkoff, C. A. Perfetti, and K. Collins-Thompson. Predicting robust vocabulary growth from measures of incremental learning. *Scientific Studies of Reading*, 15(1):71–91, 2011.

[10] T. K. Landauer. *Latent Semantic Analysis*. Wiley Online Library, 2006.

[11] Y. Li, L. Xu, F. Tian, L. Jiang, X. Zhong, and E. Chen. Word embedding revisited: A new representation learning and explicit matrix factorization perspective. In *Proceedings of the 24th International Joint Conference on Artificial Intelligence, Buenos Aires, Argentina*, pages 3650–3656, 2015.

[12] Y. Ma, L. Agnihotri, M. H. Education, R. Baker, and S. Mojarad. Effect of student ability and question difficulty on duration. In *Educational Data Mining*, 2016.

[13] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119, 2013.

[14] G. A. Miller. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41, 1995.

[15] S. Nam. Predicting off-task behaviors in an adaptive vocabulary learning system. In *Educational Data Mining*, 2016.

[16] C. E. Osgood, G. J. Suci, and P. H. Tannenbaum. *The Measurement of Meaning*. University of Illinois Press, 1957.

[17] K. Ostrow, C. Donnelly, S. Adjei, and N. Heffernan. Improving student modeling through partial credit and problem difficulty. In *Proc. of the Second ACM Conference on Learning@Scale*, pages 11–20. ACM, 2015.

[18] P. I. Pavlik and J. R. Anderson. Practice and forgetting effects on vocabulary memory: An activation-based model of the spacing effect. *Cog. Science*, 29(4):559–586, 2005.

[19] E. G. Van Inwegen, S. A. Adjei, Y. Wang, and N. T. Heffernan. Using partial credit and response history to model user knowledge. *International Educational Data Mining Society*, 2015.

[20] L. M. Yonek. *The Effects of Rich Vocabulary Instruction on Students' Expository Writing*. PhD thesis, University of Pittsburgh, 2008.