Behavior Research Methods
2008, 40 (4), 907-925
doi: 10.3758/BRM.40.4.907

Measuring incremental changes in word knowledge: Experimental validation and implications for learning and assessment

Gwen A. Frishkoff
University of Pittsburgh, Pittsburgh, Pennsylvania

Kevyn Collins-Thompson
Carnegie Mellon University, Pittsburgh, Pennsylvania

Charles A. Perfetti
University of Pittsburgh, Pittsburgh, Pennsylvania

and

Jamie Callan
Carnegie Mellon University, Pittsburgh, Pennsylvania

The goal of this study was to test a new technique for assessing vocabulary development. This technique is based on an algorithm for scoring the accuracy of word definitions using a continuous scale (Collins-Thompson & Callan, 2007). In an experiment with adult learners, target words were presented in six different sentence contexts, and the number of informative versus misleading contexts was systematically manipulated. Participants generated a target definition after each sentence, and the definition-scoring algorithm was used to assess the degree of accuracy on each trial. We observed incremental improvements in definition accuracy across trials. Moreover, learning curves were sensitive to the proportion of misleading contexts, the use of spaced versus massed practice, and individual differences, demonstrating the utility of this procedure for capturing specific experimental effects on the trajectory of word learning. We discuss the implications of these results for measurement of meaning, vocabulary assessment, and instructional design.

Knowledge of words is essential to success in early education, and it provides an important foundation for learning throughout adulthood (NRP, 2000; Stahl & Fairbanks, 1986). For example, Feingold (1983) found that vocabulary scores in high school predict 35%–40% of the variance in college GPA, which suggests that the breadth of a person's vocabulary is critical to academic success. Vocabulary also plays an important role in the mastery of particular subjects: From astronomy to zoology, practically every domain requires a thorough understanding of words and their meanings.

Given the importance of word knowledge, a fundamental question is how word meanings are acquired. When and how are specific cognitive and linguistic processes engaged during meaning acquisition? What conditions lead to successful word learning? These questions open up a rich set of topics with relevance for semantic theory, language learning, and instructional tools to support vocabulary development.

In the present article, we describe a new method for the assessment of word knowledge, the Markov estimation of semantic association (MESA), and show how this method can be used to expose the trajectory of word learning over time, providing a sensitive tool for studying psychological processes in word learning. We describe results using MESA scores from an experiment that was designed to elicit changes in word knowledge across multiple sentence contexts. After learners had read each sentence, they were asked to generate the meaning of the target word. The accuracy of the meaning derivation was scored using a continuous scale, in which accuracy was a function of how closely the response resembled the correct (dictionary) meaning. Because the definition-scoring measure that is generated by MESA is continuous, it can represent gradual changes in knowledge. We discuss how the trajectory of word learning is affected by the manipulation of instructional contexts, with a focus on the robustness of learning in the presence of misleading contexts (i.e., malapropisms). We further consider how these effects may vary as a function of individual differences in vocabulary and reading-comprehension skill.

The Incremental Nature of Word Learning

One important aspect of word learning is that knowledge of word meanings typically accrues gradually, over

G. A. Frishkoff, gwenf@pitt.edu

multiple exposures to words in context (Nagy, Anderson, & Herman, 1987; Stahl, 2003). Learners acquire some semantic features, and strengthen existing associations, each time that they encounter a word in context. The gradual nature of word learning is a feature of learning in general: Memory representations become more robust as they are strengthened by repeated exposure to a word or other object (Reichle & Perfetti, 2003).

The trajectory of learning is shaped not only by frequency of exposure, but also by the encoding of new information that modifies the semantic content of word representations. A complete and accurate representation of word meaning results from the selection of defining (and the pruning of irrelevant) semantic features that accompany a word in its surrounding context. Thus, a young child learns that the word *dog* applies to canines of various shapes and sizes, but not to cows, which share many, but not all, of the physical features of canines. Similarly, adult comprehension of an abstract word such as *puerile* includes an awareness of the word's valence (good or bad), as well as the specific dimensions of meaning that distinguish *puerile* from related words, such as *young* and *irresponsible* (Osgood, Suci, & Tannenbaum, 1957). Because the different dimensions, or features, of word meaning can be acquired separately, at any given time, word knowledge may be partial rather than "all or none." An adequate theory of word learning should account for this fact and, ultimately, for the processes that determine the quality and robustness of word representations at each stage of meaning acquisition (Perfetti & Hart, 2001, 2002).

In the word-learning literature, words that are partially known have been designated as *frontier words* (Durso & Shore, 1991). Frontier words are familiar in form, but their semantic representation is incomplete, unstable, or both. Such words may be ripe for learning: Because their form (phonological and/or orthographic representation) is familiar, one can dedicate attentional resources to learning the word's meaning, including subtle aspects of its connotative meaning or usage that may require multiple exposures to the word across a variety of contexts. Recent work on the neurocognition of word-semantic processing has described how frontier words may engage qualitatively different processes than do words that are either well known or completely novel (Frishkoff, Perfetti, & Westbury, in press; Ince & Christman, 2002). This work illustrates the complexity of semantic word learning: The acquisition of robust semantic knowledge may rely on multiple neurocognitive stages and processes that do more than simply strengthen existing word representations.

Although the idea of partial word knowledge is not new (Baumann, Kame'enui, & Ash, 1998; Durso & Shore, 1991; McKeown, Beck, Omanson, & Pople, 1985), few empirical studies have examined changes in word knowledge over multiple trials to address specific questions about the trajectory of meaning acquisition. How much and what aspects of knowledge do individuals acquire at any one time? What conditions lead to optimal word learning? Do these conditions depend on the type of word to be learned (e.g., abstract vs. concrete)? How do individual differences in reading skills—including vocabulary and comprehension skills—affect word learning in different contexts?

To help address these questions, we have designed the MESA method for assessing knowledge of individual words. This method 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). The resulting estimate is used to assess word knowledge along a continuous scale. This method has advantages over existing measures, because it can be used for automated and objective assessment of word learning. Furthermore, because the scores are continuous, in principle they can capture "degrees" of word knowledge, leading to fine-grained measurement of individual word representations as they change over time.

