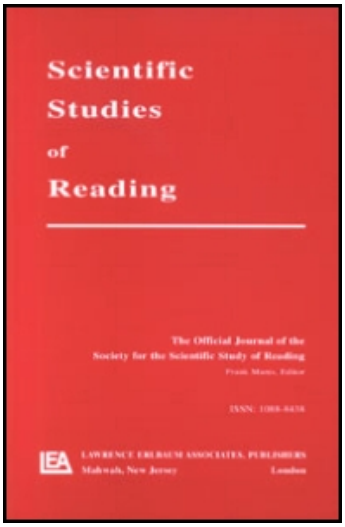


Scientific Studies of Reading
Publication details, including instructions for authors and subscription information:
http://www.informaworld.com/smpp/title~content=t775653700

Predicting Robust Vocabulary Growth from Measures of Incremental Learning

Gwen A. Frishkoff[a]; Charles A. Perfetti[b]; Kevyn Collins-Thompson[c]
[a] Georgia State University, [b] Learning Research and Development Center, University of Pittsburgh, [c] Microsoft Research,

Online publication date: 18 January 2011

To cite this Article Frishkoff, Gwen A. , Perfetti, Charles A. and Collins-Thompson, Kevyn(2011) 'Predicting Robust Vocabulary Growth from Measures of Incremental Learning', Scientific Studies of Reading, 15: 1, 71 — 91
To link to this Article: DOI: 10.1080/10888438.2011.539076
URL: http://dx.doi.org/10.1080/10888438.2011.539076

SCIENTIFIC STUDIES OF READING, *15*(1), 71–91

ISSN: 1088-8438 print / 1532-799X online
DOI: 10.1080/10888438.2011.539076

Predicting Robust Vocabulary Growth from Measures of Incremental Learning

Gwen A. Frishkoff
Georgia State University

Charles A. Perfetti
Learning Research and Development Center, University of Pittsburgh

Kevyn Collins-Thompson
Microsoft Research

We report a study of incremental learning of new word meanings over multiple episodes. A new method called MESA (Markov Estimation of Semantic Association) tracked this learning through the automated assessment of learner-generated definitions. The multiple word learning episodes varied in the strength of contextual constraint provided by sentences, in the consistency of this constraint, and in the spacing of sentences provided for each trained word. Effects of reading skill were also examined. Results showed that MESA scores increased with each word learning encounter. MESA growth curves were affected by context constraint, spacing of practice, and reading skill. Most important, the accuracy of participant responses (MESA scores) during learning predicted which words would be retained over a 1-week period. These results support the idea that word learning is incremental and that partial gains in knowledge depend on properties of both the context and the learner. The introduction of MESA presents new opportunities to test word-learning theories and the complex factors that affect growth of word knowledge over time and in different contexts.

Correspondence should be sent to Gwen A. Frishkoff, Department of Psychology, Georgia State University, 140 Decatur Street, Urban Life, Room 712, Atlanta, GA 30303-3083. E-mail: frishkoff@gsu.edu

Reading researchers have been served up a decisive challenge: to discover what factors lead to robust word knowledge and how this knowledge interacts with reading comprehension. In general terms, we know the answer: Learners require multiple, high-quality encounters with words in a variety of meaningful contexts (Beck, McKeown, & Kucan, 2002). However, what counts as a high-quality encounter—that is, a successful learning episode—may vary for different words at different points along the learning trajectory (Frishkoff, Collins-Thompson, Perfetti, & Callan, 2008). At this time little is known about the shapes of these trajectories or about how they vary in different learning contexts. Each encounter matters in ways not yet fully understood.

Progress in addressing these issues is likely to benefit from the use of robust methods for measurement of partial word knowledge. Here, we propose a new measure and test its sensitivity to factors that have previously been shown to play a role in word learning. Our central aim is to show that this method has promise as a tool for assessing changes in word knowledge and the conditions that produce these changes over time.

MESA: A NEW TOOL FOR CAPTURING PARTIAL WORD KNOWLEDGE AND INCREMENTAL LEARNING

We assume, with many others, that growth of word knowledge is incremental (e.g., Frishkoff et al., 2008; Frishkoff, Perfetti, & Collins-Thompson, 2009; Reichle & Perfetti, 2003; Stahl, 2003) and that word knowledge itself is often passive, unstable, and partial (Brown, Frishkoff, & Eskenazi, 2005; Durso & Shore, 1991). Theories of partial word knowledge differ in detail, but each captures the idea that word knowledge develops along several dimensions: familiarity with word forms increases, different aspects of meaning (both denotative and connotative) are revealed through exposure to words in different contexts, and associations between form and meaning are strengthened. The result is that word processing that is faster, more fluent, and less context-bound.

A model that captures these changes is the word experience model of Reichle and Perfetti (2003). In this model, word learning is viewed as a series of word episodes that occur over time, leaving memory traces that include both word forms and their contexts (including physical, modality-specific, linguistic, and affective contexts). The meaning of a word is the shared residue of these episodes, extracted from various contexts. The difference between a mature meaning representation and partial knowledge arises from both the quantity and quality of these episodes, that is, how well they promote the extraction of a core meaning and the range of its variations.

Although we have theoretical models (e.g., Reichle & Perfetti, 2003) and now powerful tools to simulate the growth of partial knowledge (Landauer, Kireyev, &

Panaccione, 2011 [this issue]), there has been less empirical work that tracks the incremental changes in word knowledge. One exception is Frishkoff et al. (2008), which described the use of MESA (Markov Estimation of Semantic Association) to capture changes in word knowledge over a 2-hr learning session. MESA uses a statistical model of word relations to score the accuracy of word definitions, where "accuracy" is defined as the estimated distance between participant-generated and target meanings (Collins-Thompson & Callan, 2007). In Frishkoff et al. (2008), MESA was used to score the accuracy of subject-generated definitions for very rare ("target") words that were presented in a variety of sentence contexts. Each word-learning episode (trial) consisted of viewing a target word in a single sentence context. After each context, participants generated a one-word definition. On most trials, contexts were semantically constraining, supporting valid inferences about a word's meaning (e.g., "Fresh air, exercise, and a good diet are part of a salubrious lifestyle."). On a few trials, however, the contexts were actually malapropisms designed to activate a similar-sounding (distractor) word representation and thus to promote invalid inferences about the target word meaning (e.g., "Mary was disgusted by the man's lewd gestures and salubrious remarks."). MESA estimated the distance from the response to the target region of its computed semantic space, revealing a gradual increase in word knowledge across trials, consistent with the assumption that meaning is acquired incrementally. In addition, word learning was modulated by *context quality*, that is, by the presence of more or less informative cues to word meaning. This context quality effect was modulated by spacing of practice and by individual differences in reading skill. High-skilled readers recovered more effectively than low-skilled readers from misleading contexts, as evidenced by the higher quality of their responses—but only in the narrow spacing condition. Frishkoff et al. reasoned that narrowly spaced trials favor a memory for past word episodes, allowing comparisons with the current episode. This suggests that the high-skilled readers benefited from their ability to detect inconsistencies in word usage and to adopt corrective strategies.