The Use of Free Response Data for Assessment of Word Knowledge

An important feature of our measurement method is that it relies on the scoring of free response data. In this respect, it can be classified as a generative—or expressive—measure of word knowledge. The method involves the use of data from tasks that require participants to actively express their knowledge. Other examples of generative tasks include filling in a missing word on a cloze (sentence completion) task, saying the name of a picture, or explaining a word's meaning or definition. Generative tests can be contrasted with receptive tests of knowledge, which require participants to select an answer from among a list of options, rather than to retrieve the answer from memory. Some researchers have proposed that generative tests of knowledge may promote "deeper" or more elaborated processing of stimuli than do receptive tests (Schmidt & Bjork, 1992). The very act of retrieving information from memory requires more active processing, because one needs to select the response from a very large set of possible answers. In turn, active processing has been linked to enhanced learning and retention (for a review, see Roediger & Karpicke, 2006).

In addition to promoting the learning and retention of new knowledge, generative tasks can provide a rich source of information about the "quality" or completeness of this knowledge. This is especially true when the variability in free response data is fully analyzed to reveal different dimensions and gradations of knowledge. Unfortunately, this same variability can present problems for analysis: Traditionally, free response data are coded by hand. As a result, postprocessing and analysis can be difficult, time-consuming, and subjective. Furthermore, manual scoring can be insensitive—that is, limited in its ability to detect subtle differences in response accuracy. This is due to the inherent trade-off between the sophistication of a coding scale, on the one hand, and the ease and consistency with which human judges can use the scale, on the other. For example, a participant may have learned that a particular word has negative connotations, but little more (J. C. Brown, Frishkoff, & Eskenazi, 2005). The usual binary or ternary (2- or 3-point) score may be too coarse grained to reveal these effects. Consequently, manual scores may be limited in their ability to reveal partial word knowledge.

Despite these challenges, a few word-learning studies have used open-ended vocabulary questions, including requests for participant-generated "definitions" of words or, more generally, meaning derivation. For example, van Daalen-Kapteijns, Elshout-Mohr, and de Glopper (2001) asked 11- and 12-year-old participants to derive the meanings of unknown words. Each unfamiliar word appeared in three consecutive contexts. After each context, the participants were asked to explain the meaning of the word. The contexts provided overlapping clues to the word's meaning. Each response was scored manually on a 4-point scale for decontextualization and stability of representations over trials (cumulative testing). In addition, learners were asked to generate a dictionary-like definition after all three contexts had been presented. The accuracy of the definition was scored using a 3-point scale (0 = *inadequate*, 1 = *fairly adequate*, 2 = *adequate*) in which adequacy was determined subjectively, on the basis of the number of semantic "elements" of the target word definition that were judged to have been present in the learner-generated definition. According to the analysis of their results, good readers outscored less skilled readers on all three measures of success (decontextualization, stability of meaning generation over time, and accuracy of dictionary definition). It is interesting to note that van Daalen-Kapteijns et al. did not describe how scores changed as learners were exposed to a word across multiple contexts; scores were aggregated across trials, and a single mean value was reported. Although averaging has the advantage of providing more robust statistics, it also obscures changes in response over time that could reveal important information about word-learning processes.

Two more recent studies have explored the use of meaning-generation tasks for the assessment of word learning. Swanborn and de Glopper (2002) used a free response, definition-generation task, and they scored the accuracy of the participant responses on a 4-point scale. Similarly, Cain, Oakhill, and Lemmon (2004) used a 3-point coding scale. Although both studies reported interesting results with relevance for individual differences in word learning (see below for discussion), a reliance on manual scoring of free response data limits the possibility of replicating and, ultimately, of explaining the experimental results.

MESA

More recently, Collins-Thompson and Callan (2007) presented an automated method, MESA, for scoring the quality of participant definitions along a continuous scale. The MESA score for a free response to a target word is derived from the logarithm of the probability that the response describes the correct word definition. The result is a normalized scale ranging from −1 (the learner's knowledge is far from the target definition) to 0 (an ideal match). This scoring method treats word relations as a conjunction of features such as synonymy, morphology, associative strength, and co-occurrence with other words. These word relations form a network, which is used to define a semantic similarity function between sets of words by analyzing the connection strength between words. Each feature of a word relation can be independently assessed using computational measures that are based on "random walks" on the network (called "Markov chains"; see Toutanova, Manning, & Ng, 2004). In this way, the algorithm can yield continuous and fine-grained measures of partial word knowledge by providing a similarity score that represents the relationship between the words in a target definition and the words in the participant's response.

In an evaluation study, Collins-Thompson and Callan (2007) used MESA to score free response data that had been acquired in a definition-generation task. MESA scores were compared with human judgments that used a 4-point scale, with 0 used to indicate that the response was completely wrong and 3 used to indicate a complete (and accurate) response. Three independent coders scored the data; interrater reliability was moderate to good (.64–.72). The human judgments were adopted as the "gold standard" and were used to compare the accuracy of several automated procedures for scoring the free response data, including latent semantic analysis (LSA; Landauer, Foltz, & Laham, 1998) and the MESA procedure. Results showed a substantially better rank correlation between the MESA scores and the human judgments ($r \approx .61$) than between LSA scores and the human judgments ($r \approx .49$).

PRESENT STUDY

The goal of the present study was to use the MESA algorithm to measure changes in word knowledge across multiple contexts. By allowing measurement of incremental changes in meaning, we intended to examine the effects of context quality, spacing of practice, and skill differences on the trajectory of word learning.

Study Design: Word Learning Across Multiple Contexts

In the present study, adult participants were exposed to very low-frequency words, such as *abrogate*, in the context of visually presented sentences. On most learning trials, contexts were selected to be highly semantically constraining, in order to support accurate inferences about the word's meaning. For example, consider the following sentence:

> This system has been weakened since 1983, and the current Liberal party government seeks to further weaken or *abrogate* it.

A number of words in this context ("weaken," "government," "seek") provide important clues to the meaning of the target word, *abrogate*. In our task, participants saw each target word in six different contexts. After they had seen each context, they were asked to type the meaning of the word (i.e., a close synonym). In general, multiple contexts provide overlapping cues that converge to strengthen the target meaning (Stahl, 2003). Therefore, although we expected to see variability in performance across trials, we expected the change in the definition accuracy, as reflected in the MESA scores, to increase on average with additional exposures to a word in context.

We estimated the overall improvement in definition accuracy for a given word by calculating a robust estimate of the difference between the definition-accuracy scores for the first and last trials (see the Method section for details).

Beyond the expectation that we would see significant changes in performance over time, we had three, more specific, hypotheses regarding the influence of context quality, spacing of practice, and vocabulary skill on measures of word learning.

Hypothesis 1: Effects of Context Quality

Our first hypothesis was that definition accuracy (MESA scores) would vary as a direct function of *context quality*. To test this idea, we varied context quality—the extent to which a given sentence context provides clues to the target word's meaning. Some experimental conditions included sentence contexts that were designed to mislead the learner in a consistent way toward an alternative definition by priming concepts that are associated with a distractor word. Distractor words share a similar form (spelling and sound) with the target word, but the two words have different meanings. An example is the following sentence:

> Traditional distributors . . . *abrogate* [sic] to themselves the role of determining what's proper for their customers to read.

This sentence illustrates a semantically based error in language production, known as a *malapropism*: The writer clearly intended to use the word *arrogate* but instead used a similar-sounding word ("abrogate"), resulting in a semantic error or incongruity (Vitevitch, 1997). In the present study, we introduced malapropisms, or semantic errors, to manipulate the probability that target words would be successfully learned. Some words were presented only in contexts that were supportive—that is, that contained strong (and partially overlapping) cues to the meaning of the target. A second set of words was presented in a series of contexts in which one of the six sentences (17%) was misleading. Finally, a third set of target words occurred in misleading contexts three out of six times (50%), increasing the likelihood of target–distractor confusion and decreasing the likelihood that a participant would learn the correct meaning of the targets. By introducing different ratios of supportive to misleading contexts, we aimed to promote either more-effective or less-effective learning. We expected, in turn, that MESA measures of definition accuracy would vary in response to these manipulations.

Manipulation of context quality enabled us to address questions of practical, as well as theoretical, importance: How robust is word learning from context, when a learner encounters some contexts that are uninformative or, as in the present case, misleading? This question is interesting in theoretical terms, because an adequate model of word learning should be able to account for the effects of context quality on learning. It is also of practical importance, because the presence of "noise," in the form of low-quality contexts, is likely to affect word learning in real-world situations.

Hypothesis 2: Dissociation of Processes Underlying Immediate Versus Delayed Recall

Our second hypothesis was that spacing of practice would have distinct effects on immediate performance (during learning) and delayed recall (1 week later). To provide some background for this hypothesis, we give a brief overview of research on optimal spacing of practice and implications for word learning.

In the past several years, there has been growing interest in identifying the conditions that lead to robust learning (Koedinger & Aleven, 2007; Pavlik & Anderson, 2005). A consistent finding is that conditions that promote increased attention or engagement during learning result in better long-term retention of material. Bjork and Linn (2006) have referred to these test conditions as "desirable difficulties." For example, taking a test not only yields data for assessment, but also leads to more robust learning when compared with passive study. In Roediger and Karpicke (2006), two groups of students were exposed to prose passages and were later asked to recall ideas from each passage. Before the final test, the control group had two additional opportunities to study the passage. The experimental group had one additional opportunity to study the passage and, in lieu of the second opportunity, participants were tested on their recall of the passage content. Immediately (within 5 min) after the study session, the control group performed slightly better than the experimental group. The experimental (retrieval practice) group recalled significantly more ideas from the passage when tested 1 week later, however. In a second experiment, Roediger and Karpicke replicated Wheeler and Roediger (1992) and found a benefit of repeated testing as compared with repeated study opportunities.

Comparisons of widely spaced versus closely spaced (or "massed") practice have likewise yielded different results on immediate versus delayed tests of memory. When a knowledge representation is activated in memory, the strength of the representation decays over time. It is therefore not surprising that testing immediately after practice results in better performance than does testing after a delay. When a test is administered after a significant delay (typically 1 day or more, depending on the instructional and test conditions), however, the effect of spaced versus massed practice is reversed: Studying an item over widely spaced intervals leads to better long-term retention (Pavlik & Anderson, 2005; Schmidt & Bjork, 1992). Spaced practice thus appears to constitute another "desirable difficulty."

The effects of repeated testing and spacing of practice have implications for word learning. In a recent study, Karpicke and Roediger (2007) combined repeated testing and spacing of practice in an associative word-learning paradigm. Participants first studied a pair of words (e.g., *sobriquet–nickname*). One word (*sobriquet*) was the target (unfamiliar, trained) word, and the second (*nickname*) was a familiar word that is a close synonym of the target. Participants then completed three tests of cued recall: The test word was presented, and they were asked to recall the meaning. In the massed-practice condition, the three practice trials for a given word were blocked. In the spaced-practice condition, practice was interspersed

within trials, with an average of five trials separating retrieval practices for a given word. (There were actually two different spaced-practice conditions, but the comparison of these two is not relevant to the present discussion.) Not surprisingly, performance during the learning phase was better in the massed-practice condition than in the spaced-practice conditions: The mean performance over the three tests of cued recall was ~98% in the massed condition versus ~77% in the spaced conditions (Karpicke & Roediger, 2007, Table 1). In addition, reaction times were faster in the massed-practice condition (~2.5 sec vs. ~3.2 sec in the two spaced-practice conditions). The availability of newly encoded information in short-term memory thus led to superior performance, a finding that is consistent with a large body of prior work (Schmidt & Bjork, 1992). By contrast, retrieval practice in the spaced-practice condition produced more robust learning: When participants were tested after a 10-min delay, recall in the massed-practice condition fell to 47%, whereas recall in the spaced-practice condition was ~67% (averaging across the two spaced-practice conditions). The advantage of spaced versus massed practice remained significant when participants were tested two days later.

The lesson from this prior work is that scheduling of practice matters for learning. In fact, it has specific, and sometimes opposite, effects on tests of immediate versus delayed recall. For this reason, we felt that testing for the effects of spacing would provide persuasive evidence that the new measure is sensitive enough to detect important effects in word learning; therefore, we systematically varied the practice schedule for words in our task, rather than randomizing trial order. On the basis of prior vocabulary studies, we predicted that performance during learning in massed practice would be comparable (or even superior) to that in spaced practice, and that the MESA measure would capture this effect. Performance on a second, online task (sentence-level semantic-congruity judgment) provided a secondary measure of short-term learning. We predicted that spaced practice, in contrast to massed practice, would lead to superior long-term retention. We examined this latter hypothesis using a test of target-word knowledge that was administered approximately 1 week before and 1 week after instruction (see the Method section for details on pre- and posttest measures).