The present study extends this work and provides a new test for MESA using a different set of context variables. Instead of word learning episodes that provide conflicting information, the present study uses only contexts consistent with the meaning of the target word. We expected high constraint contexts to support more accurate meaning abstraction than sentences with low constraint. We were further interested in the effect of mixed constraints, in which some contexts were highly constraining, whereas others were not, given that some models of learning would predict an advantage of mixed versus all-high constraint sentences for long-term retention and transfer of learning to new contexts. A second question for the present study was the pattern of retention of word knowledge after a 1-week delay. Short-terms gains in word knowledge are easily observed, but long-term gains are the goal of word instruction. We tested both short- and long-term gains and also

used a transfer task, involving two-word semantic judgments, to assess robust learning after the delay (Frishkoff, Perfetti, & Collins-Thompson, 2010; Perfetti, Wloko, & Hart, 2005). Finally, as in the previous study by Frishkoff et al. (2008), we examined effects of vocabulary and comprehension skill to assess individual differences (i.e., learner characteristics) using the MESA methodology.

CONTEXT, SPACING, AND READING SKILL IN ROBUST WORD LEARNING

Robust knowledge implies retention over time. Indeed, one type of instruction may lead to higher scores on short-term measures of learning but prove inferior on delayed measures (Karpicke & Roediger, 2007; Roediger & Karpicke, 2006). In the case of word learning, we can ask whether the greater immediate learning that comes from high-constraint contexts survives over longer retention periods. A straightforward account is that learning conditions that support short-term retention also support longer term retention, with forgetting being constant over conditions. Thus high-constraint contexts should produce better learning and long-term retention than low-constraint contexts.

However, if we add a condition that contains a more realistic mix of high- and low-constraint contexts the picture may become more complex. Low-constraint contexts provide less support for meaning generation, and thus greater effort is required on these trials. Effortful retrieval is considered supportive for robust learning (Pavlik & Anderson, 2005; Schmidt & Bjork, 1992). Accordingly, an alternative prediction is that a mixture of high- and low-constraint sentences will bring about more robust learning than a set of consistently high-constraint contexts. Indeed, a prior study by Lampinen and Faries (1994) found that words trained in medium-constraint contexts, or in a mixture of high- and low-constraint contexts, showed more robust gains than words trained exclusively in high-constraint contexts.

Massed Versus Distributed Word Learning Episodes

The strength of a knowledge representation decays over time, a fact that leads to the following prediction for word learning episodes: Episodes that are experienced close in time (closely spaced or "massed" practice) will produce stronger memory representations and thus lead to better retention than episodes that are widely spaced ("distributed"). However, when a test is administered after a significant delay (typically 1 day or longer), the effect of massed versus distributed practice is reversed: Distributed practice produces greater long-term retention (Karpicke & Roediger, 2007; Pavlik & Anderson, 2005; Schmidt & Bjork, 1992). Frishkoff et al. (2008) provided some of the first evidence for spacing effects in word learning from context (vs. direct instruction): Accuracy on a delayed posttest

(semantic judgment for target words) was greater for words encountered in widely spaced versus grouped contexts—but only for more skilled readers (the advantage was nonsignificant for less-skilled readers). Of interest, MESA scores showed a Spacing × Skill interaction, such that more skilled readers (but not less skilled readers) were better able to recover from malapropisms when spacing was narrow. Thus, measures of performance during training suggest a slight advantage for narrowly spaced practice, and the reverse is true for tests of long-term retention. In both cases, however, spacing effects are only observed for more skilled readers. Given these prior results, it remains to be seen whether this reversal pattern of spacing effects over retention interval will observed with a more realistic range of supportive low- and high-constraint contexts.

Reading Skill

Word learning may be affected by individual reader skills, vocabulary knowledge, or both (Frishkoff et al., 2008; Perfetti et al., 2005; Stanovich, 1986). With a population comparable to what we sample in the present study, Perfetti et al. (2005) found that skilled comprehenders showed better learning from simple definition-type episodes than did less skilled comprehenders and showed evidence in event-related potentials that their memories for the word-learning episodes were stronger. We expect to see effects of comprehension skill and vocabulary knowledge in learning from context as well (Frishkoff et al., 2008) and to test whether the responsiveness to contextual constraint and spacing depends on these skill and knowledge factors.

The Present Study

The present study tests the above hypotheses concerning effects of different types of word episodes and reading skill on the incremental learning of word meanings. In particular, the study aims to address two questions. First, Can MESA scores reveal different rates of learning as a function of context constraint and reading skill? And second, Can MESA scores predict which words will be retained over time? In a broader context, the study aims to demonstrate the value of MESA in exposing variability in learning trajectories as a function of context and learner characteristics.

METHOD

Participants

Participants were recruited from a prescreened pool of college adults who had completed a battery of reading-related tests administered by the Reading and

Language Lab at the Learning Research and Development Center (see Frishkoff et al., 2008, Appendix B). Participants received payment ($7 per hour), academic course credit, or a combination, for the prescreening task and the word learning experiment.

The word learning experiment involved 34 participants from the prescreening pool. Four participants were excluded: 1 for missing Session 2, 1 for failing to respond on more than 40% of the trials, and 2 because of errors in test administration. All 30 remaining participants (11 male, 19 female) were monolingual native English speakers with normal or corrected-to-normal vision with average age of 23.3 years (range = 18 to 53). No participants reported a history of any reading or language disorder.

Table 1 shows mean scores on the two prescreening tests used in the present study, the Nelson-Denny Vocabulary and Nelson-Denny Reading Comprehension tests. These scores were used to group participants into three levels of reading skill, based on a composite of their standardized scores on these two tests. The highest level ($n = 10$, HiSkill group) had strong vocabulary and reading comprehension skills (.5 *SD* higher than the mean scores of the vocabulary and/or comprehension test). The lowest level ($n = 10$, LoSkill group) scored at least .5 *SD* below the mean on one or both tests. The remaining participants ($n = 10$, MedSkill group) had scores within the middle one third of the distribution on both tests, except for two participants, who scored in the low range on the vocabulary test and in the high range on the comprehension test.