Hypothesis 3: Individual Differences in Word Learning From Context

The third factor we considered was the effect of individual differences in word knowledge, as measured by participant scores on a standardized test of vocabulary knowledge (J. I. Brown, Nelson, & Denny, 1973). Such an effect may have important implications for theories of word learning, since it may provide clues to the cognitive processes underlying skilled versus less-skilled performance. For example, if different skill groups respond differently in the spaced-practice condition versus the massed-practice condition, it could motivate additional studies that test hypotheses about the role of working memory, forgetting, or active strategies in word learning—processes that are likely to be reflected in measures of vocabulary and reading comprehension and are also likely to mediate successful word learning.

Skill differences in word learning also have practical implications: The optimal method for vocabulary instruction may be different for different learners, depending on their prior knowledge and skills. Indeed, some studies have shown that learning new words from context can lead to different outcomes for different groups of learners (Daneman & Green, 1986; McKeown, 1985). A typical finding is that older and more-skilled learners can benefit from multiple contexts, whereas younger and less-skilled learners show smaller gains. This finding may reflect the task demands on memory and comprehension, which can affect less-skilled readers disproportionately. For example, Swanborn and de Glopper (2002) reported that low-skilled readers in their task showed only a nominal gain in word knowledge during "incidental" (natural) reading. By contrast, high-skilled readers showed sizeable gains, although outcomes varied depending on the task (e.g., reading for general comprehension vs. reading to learn about a new topic). Similarly, Cain et al. (2004) found that children's ability to infer word meanings from context was related to comprehension and working-memory ability.

Given this prior work, we hypothesized that learners who differed on tests of reading skill—particularly vocabulary knowledge—would show different trajectories of word learning and different effects of intertrial spacing and context quality. We had three specific predictions. First, we predicted that more high-verbal participants would show faster and more robust learning overall (i.e., an interaction between time and skill). This hypothesis is related to the reciprocal effects of vocabulary knowledge and text comprehension that have been deemed the "Matthew effect" (Walberg & Tsai, 1983): Skilled readers acquire more words, which in turn makes it easier for them to learn new words from text—a variation on the adage from the Book of Matthew that the rich get richer, and the poor get poorer (see Biemiller, 2004; Stanovich, 1986).

Second, we expected that more highly skilled readers would be more accurate in detecting misleading contexts and would therefore recover more effectively from the presence of such contexts.

Finally, we conjectured that strong and weak readers would respond differently in the massed-practice condition versus the spaced-practice conditions. Reading ability has been correlated with differences in working memory and has been linked to different reading strategies (Cain et al., 2004; McKeown, 1985; van Daalen-Kapteijns & Elshout-Mohr, 1981); therefore, readers with different skill levels may require different practice schedules to support optimal learning and retention of new words. Evidence for this effect would also provide strong confirmation of the degree of specificity, and sensitivity, of the new method for assessment of word learning.

METHOD

Participants

Participants were recruited from a reading and language prescreening pool, which comprises younger and older adults who have completed a battery of language tests that were developed and

Table 1
Subgroup Scores on Lexical-Knowledge Battery (Perfetti & Hart, 2001; See Also Appendix B)

	High Skilled		Low Skilled		
	M	*SD*	*M*	*SD*	*p*
Vocabulary (NDV)	64.44	4.84	35.56	1.62	<.001
Comprehension (NDC)	24.55	2.21	16.56	1.99	<.001
Nonverbal reasoning (RM)	7.05	3.02	7.42	3.97	n.s.
Spelling (SP)	.36	.19	.38	.13	n.s.
Phonological awareness (PhAT)	.88	.09	.77	.15	n.s.
Decoding skill (PSH)	.46	.23	.36	.21	<.05

administered within the Perfetti Reading and Language Lab (see Appendix B). Prescreening participants were recruited through introductory courses in psychology at the University of Pittsburgh, and through ads that were posted in the campus newspaper. Payment ($7 per hour), academic course credit, or a combination of the two was given in exchange for participation.

A total of 37 individuals who were recruited from the prescreening pool completed both the pretest session and the word-learning session. Twenty-one of these individuals also returned for the third and final (delayed posttest) session 1 week later. All participants were monolingual native English speakers, were right-handed, and had normal or corrected-to-normal vision. Participant ages ranged from 18 to 52; the mean age was 24.10 years ($SD = 10.07$). Four out of 21 participants were male. One participant was left dominant; the rest of the participants were right dominant. Payment ($20 per hour), academic course credit, or a combination of the two was given in exchange for participation.

Table 1 shows the mean scores on all prescreening tests for participants who completed all three experimental sessions ($n = 21$). On the basis of their Nelson–Denny vocabulary scores (see below), participants were divided into two groups: Participants with strong vocabulary skills ($n = 11$) scored in the top half of the distribution, and participants with weak vocabulary skills ($n = 10$) scored in the bottom half. Participants with strong vocabulary skills and those with weak vocabulary skills (henceforth, "high-skilled" and "low-skilled" participants) had significantly different scores on the Nelson–Denny test of comprehension ($p < .001$) and on a pseudohomophone test ($p < .05$). High-skilled participants also had higher mean accuracy on a test of phonological awareness (difference of 11%), but, because there was substantial variability in scores, the group difference on this measure did not reach significance. There was no difference between the two groups in nonverbal awareness or in spelling ability. See Appendix B for additional information about each of these tests, including the number of items and the scoring criteria. No participants reported a history, or a current diagnosis, of any reading or language disorder. All participants had normal or corrected-to-normal vision and no history of hearing loss.

Stimuli and Design

For the word-learning task, we selected 60 pairs of words that are overlapping in form (phonology and orthography) and occur with very low frequency or are rare. The less frequent member of each word pair was designated as the *target* word that was to be learned by our study participants, and the more frequent word was designated as the *distractor*. The distractor was never explicitly presented in the experiment. Instead, it was used to create misleading contexts, or malapropisms (Vitevitch, 1997). Such contexts are not merely unsupportive of the target meaning; they actively pull the participant away from the target meaning and toward the distractor meaning. These "bad" or misleading contexts represent misuses that actually occur in Web and other text corpora and are therefore of practical interest to researchers who are interested in search and retrieval of texts to support research and eductional applications (Collins-Thompson & Callan, 2004). Table 2 shows examples of target–distractor pairs with corresponding congruent (target) and noise (distractor) contexts.

REAP retrieval of contexts. The contexts (sentences) used in this experiment were selected from texts that are freely available on the World Wide Web. REAP software was used for automated and constrained text retrieval (Collins-Thompson & Callan, 2004; reap.cs.cmu.edu/). For the present experiment, REAP was provided with the target words, together with two or three "cohorts" (synonyms, or near-synonyms, of targets). Most of the target words were unambiguous (had a single meaning). For the few words that were polysemous (e.g., *otiose* and *recondite*), the cohorts were sufficient to constrain retrieval to fit the intended meaning of the target word. There were several additional constraints on text retrieval: Sentence length was constrained to fall between 7 and 50 words; only contexts in which the target word appeared as the correct part of speech were selected; and contexts were constrained so that one or two at most were taken from any given document (if possible), to increase context "background topic" diversity.

For each target word, up to 30 sentence contexts were retrieved. These contexts were then sorted according to automatically computed scores that indicated context "informativeness"—that is, degree of sentential constraint. We selected the most constraining contexts, on the assumption that these contexts would provide more clues to the target-word meaning (see the Discussion section for a description of ongoing work in which we are explicitly testing this hypothesis).

In a final pass, the authors modified extraneous or distracting portions of some sentences. For example, obscure names were replaced with more familiar or generic nouns (e.g., *city*, *country*, *president*).

Experimental conditions. The experimental conditions were characterized by a 3 (context quality) × 2 (intertrial spacing) fully crossed design. Context quality was based on the ratio of "good" to "bad" contexts. A good context is supportive of the target meaning, whereas a bad context is supportive of the distractor meaning. In the *NoError* condition, all contexts were supportive. In the *LoError*

Table 2
Example Stimuli for the Experiment

Word Pair	"Good" Context (Congruent With Target)	"Bad" Context (Congruent With Distractor)
Target: ACIDULOUS Distractor: ASSIDUOUS	At times, the playwright allows an ACIDULOUS tone to enter their conversation, criticizing everyone and especially each other.	He had little education, but sought distraction from the dull routine of his job through ACIDULOUS study.
Target: FLAGRANT Distractor: FRAGRANT	The bombing in broad daylight was a FLAGRANT violation of the ceasefire.	The FLAGRANT odor of sandalwood is often used to create sweet-smelling perfumes.
Target: ABDITIVE Distractor: ADDITIVE	His function in the agency was an ABDITIVE one, sometimes even requiring the use of disguise.	The income tax statement clearly showed that the charges were ABDITIVE and would continue increasing every year until his debt was paid.

condition, one of the six contexts was misleading. In the *HiError* condition, half (three of six) of the contexts were misleading.

The order of presentation of good and bad contexts was randomized, subject to constraints on wide versus narrow intertrial spacing. In the narrow-spacing (massed practice) condition, contexts that contained a particular target word occurred 3–5 trials apart. In the wide-spacing (spaced) condition, 14–25 trials separated contexts that contained a particular target word. Each target word was assigned to one of the six experimental conditions, with assignment counterbalanced across study participants. The 60 target–distractor pairs were evenly distributed across the six conditions.

Pre- and Posttraining Tasks

Participants completed three experimental sessions. In the first and last sessions, they completed several tests that were designed to assess their familiarity with the lexical (word or nonword) status and meaning of the target words. The change in accuracy from the pretraining session to the posttraining session was used to assess learning and retention of new knowledge about the target words. These measures provided independent data to be used for cross-validation of the new (MESA) word-learning measure. Although we did not expect to find a perfect correlation between the online (MESA) and offline (pre- to posttraining) measures, we did expect that the two measures would show reliable effects of time (i.e., accuracy gains with training) and context quality (inverse relation between learning gains and number of misleading contexts). We were less certain whether these measures would show effects of intertrial spacing. We theorized that if spacing effects were reliable, then the delayed posttraining measures would show a benefit of spaced practice.

Pretraining session. The pretraining session took place approximately 1 week before the word-learning session and approximately 2 weeks before the posttraining assessment. Participants completed two computer-based tasks in the first session. The first was a lexical-decision task, in which participants were asked to make speeded word–nonword judgments. Word and nonword stimuli were closely matched in orthographic form (word length, orthographic neighborhood), and instructions placed a greater emphasis on accuracy versus speed. These two controls have been shown to promote attention to word meaning versus word form in lexical-decision tasks (Binder et al., 2003). After completing the lexical decision, participants were prompted to indicate their confidence in the word–nonword judgment on a scale from 1 to 4. Word stimuli for the lexical-decision task consisted of the 60 target words, and nonword stimuli were matched in length and orthographic neighborhood with the word stimuli. Nonwords were extracted from the MCWord online database (www.neuro.mcw.edu/wordgen/), using the following search criteria. First, we searched for constrained trigram-based strings, 5–11 letters in length, and we excluded words and repeats. This search yielded 500 pseudowords, 60 of which were selected to be used in the lexical-decision task. Selection was random, except for the constraint that each pseudoword needed to be matched in length and orthographic neighborhood with a corresponding target word. No homophones were used, and all words approximately matched in orthographic frequency. Lexical decisions provide the weakest measure of vocabulary knowledge, since decisions can be based on familiarity with the word form or meaning, even when the meaning of a word is not well known. These data therefore provide clues to participant familiarity with the target words prior to training.

After they had completed the lexical-decision task, participants were asked to complete a multiple-choice test that required them to select the word that they believed was closest in meaning to each target word. There were five possible answers for each item. In each instance, one of the incorrect responses corresponded to the meaning of the distractor word. For example, the correct response for the word *abrogate* in the example below is "yield"; the answer corresponding to the distractor word is "seize":

ABROGATE 1. void 2. seize 3. start 4. finish 5. yield

After they had completed the synonym judgment, participants were asked to indicate their confidence on a scale from 1 (*just guessing*) to 4 (*very confident*). This task was designed to yield two types of information about participant knowledge of the target words: (1) Does the participant know the meaning of the target word? and (2) Is the participant confused about the difference between the target and distractor words? Clearly, if the answer to (1) is "yes," then the answer to (2) is "no." If participants did not know the correct meaning of the target word, it was possible that they had the target and distractor words confused. It was important to assess this potential confusion prior to the training task in order to begin with an accurate baseline measure of target-word knowledge.