Experiment Protocol

Participants completed three sessions: a pretest session, a training session with an immediate posttest, and a delayed posttest session. The pretest (Session 1) and training (Session 2) were scheduled 4 to 7 days apart. The delayed posttest (Session 3) was exactly 7 days after the training session.

Pretest (Session 1). The pretraining session included two computerized assessments designed to test participant familiarity and partial knowledge of

TABLE 1
Subgroup Scores on Vocabulary and Comprehension Subtests of the Lexical Knowledge Battery[a]

	High Skilled		*Medium Skilled*		*Low Skilled*	
	M	*SD*	*M*	*SD*	*M*	*SD*
Vocabulary	70.5	8.9	55.0	13.2	33.5	10.9
Comprehension	27.7	4.4	23.2	3.2	13.6	6.4

[a]See Frishkoff et al. (2008; Appendix B) for details.

the target (rare-trained) words. In the *Familiarity Test*, participants rated their familiarity with each word on a scale from 1 to 3.

- 1 = "I don't remember having seen or heard this word before."
- 2 = "I have seen or heard this word before, but I don't know what it means."
- 3 = "I have seen this word before, and I think I know what it means."

Following each rating, participants were prompted to generate a one-word definition or synonym for the target word, even if they were uncertain about its meaning. There were 10 s allowed for a response prior to the next trial. The Familiarity pretest included the 60 rare-trained words and 30 low-frequency control words that are typically judged as "known" (mean rating close to 3 on the 3-point familiarity rating scale).

Participants also completed a *Synonym Judgment Task* of 90 multiple-choice items (60 rare-trained words and 30 low-frequency words). For each item, participants selected the word that was closest in meaning to the target word. The five choices included four distracters from the same word frequency range, average length, and part of speech as the correct response. Participants also rated the confidence of their judgments on a scale from 1 (*just guessing*) to 7 (*very confident*). For a more sensitive measure that can capture partial word knowledge, we combined the accuracy and confidence scores to derive a single *Degree of Knowledge* (DoK) score for each word, computed as follows. DoK = 1 when the response was wrong *or* confidence rating was less than 3. DoK = 2 when the response was accurate *and* confidence rating was between 3 and 5. Finally, DoK = 3 when the response was accurate *and* confidence rating was 6 or 7.

Training (Session 2). In Session 2, participants viewed each of the 60 target words in six different, interleaved contexts (sentences), for a total of 360 learning trials. After each sentence, participants generated a one-word meaning (synonym or near-synonym) for the target word.

The 360 trials included the six conditions of a 3 (context constraint) × 2 (intertrial spacing) factorial design. *Context constraint* was the number of unique cloze completions for each context and was determined in a different study with a comparable sample (Frishkoff et al., 2010). Each rare trained word was presented in six different sentences, each either high constraint (HiConstr) or low constraint (LoConstr). In the AllHi condition, all six sentences for a particular rare trained word were HiConstr. In the AllLo condition, all six sentences were LoConstr. The Mixed condition contained three high- and three low-constraint contexts. Each target was assigned to one of these three conditions, counterbalanced across participants.

Intertrial spacing was defined by the number of trials separating the presentation of contexts for a given rare trained word. In the Narrow spacing condition,

contexts were separated by an interval of 3 to 5 trials. In the Wide spacing condition, the interval was 14 to 25 trials. The order of trials was randomized, subject to the constraints on intertrial spacing.

The session contained one practice block (9 trials) followed by four blocks of experimental trials (90 trials per block).

Immediate posttest (Session 2). Immediately after the word training in Session 2, participants completed the synonym judgment test (DoK).

Delayed posttest (Session 3). The final session took place 1 week after training and included three types of assessments. The Familiarity and Synonym Judgment tasks were repeated. As a transfer task, we modified the *Semantic Priming* paradigm of Perfetti et al. (2005). Participants saw a sequence of two words and decided whether they were semantically related. The first ("prime") word, exposed for 1,000 ms, belonged to one of three categories: (a) Rare Trained words ($n = 60$), (b) Familiar Untrained words ($n = 30$), or (c) Rare Untrained words ($n = 30$). The second ("target") word, which followed immediately, was a familiar word, either a near-synonym for the prime word or an unrelated word. Participants indicated their decision with a key press, using the 1 key (right index finger) and the 2 key (right middle finger) on the right side of the keyboard. "Yes"/"No" assignment of keys was counterbalanced across participants. Accuracy was emphasized over speed. The unrelated word pairs were created by reshuffling the 60 related prime-target pairs. The related and unrelated word pairs were the same for every participant, with order randomized.

Experiment Stimuli

Target words. Stimuli in the pretest session were 60 very rare words and 30 familiar (low-frequency) words. Rare words included 17 nouns, 9 verbs, and 34 adjectives, with a mean length of 7.1 letters ($SD = 1.1$). No rare word occurred on the Francis and Kucera (1982) norms. Thus, all rare words had a frequency of less than 1 per million. Pretesting confirmed the unfamiliarity of these words, as participants were barely above chance levels in selecting their meanings (mean accuracy = 30%; chance accuracy = 20%).

The 30 familiar words were from a larger set of 60 relatively low-frequency words from a previous word recognition experiment (Frishkoff et al., 2008). The mean written frequency for these words was 3.38 per million ($SD = 8.15$). Average word length was 6.9 letters ($SD = 1.32$), not significantly different from the average for the rare words ($p > .7$).

The delayed posttest included these same 90 words (60 rare and 30 familiar words). In the transfer (semantic priming) task, the prime words included these same 60 Trained Rare and 30 Untrained Familiar words plus 30 Untrained Rare

words (e.g., "accolent") to establish a baseline. All untrained rare words had a frequency of less than 1 per million and averaged 7.37 letters ($SD = 0.92$). Target words for semantic priming were 120 familiar words, near-synonyms of the prime words. Target words averaged 5.00 letters ($SD = 1.25$) and 90.65 per million ($SD = 134.51$) written word frequency. Each participant saw the same target twice, paired once with a closely related prime and once with an unrelated prime.

Training contexts. There were 720 sentences constructed to be either high or low constraint. Their classification was subsequently validated in a cloze completion task (Taylor, 1953) by a separate group of 60 adult, monolingual English-speaking participants. In this task, the rare word in each sentence was replaced by an underscore (blank space), and participants were asked to provide a one-word completion ("cloze" response), as illustrated next.