Posttraining session. One week after the word-learning experiment, participants returned to complete the third, and final, session. The synonym-judgment task was presented in both the pretraining and posttraining sessions (Forms A and B contained different stimulus orders and different answers, and they were counterbalanced in order for each participant).

Word-Learning Task

In the second session, participants viewed the target words embedded in six different, semantically constraining contexts. The target words were capitalized. Participants made two kinds of responses on each trial: a semantic-congruity (sentence) judgment, and a synonym or definition generation.

Semantic-congruity judgment. As each word-learning context was presented, participants were asked to determine whether the sentence was coherent and to indicate their judgment by pressing one of three keys. The specific instructions were as follows:

> Your task will be to indicate with a button press whether the sentence makes sense. Press:
>
> 1 = if the sentence meaning is CONSISTENT with the meaning of the capitalized word;
>
> 2 = if the sentence meaning is NOT CONSISTENT with the meaning of the capitalized word (i.e., if the sentence as a whole does not makes sense); &
>
> 3 = if you DO NOT KNOW the meaning of the capitalized word.

The following sentence was provided as an example:

> He worked OSTENSIBLY in an effort to finish before they were asked to hand in the tests.
>
> 1 = Makes sense
>
> 2 = Does not make sense
>
> 3 = Don't know

The purpose of the sentence-judgment task was threefold: (1) to encourage deep (semantic) processing of each sentence, (2) to give a measure of error detection, and (3) to give a measure of target-word knowledge.

To keep participants engaged with the task, we informed them that they would receive payment that was based on the accuracy of their responses on the sentence-judgment task. Each correct detection of an error (misleading context) received $0.05, with a maximum $4.00 reward (in addition to the base compensation). Failure to detect an error led to a $0.02 penalty, with a maximum $1.60 penalty. Each incorrect error response (false alarm to good contexts) received a deduction of $0.05, with a maximum $14.00 penalty and a minimum possible score of 0. Correct recognition that good context was supportive resulted in a $0.02 reward, with a maximum $5.60 reward. No response or a response of "I don't know" (0) had no effect on the score. Note that the detection of errors was weighted more heavily than was recognition of supportive contexts and that there was a built-in penalty for guessing, particularly for false detection of errors. This weighting was designed to encourage participants to respond "yes" or "no" only when they were relatively certain of the correct answer. Guessing was penalized, particularly in the detection of incongruous sentences (i.e., misleading contexts). There

was a short practice block before the experiment began, to illustrate scoring to participants and to demonstrate the definition-generation task.

Definition-generation task. After participants had made the sentence judgment, they were asked to generate a succinct definition or synonym for the target word and to type their answer in response to a probe. They were strongly encouraged to give a response on each trial, even if they were uncertain about the meaning of the target word. If there was no response after 10 sec, the experiment advanced to the next learning trial. Participants completed 9 practice trials, followed by four blocks of experimental trials (90 trials per block).

Definition-Scoring (MESA) Algorithm

Participant responses on the definition-generation task were corrected offline for spelling, using the spell-checker in Microsoft Excel. If a response was not only misspelled, but also unintelligible (e.g., "wti"), it was omitted from the analysis.

Spell-corrected responses were then entered into the MESA definition-scoring algorithm. The MESA scores (measures of semantic distance between the participant's response and the correct response) were computed using a statistical model of semantic similarity between texts, as described in Collins-Thompson and Callan (2007). This 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. The relationships are represented by pairs of nodes that are joined by multiple weighted edges, with each edge corresponding to a different link type. Link types include the following.

Stemming. Words are based on common morphology (e.g., *stem* and *stemming*).

Synonymy. Words share a high degree of semantic-feature overlap (e.g., as characterized in WordNet; Miller, 1995) (e.g., *quaff* and *drink*).

Co-occurrence. Words tend to appear together in the same contexts (e.g., *politics* and *election*).

Hypernymy and hyponymy. Categorical elations such as "X is a kind of Y," as obtained from WordNet or other thesaurus-like resources (e.g., *airplane* and *transportation*).

Associative strength. A relation that is defined by the fact that a person is likely to give one word as a free-association response to the other (e.g., *disaster* and *fear*).

Graphs provide a rich model for representing multiple word relationships. They typically use nodes of words, with word labels at the vertices, and with edges denoting word relationships. In our model, the dependency between two words represents a single inference step in which the label of the destination word is inferred from the source word. Multiple inference steps may then be linked together to make long-range inferences about word relations. Each inference step uses a mixture of synonym, stem, co-occurrence, and free-association links. In this way, we can infer the similarity of two terms without requiring direct evidence for the relations between that specific pair. To evaluate this method, Collins-Thompson and Callan (2007) compared the Markov chaining method with three other automated methods for scoring the accuracy of word definitions. The gold standard for this comparison was a set of human ratings (mean kappa $\approx$ .68). The Markov chaining method gave the most accurate results (Collins-Thompson & Callan, 2007).

The present study used the Markov procedure to compute two complementary measures of learning on each trial. First, a measure of target accuracy was estimated using the semantic distance between the free-response and the correct (target) definition. Second, a measure of target–distractor confusion captured the semantic distance between the participant's response and the distractor-word meaning. This latter score allowed us to measure the extent to which the malapropisms (bad contexts) had effectively misled participants.

After the responses were scored, we measured the change in scores across trials: The amount of change was treated as an outcome measure that indicated degree of word learning. Rather than computing the absolute difference between scores for the first and last trial, we performed a linear regression using all the trial values and using the resulting slope as the change measure. To further reduce parameter bias, this regression was implemented using a simple, widely used estimator called the *jackknife estimator*. If there are N trials, the jackknife estimate of the slope is simply the average of the slopes from N separate regressions, where each subregression leaves out a different trial-data point. These robust estimates are less sensitive to noise and to missing values.

RESULTS

Our primary goal in the present set of analyses was to evaluate word-learning outcomes as indicated by the incremental learning measure. To ground our interpretation of these data, we began by analyzing effects from the pre- and posttest vocabulary assessments. Because the posttest session was administered 1 week after the word-learning session, we refer to increases in accuracy from pre- to posttest as *long-term retention* (or, conversely, as *long-term forgetting* if there is a decrease in performance from pre- to posttest). The retention measure provides cross-validation of evidence for learning. More generally, the retention measure provides independent evidence regarding the effectiveness of the experimental manipulations.

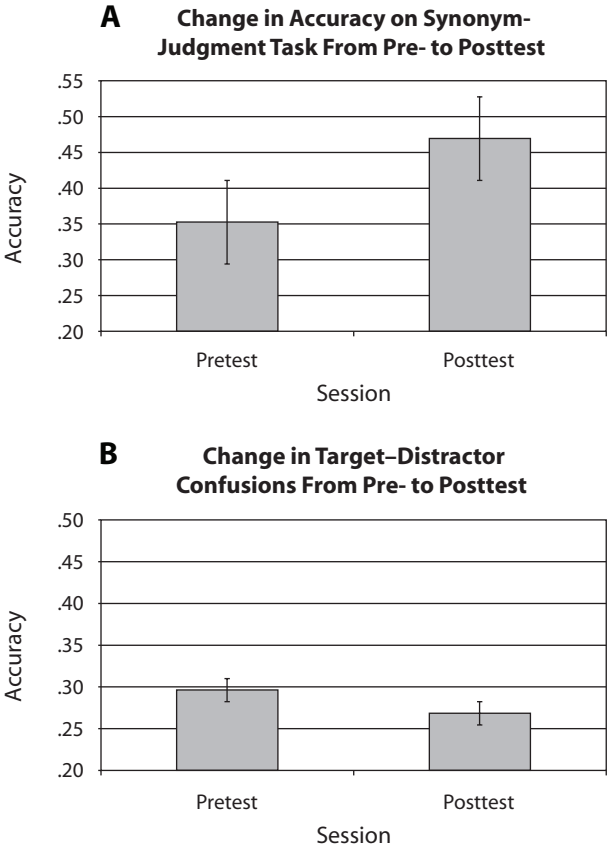


Figure 1. (A) Mean accurate responses and (B) mean target–distractor confusions on the synonym-judgment task. The *y*-axis on both graphs extends from .20 (chance level) to ~.50.

Pre- and Posttest Results for Synonym-Judgment Task

We subjected scores on the pre- and posttest tasks to a mixed 2 (session) × 2 (spacing) × 3 (context quality) ANOVA, with vocabulary skill (high vs. low) as the between-subjects variable. Recall that the design of the synonym-judgment task allowed us to identify changes in target–distractor confusions, as well as changes in accuracy of target-word knowledge, from pre- to posttest. We therefore analyzed these two measures in separate ANOVAs.

Pre- to posttest gains in target-word knowledge. As we expected, there was a reliable increase in target-word knowledge from pre- to posttest [session main effect, $F(1,20) = 21.83$, $MS_e = .828$, $p < .001$] (Figure 1A). On average, there was an increase of ~12% in accuracy from pre- to posttest, corresponding to the addition, and retention (over 1 week), of approximately seven new words. There was also a marginal decrease in target–distractor confusions [session main effect, $F(1,20) = 4.11$, $MS_e = .046$, $p = .057$] (Figure 1B). This effect corresponds to the elimination, on average, of two target–distractor confusions from pre- to posttest.

Effects of context quality. Both the target accuracy and target–distractor confusion scores showed effects of context quality. As illustrated in Figure 2A, the increase in accuracy from pre- to posttest interacted with the number of misleading contexts [session × context quality, $F(1,20) = 4.13$, $MS_e = .049$, $p < .05$]. Figure 2B shows the complementary effect of context quality on changes in target–distractor confusion: As the number of misleading contexts increases, there is a corresponding increase in confusion [session × context quality, $F(1,20) = 4.13$, $MS_e = .049$, $p < .05$]. Figure 2B also shows a marginal effect of spacing on this two-way interaction [spacing × session × context quality, $F(1,20) = 2.29$, $MS_e = .025$, $p = .11$]. Note that although the three-way interaction with session failed to reach statistical sig-

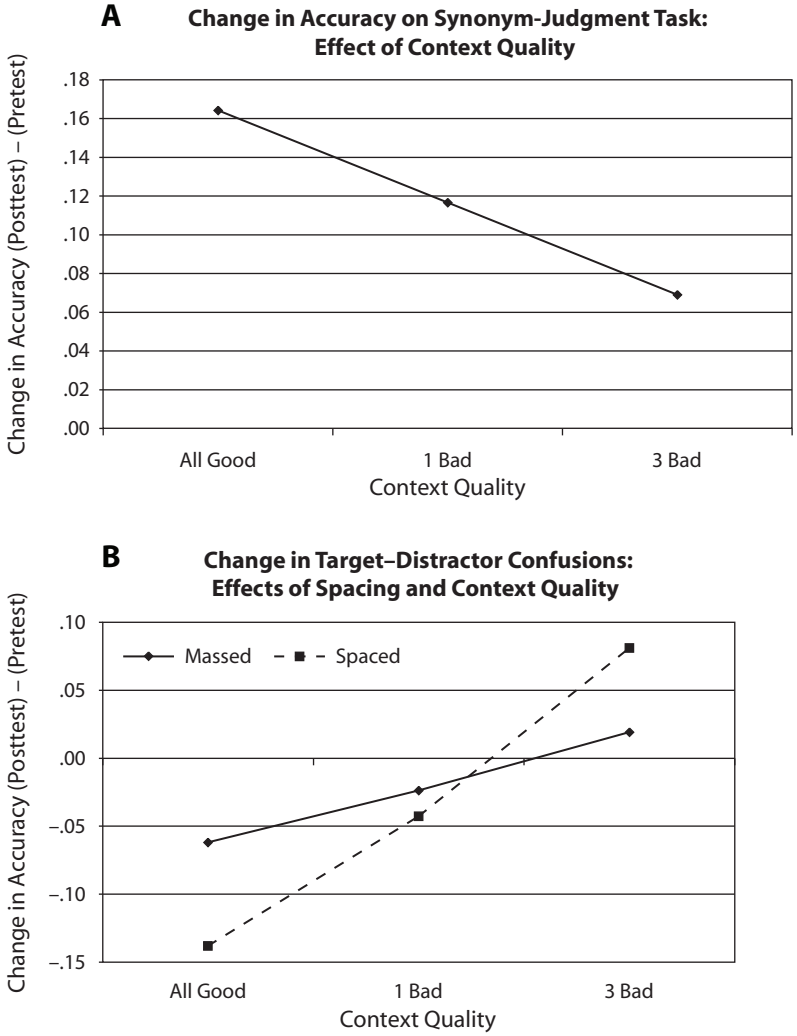


Figure 2. (A) Effects of context quality on change in accuracy on the synonym-judgment task from pre- to posttest. (B) Effects of context quality on change in target–distractor confusions on the synonym-judgment task from pre- to posttest. The solid line indicates closely spaced ("massed") trials. The dotted line indicates widely spaced trials.

nificance, the interaction of spacing × context quality did reach significance [$F(1,20) = 3.87$, $MS_e = .097$, $p < .05$]. In general, there were larger training (session) effects in the spaced condition than in the massed condition, a finding that is consistent with previous work on scheduling of practice during vocabulary instruction (Pavlik & Anderson, 2005).