1. The _____ firefighter ran into the burning house.
2. His friends did not consider him a very _____ man.

Context 1 provides strong clues to the meaning of the target word (*brave*) "impavid," as evidenced by 64% of respondents providing "brave" as the response to Context 1. By contrast, Sentence 2 is fairly low-constraint: Practically any adjective can be used in this context, as long as it can refer to a personal trait. The most common cloze response for Sentence 2 was "nice," with a relatively low cloze probability of 9%. Table 2 shows two examples of high- and low-constraint contexts for two of the trained (rare) words in this study.

TABLE 2
Sample Target (Rare) Words, Cohort Words, and Training Contexts

Target Word	*Cohort Words*	*High Constraint Contexts*	*Low Constraint Contexts*
impavid	fearlessous brave courage	The impavid firefighter ran into the burning house.	Her impavid actions were the subject of a lot of discussion.
		Policemen must be impavid to fight crime every day.	His friends did not consider him to be a very impavid man.
roodge	lift hoist move	Only a superhero could roodge a car above his head.	They were eventually able to roodge all the objects.
		The weight lifter was able to roodge a hundred pounds.	My brother is able to roodge more things than I can.

To capture the average level of constraint for each sentence context, we examined several measures related to the cloze probability for the target word meaning (see Frishkoff et al., 2010, Table 1, for a full set of results). According to each of these measures, the high-constraint sentences were more semantically constraining than the low-constraint sentences. For example, according to one measure of constraint, which considered all synonyms of the target word as correct (target) responses, 39% ($SD = 15\%$) of participants provided the expected response for high constraint, compared with 3% ($SD = 7\%$) for low constraint. High and low constraint sentences were matched in average length (mean number of words/sentence = 10.22, $SD = 1.50$).

Definition Scoring (MESA)

Participant responses on the definition-generation task were corrected for spelling, and unintelligible responses (e.g., "wti") were discarded. Spelling-corrected responses were then entered into the MESA definition-scoring algorithm. The MESA scores are measures of the semantic distance between the participant's response and the correct response (near-synonym). These scores were computed using a statistical model of text semantic similarity, as described in Collins-Thompson and Callan (2007) and implemented in a similarity network as in Frishkoff et al. (2008). The model uses Markov chains on a graph of individual word relations to compute the distance between word semantics. This graph is constructed from a weighted combination of links, where each link defines a particular type of relationship between words (e.g., stemming, synonymy, co-occurrence, hypernomy, hyponymy, and associative strength).

RESULTS

We report analysis results for four measures of word semantic learning: (a) self-rated knowledge of target words (Familiarity Test), (b) accuracy on the Synonym Test, (c) speed and accuracy on the Semantic Priming (Transfer) Test, and (d) quality of participant definitions (MESA scores). Each measure was tested in a repeated measures analysis of variance with two within-subjects factors: Spacing (Spaced vs. Massed practice) and Context Constraint (AllHi, Mixed, AllLo). Reading skill (HiSkill vs. MedSkill vs. LowSkill Group), the composite of vocabulary and comprehension, was used as the between-subjects measure for each analysis. Session effects included two levels (pre- vs. delayed posttest) for Familiarity scores and three levels (pretest vs. immediate posttest vs. delayed posttest) for Synonym test scores. The corresponding factor in the MESA analysis

is Time (Trials 1, 2–3, 4–5, and 5–6). In addition, a secondary MESA analysis included long-term Retention (retained vs. forgotten) as a measure. Finally, Relatedness (related vs. unrelated prime-target pairs) is unique to the semantic judgment transfer test in Session 2.

Pre- and Posttest Results for Familiarity Task

For pretest performance on words, there was a small difference (nonsignificant) in familiarity between words assigned to the narrow versus wide spacing groups and no differences among the three reading groups ($p > .5$). Posttest results showed a main effect of Session, $F(1, 26) = 359.02$, $MSE = .241$, $p < .001$; a Session × Group interaction, $F(2, 26) = 5.29$, $MSE = .241$, $p < .05$; and a Session × Spacing interaction, $F(1, 26) = 6.31$, $MSE = .020$, $p < .05$. Post hoc comparisons revealed small differences between groups after training (LoSkill $M = 2.0$, MedSkill = 2.2, and HighSkill = 2.3) that were not reliable (high vs. low, $p > .05$). (Although only rare trained words were in this analysis, Figure 1 shows familiarity ratings for known words and for untrained rare words to provide ceiling and baseline reference points.)

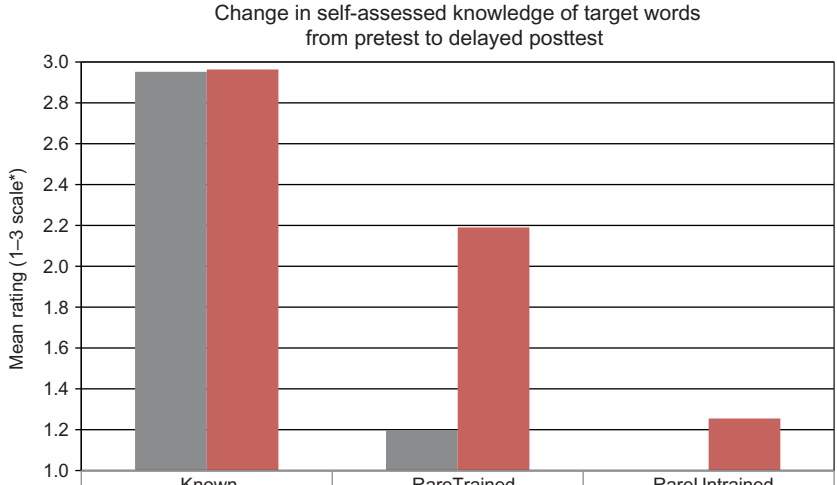


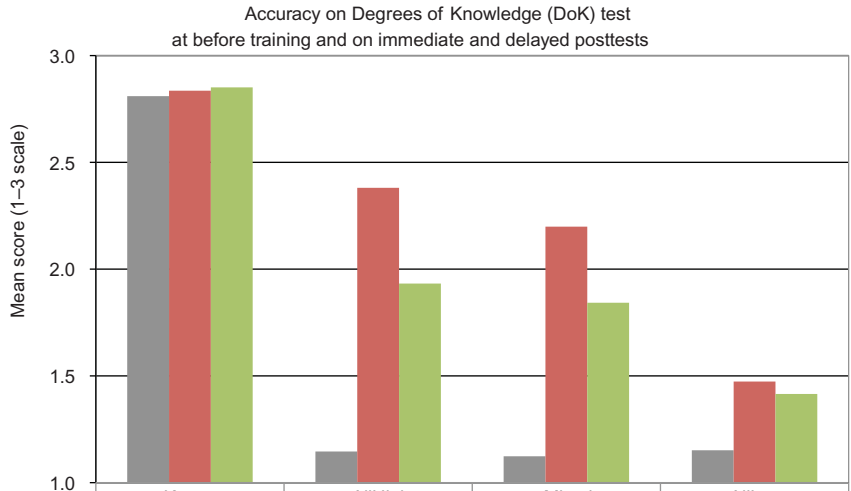
FIGURE 1 Mean self-rated familiarity with Known words, TrainedRare words, and UntrainedRare words in the pretest (medium grey) and delayed posttest (dark grey) sessions (∗1 = word is unfamiliar; 2 = word is familiar, but I don't know what it means; 3 = I know this word). *Note*: RareUntrained words were not presented at pretest.

Pretest Versus Immediate and Delayed Posttest Results for Synonym (DoK) Task

Pre- to posttest gains in target word knowledge. A reliable increase in the DoK measure occurred across the three sessions, $F(2, 52) = 182.90$, $MSE = .190$, $p < .001$. On average, DoK scores increased about 75% from pretest ($M = 1.14$, $SE = .033$) to the immediate posttest ($M = 2.02$, $SE = .039$). From the immediate to the delayed posttest approximately 35% of these gains were lost.

Effects of context constraint. Context Constraint affected word learning, $F(2, 52) = 128.97$, $MSE = .084$, $p < .001$. Figure 2 shows this effect was unequal across sessions, Session × ContextConstraint, $F(4, 104) = 75.76$, $MSE = .046$, $p < .001$. AllHigh and Mixed conditions led to substantial gains in partial word knowledge over sessions, whereas the AllLow condition resulted in only small, nonsignificant gains.

DoK scores on the immediate posttest were substantially greater for the Mixed condition versus the AllLow condition (M difference = .32, $p < .001$). There was also a smaller, but significant, advantage for the AllHigh versus Mixed condition (M difference = .12, $p < .05$). At the delayed posttest, the difference between



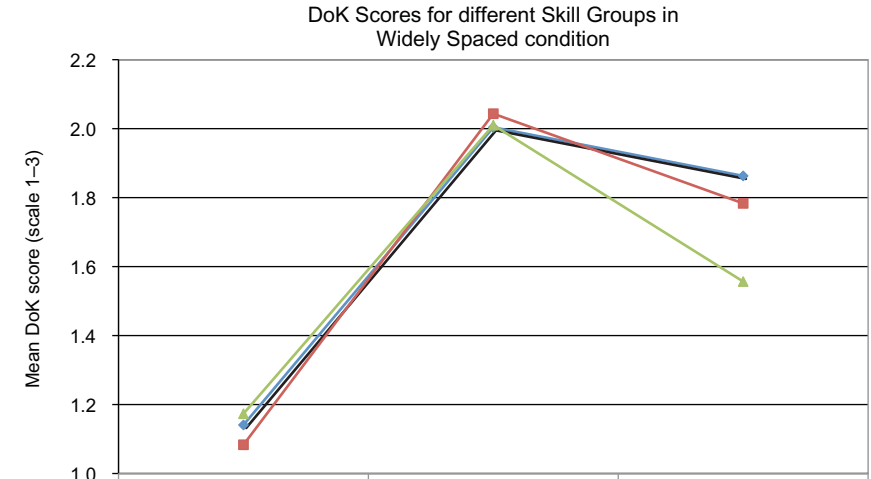
	Known	AllHigh	Mixed	AllLow
PreTest	2.81	1.15	1.12	1.15
ImmedPost	2.84	2.38	2.20	1.47
DelayedPost	2.85	1.93	1.84	1.42

FIGURE 2 Mean scores on Degrees of Knowledge (DoK) test before training (medium grey), on the immediate posttest (dark grey), and on the delayed posttest (light grey). See text for an explanation of how DoK scores are computed.

the Mixed versus AllLow condition remained significant (M difference = .18, $p < .001$); however, the advantage for the AllHigh versus Mixed condition largely disappeared (M difference = .06, $p > .05$).

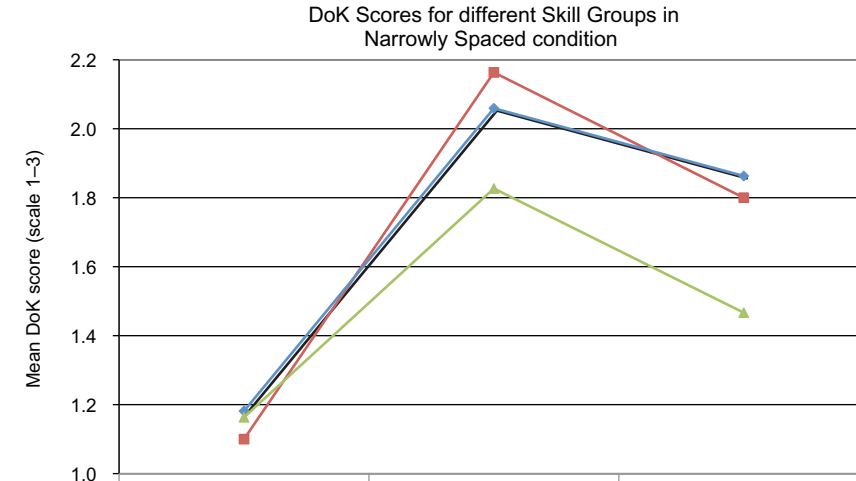
Effects of spacing. There was no main effect of spacing ($p > .5$). However, as described next, there were significant interactions of spacing with skill differences (Group) and with group differences in word learning (i.e., a Session × Spacing × Group interaction).

Effects of reading skill. Although the main effect of reading skill (Group) was not significant ($p > .1$), reading skill modulated the main effect of session, Group × Session, $F(4, 52) = 4.13$, $MSE = .785$, $p < .01$. Further, although the main effect of spacing was not significant (see earlier), there was a Group × Spacing interaction, $F(4, 52) = 5.85$, $MSE = .306$, $p < .01$. In addition, the three-way interaction of Session × Spacing × Group was also significant, $F(4, 52) = 3.61$, $MSE = .022$, $p < .05$. As shown in Figure 3, LowSkill readers appeared to have greater difficulty in the narrow space condition. This effect was confirmed in follow-up analyses, which showed between-group differences for words trained in the narrow space condition, both on the immediate posttest



	PreTest	ImmedPost	DelayedPost
HiSkill	1.141	2.004	1.863
MedSkill	1.083	2.043	1.783
LoSkill	1.173	2.010	1.557

FIGURE 3a Mean scores on DoK test for different skill groups in wide spacing condition.



	PreTest	ImmedPost	DelayedPost
HiSkill	1.181	2.059	1.863
MedSkill	1.100	2.163	1.800
LoSkill	1.163	1.827	1.467

FIGURE 3b Mean scores on DoK test for different skill groups in narrow spacing condition.

($p < .01$) and on the delayed posttest ($p < .01$), but not on the pretest ($p > .5$) and not on the pre- or posttests for words trained in the wide space condition (all $p > .1$). For words trained in the narrow space condition, post hoc tests confirmed that the low-skilled group performed substantially worse than the medium- and high-skilled groups ($p < .05$ for both group comparisons) on the immediate posttest, whereas the medium- and high-skilled groups did not differ ($p > .3$).

Semantic Priming Results (Delayed Posttest)

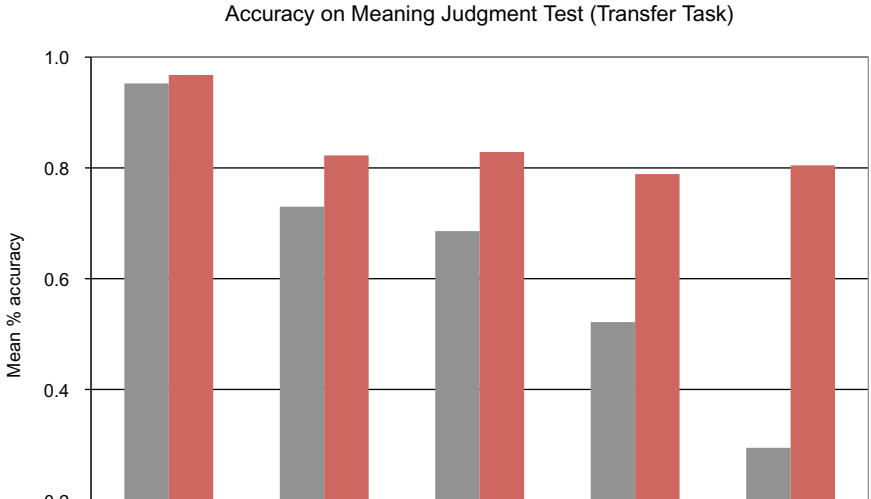
Accuracy and response time results on the semantic priming task were remarkably similar to results from a previous study (Frishkoff et al., 2010). There was greater accuracy of responses to related targets, but not unrelated targets, following trained rare words (main effect of Relatedness), $F(1, 25) = 33.37$, $MSE = .070$, $p < .001$. A strong "no" response bias is evidenced by the relatively high accuracy for unrelated versus related targets. Accuracy analyses revealed a main effect of ContextConstraint, $F(2, 50) = 23.63$, $MSE = .020$, $p < .001$, and an interaction of Relatedness × ContextConstraint, $F(2, 50) = 10.60$, $MSE = .021$, $p < .001$. The AllHigh condition produced better accuracy than the Mixed constraint condition (M difference $= .11$, $p < .01$). Mixed and AllLow contexts did not differ ($p > .3$).

Finally, reading skill differences appeared on the semantic priming task—Group, $F(2, 25) = 4.73$, $MSE = .066$, $p < .05$. Accuracy was 78%, 73%, and 67% for HiSkill, MedSkill, and LoSkill, respectively. Thus the priming task showed effects of training (Session differences), context and reading skill. Most interesting is that trials with high constraint led to better transfer to this task compared with both low constraint and mixed constrain conditions.

MESA Results

For analysis of responses in the definition-generation task, we treated the MESA scores—the MESA-computed distance between the participant's response and the target word meaning—as the dependent measure. To increase statistical power, the MESA scores were averaged across Trials 1–2, 3–4, and 5–6. Pretest scores were included, resulting in four time points for analysis. Results showed main effects of ContextConstraint, $F(2, 50) = 210.33$, $MSE = .012$, $p < .001$, and Time, $F(3, 75) = 226.87$, $MSE = .006$, $p < .001$, and an interaction of Time × ContextConstraint, $F(6, 150) = 68.76$, $MSE = .004$, $p < .001$. No other interactions were significant.

As illustrated in Figure 4, there was little change in MESA score for the AllLow condition. However, MESA scores increased substantially from pretest



	Known	AllHigh	Mixed	AllLow	UntrainedRare
Related	0.95	0.73	0.69	0.52	0.29
Unrelated	0.97	0.82	0.83	0.79	0.80

FIGURE 4 Mean accuracy on Meaning Judgment (Transfer) Task.

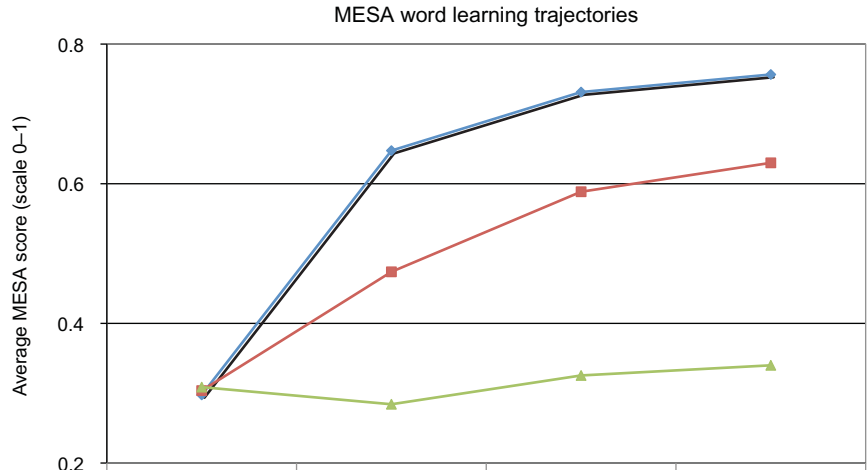
to Trials 1–2 for the AllHigh and Mixed conditions and continued to show gains across Trials 3–4 and Trials 5–6. The difference between AllHigh and Mixed conditions was not significant.

Skill differences were also observed in MESA scores: .51 for HiSkill, .47 for MedSkill, and .45 for LoSkill; Group, $F(2, 25) = 4.09$, $MSE = .023$, $p < .05$. Group did not interact with other factors.

The final analysis focused on MESA scores for words that were forgotten or retained over a 1-week period, as determined from accuracy on the delayed synonym posttest. MESA scores showed a significant interaction of Time × Retention, $F(3, 60) = 3.88$, $MSE = .015$, $p < .05$. As shown in Figure 5, MESA scores were higher for words that would later be retained versus those that would be forgotten. Moreover, the difference for retained versus forgotten words increased over time.

Summary of Results

When learners were exposed to rare words and then actively engaged in meaning generation, multiple word-in-context episodes resulted in short- and long-term growth in word knowledge, as evidenced by scores on several tests of partial word knowledge. Self-rated familiarity increased for rare trained words (Figure 1),



	PreTest	Contexts 1–2	Contexts 3–4	Contexts 5–6
AllHigh	.298	.647	.731	.756
Mixed	.304	.474	.589	.630
AllLow	.309	.284	.325	.340

FIGURE 5 MESA word learning trajectories for three training conditions.

as did accuracy on a multiple-choice synonym test. These gains were greater for words experienced in high constraint contexts or a mixture of high and low constraint contexts than for words that appeared only in low-constraint contexts (Figure 2). Approximately 30 to 40% of these gains were lost from immediate to delayed posttest. Furthermore, less skilled readers showed smaller gains, particularly in narrowly spaced contexts (Figure 3a–b). Most important, MESA scores increased over time, with each word-in-context episode (Figure 4). Moreover, accuracy of participant responses during learning predicted which words would be retained over a 1-week period.

DISCUSSION

The present study demonstrates the value of a novel method for assessment of incremental learning, while also providing new evidence concerning the kinds of word episodes that support the learning of word meanings.

MESA Trajectories Reflect Successful and Less Successful Word Learning

MESA was able to track the incremental learning of novel words that were embedded in six different contexts. MESA revealed that the quality of learner-generated meaning responses (synonym or near-synonyms of the novel word) changed over time toward the meaning of the novel word. Thus, MESA can capture incremental changes in word knowledge. In addition, these results provide tentative answers to these questions: Can MESA scores reveal different rates of learning as a function of context constraint? Can MESA scores predict which words will be retained over time?

On the first question, MESA scores captured the hypothesized difference between high-constraint and low-constraint context. MESA scores showed a reliable interaction between context constraint and session, from pretest through training. They captured robust gains over sessions when the six-word episodes included three (mixed-constraint) or six (high-constraint) supportive contexts and showed minimal gains when all contexts were low in constraint. When a word was presented in a mixture of high- and low-constraint sentences, MESA scores were not higher than when all contexts were high in constraint (and were lower according to the transfer task results). We discuss this result further in the next section.

On the second question, MESA successfully predicted robust learning as indicated by long-term retention of learned meanings. A high MESA score at the end of the learning trials predicted higher scores 1 week later. Underlying MESA's computation is a network of associated words that define the semantic field of a

word meaning. Thus, we can suggest that long-term retention of the meaning of a new word is related to the fit of the new word to the semantic network.

We conclude that MESA tracks word learning, captures the effects of word episode differences, and predicts retention.

What Kinds of Word Episodes Support Learning?

Variation in a number of word episode factors can affect learning. We examined two factors here—the degree and consistency of constraint provided by sentence contexts and the spacing of these contexts. We found, as have others, that high constraint produces more learning than low constraint. Of interest, we did not find an advantage for episodes that intermixed high- and low-constraint sentences. According to learning theories that stress the importance of effort (Schmidt & Bjork, 1992), the mixed-constraint condition might be expected to confer a benefit, as it is more challenging to infer the target word meaning in these contexts (e.g., Lampinen & Faries, 1994). However, we find no evidence for such an effect in the present study. Under some conditions, experience with medium- or low-constraint sentences may be helpful, perhaps when enough high-constraint sentences have already occurred, allowing effortful retrieval of meaning to be successful. However, in the conditions of our study, there was no cost to learning from consistently supportive word episodes.

The second factor, spacing of contexts, showed modest effects on word learning outcomes as measured by familiarity and synonym (meaning judgment) scores. As suggested in our previous article (Frishkoff et al., 2008), spacing of practice effects have been attributed to passive decay of the memory trace, which appears to explain learning (and forgetting) curves in simple, associative learning tasks (e.g., word learning from definitions; Pavlik & Anderson, 2005). By contrast, word learning from context may engage more active processes that are not captured by this simple model. Future models may need to account for these active processes more directly to explain the effects of spacing on word learning from context.

A final question is whether highly skilled readers receive more benefit from word learning episodes than less skilled readers. We found that high- and medium-skill learners did show larger gains than lower skilled readers on several measures. However, differences were modest, possibly due to the restricted range of scores in our participant sample. In previous studies where we strategically recruited low-skilled adult readers, in addition to medium- and high-skilled readers (e.g., Frishkoff et al., 2009; Frishkoff et al., 2008; Perfetti et al., 2005), we have found somewhat larger skill effects in word learning. For example, Perfetti et al., (2005) found that learning of rare words through definitions was related to comprehension skill. Similarly, Frishkoff et al., 2008 observed a significant interaction

between word learning from context and both vocabulary and comprehension skill (Frishkoff et al., 2008, Figure 3).

The question is, Why does reading skill affect word learning? A general, but imprecise, answer is that all literacy skills, including learning new words, depend on comprehension and vocabulary. However, more specific explanations may also be possible. Knowing more word meanings, for example, may provide more links for new words. This hypothesis receives some support from our earlier findings that high- versus low-skill readers learned more words in the presence of misleading contexts and knew more of the distractor words (Frishkoff et al., 2008). It is likely that this knowledge facilitated the rejection of contexts where the distractor was used incorrectly (i.e., the malapropisms). Of importance, vocabulary differences in knowledge of both target words (pre- and posttraining) and distractor words remained significant after partialing out variance due to comprehension skill. Thus, knowledge of words does appear to support more efficient learning of words from context. That said, other factors, such as background (world) knowledge and inferencing skills—that is, "comprehension" broadly defined—are likely to explain additional variance in word learning outcomes.

Finally, we note that one kind of effective word learning episode is one in which active processing of meaning occurs. The meaning generation task is one way of encouraging such processing and may produce robust learning, as well as providing data to measure this learning (Frishkoff et al., 2010; Frishkoff et al., 2008). However, the generation of near synonyms may be more or less difficult for different kinds of words. Other tasks that require active meaning processing may also prove to be effective and will be explored in future work.

ACKNOWLEDGMENTS

The research reported in this article was made possible by grant number 101HD058566-01A1 from the National Institutes of Health (NIH) to the University of Pittsburgh. Its contents are solely the responsibility of the authors and do not necessarily represent the official views of NIH. We thank Jamie Callan and the Language Technologies Institute at Carnegie Mellon University for the use of their computing and data storage resources.

REFERENCES

Baumann, J. F., Kame'enui, E. J., & Ash, G. E. (2003). Research on vocabulary instruction: Voltaire redux. In D. C. Simmons & E. J. Kame'enui (Eds.), *What reading research tells us about diverse children's needs* (pp. 183–218). Mahwah, NJ: Erlbaum.

Beck, I., McKeown, M., & Kucan, L. (2002). *Bringing words to life: Robust vocabulary instruction*. New York, NY: Guilford.

Biemiller, A. (2001). Teaching vocabulary: Early, direct, and sequential. *American Educator, 25*, 25–28.

Brown, J., Frishkoff, G. A., & Eskenazi, M. (2005). *Automatic question generation for vocabulary assessment*. Paper presented at the Human Language Technology. Retrieved from http://www.aclweb.org/anthology/H/H05-1103

Cain, K., Oakhill, J. V., & Lemmon, K. (2004). Individual differences in the inference of word meanings from context: The influence of reading comprehension, vocabulary knowledge, and memory capacity. *Journal of Educational Psychology, 96*, 671–681.

Collins-Thompson, K., & Callan, J. (2007). Automated and human assessment of word definition responses. *Proceedings of the Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL–HLT 2007)*, 476–483.

Daneman, M., & Green, I. (1986). Individual differences in comprehending and producing words in context. *Journal of Memory and Language, 25*, 1–18.

Durso, F. T., & Shore, W. J. (1991). Partial knowledge of word meanings. *Journal of Experimental Psychology: General, 120*, 190–202.

Francis, W. N., & Kucera, H. (1982). *Frequency analysis of English usage*. Boston, MA: Houghton Mifflin.

Frishkoff, G. A., Collins-Thompson, K., Perfetti, C., & Callan, J. (2008). Measuring incremental changes in word knowledge: Experimental validation and implications for learning and assessment. *Behavioral Research Methods, 40*, 907–925.

Frishkoff, G. A., Perfetti, C. A., & Collins-Thompson, K. (2010). Lexical quality in the brain: ERP evidence for robust word learning from context. *Developmental Neuropsychology*, *35*(4), 376–403.

Frishkoff, G. A., Perfetti, C. A., & Westbury, C. (2009). ERP measures of partial semantic knowledge: Left temporal indices of skill differences and lexical quality. *Biological Psychology, 80*(1), 130–147.

Hirsch, E. D. (2003). Reading comprehension requires knowledge—of words and the world. *American Educator*, *27*, 12–20.

Karpicke, J. D., & Roediger, H. L., III. (2007). Expanding retrieval practice promotes short-term retention, but equally spaced retrieval enhances long-term retention. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 33*, 704–719.

Lampinen, J. M., & Faries, J. M. (1994). Levels of semantic constraint and learning novel words. In A. Ram & K. Eiselt (Eds.), *Proceedings of the Sixteenth Annual Conference of the Cognitive Science Society* (pp. 531–536). Hillsdale NJ: Erlbaum.

Landauer, T. K., Kireyev, K., & Panaccione, C. (2011). World Maturity: A new metric for word knowledge. *Scientific Studies of Reading, 15*, 92–108.

McKeown, M. (1985). The acquisition of word meaning from context by children of high and low ability. *Reading Research Quarterly, 20*, 482–496.

McKeown, M., Beck, I., Omanson, R. C., & Pople, M. T. (1985). Some effects of the nature and frequency of vocabulary instruction on the knowledge of use of words. *Reading Research Quarterly, 20*, 522–535.

Nagy, W. E., Anderson, R. C., & Herman, P. R. (1987). Learning word meanings from context during normal reading. *American Educational Research Journal, 24*, 237–270.

National Reading Panel. (2000). *Teaching children to read: An evidence-based assessment of the scientific research literature on reading and its implications for reading instruction*. Washington, DC: National Institute of Child Health and Human Development.

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

Perfetti, C. A., Wlotko, E. W., & Hart, L. A. (2005). Word learning and individual differences in word learning reflected in event-related potentials. *Journal of Experimental Psychology. Learning, Memory, and Cognition, 31*, 1281–1292.

Reichle, E. D., & Perfetti, C. A. (2003). Morphology in word identification: A word-experience model that accounts for morpheme frequency effects. *Scientific Studies of Reading, 7*, 219–238.

Roediger, H. L., & Karpicke, J. D. (2006). Test-enhanced learning: Taking memory tests improves long-term retention. *Psychological Science, 17*, 249–255.

Schmidt, R. A., & Bjork, R. A. (1992). New conceptualizations of practice: Common principles in three paradigms suggest new concepts for training. *Psychological Science, 3*, 207–217.

Stahl, S. A. (1986). Three principles of effective vocabulary instruction. *Journal of Reading, 29*, 662–668.

Stahl, S. (2003). Words are learned incrementally over multiple exposures. *American Educator, 27*, 18–22.

Stahl, S. A., & Fairbanks, M. M. (1986). The effects of vocabulary instruction: A model-based meta-analysis. *Review of Reading Research, 56*, 72–110.

Stanovich, K. E. (1986). Matthew effects in reading: Some consequences of individual differences in the acquisition of literacy. *Reading Research Quarterly, 21*, 360–407.

Swanborn, M. S. L., & de Glopper, K. (1999). Incidental word learning while reading: A meta-analysis. *Review of Educational Research, 69*, 261–285.

Swanborn, M. S. L., & de Glopper, K. (2002). Impact of reading purpose on incidental word learning from context. *Language Learning, 52*, 95–117.

Taylor, W. L. (1953). "Cloze" procedure: A new tool for measuring readability. *Journalism Quarterly*, *30*, 415.

Thorndike, E. L. (1908). Memory for paired associates. *Psychological Review, 15*, 122–138.

Van Daalen-Kapteijns, M. M., Elshout-Mohr, M., & De Glopper, K. (2001). Deriving the meaning of unknown words from multiple contexts. *Language Learning, 51*, 145–181.

Wheeler, M. A., & Roediger, H. L. (1992). Disparate effects of repeated testing: Reconciling Ballard's (1913) and Bartlett's (1932) results. *Psychological Science, 3*, 240–245.