Post hoc comparisons showed a nonsignificant difference in learning (posttest − pretest score) between the NoError and LoError conditions [$t(20) = 1.64$, $p = .11$]. There was a significant difference between the HiError and NoError conditions [$t(20) = 2.76$, $p < .05$]. Although learning was less robust in the HiError condition, the increase in synonym-judgment score from pre- to posttest was reliably greater than zero [$t(20) = 2.14$, $p < .005$].

Effects of vocabulary skill. As expected, participants who scored higher on the Nelson–Denny test of vocabulary knowledge had higher scores overall on the synonym-judgment test [main effect of vocabulary, $F(1,20) = 5.04$, $MS_e = .091$, $p < .05$]. In addition, there was an unreliable trend toward an interaction of vocabulary skill and gains in accuracy pre- to posttest [session × vocabulary, $F(1,20) = 2.62$, $MS_e = .099$, $p = .11$]. On average, high-skilled participants responded correctly to approximately 22 out of the 60 target words on the pretest, and approximately 32 words on the posttest, an average gain of 10 new words (note that 12 out of 60 words = chance level of responding). By contrast, low-skilled participants responded correctly to an average of 20 words on the pretest and 24 words on the posttest, an average gain, and weeklong retention, of only 4 new words. In addition, there was a statistically significant interaction of spacing × vocabulary [$F(1,20) = 11.84$, $MS_e = .179$, $p < .01$]. The most meaningful comparison is between performance

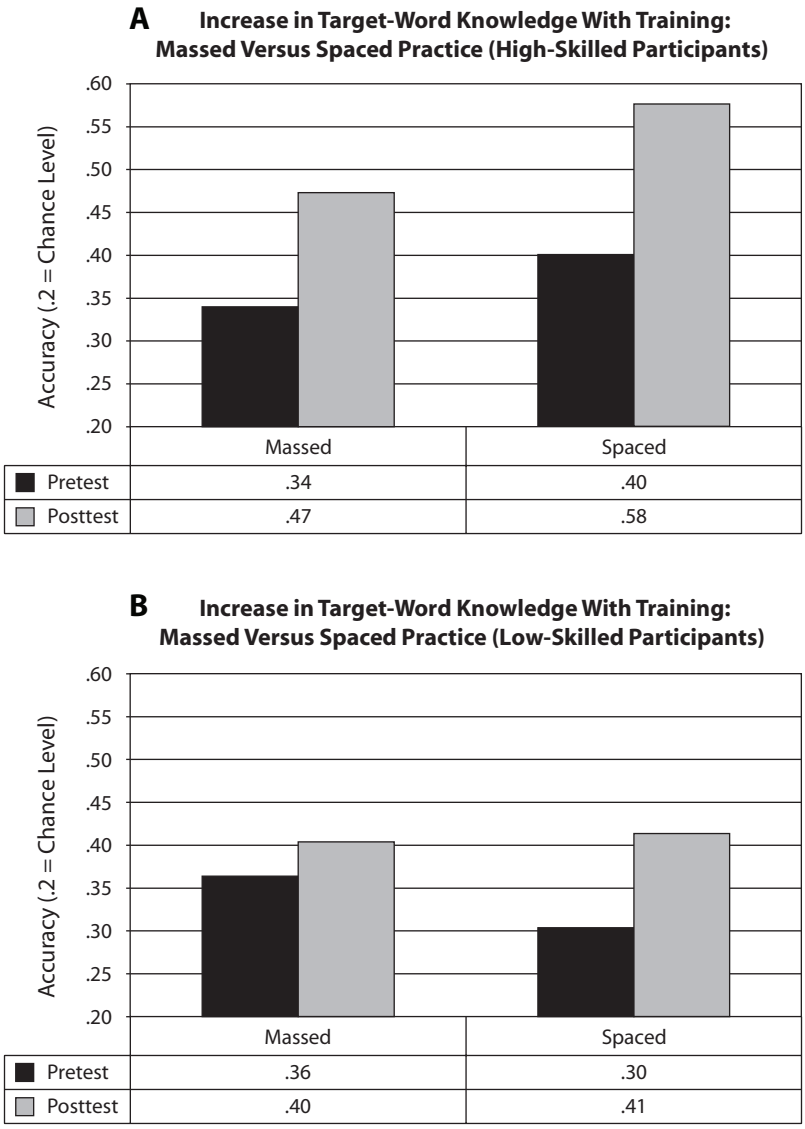


Figure 3. Mean accurate responses for (A) high-skilled participants and (B) low-skilled participants in closely spaced ("massed") versus widely spaced trials. The black bars indicate pretest scores. The gray bars indicate posttest scores. The *y*-axis on both graphs extends from 20% (chance level) to 60%.

on the posttest for words that were trained in the spaced condition versus those trained in the massed condition. Post hoc comparisons revealed that posttest scores were higher in the spaced condition than in the massed condition for high-verbal participants ($p < .01$) but not for low-verbal participants ($p > .7$). This pattern is consistent with the hypothesis that high-verbal participants would benefit more from spaced versus massed practice than would low-verbal participants. Although these results are intriguing, they should be interpreted with caution, given that the three-way interaction of spacing, vocabulary, and session was not significant.

The high-skill and low-skill groups did not differ in pre- or posttest measures of target–distractor confusion.

Effects of intertrial spacing. There was a marginal interaction between spacing and session, with larger gains in accuracy in the spaced condition than in the massed-practice condition [spacing × session, $F(1,20) = 3.74$, $MS_e = .050$, $p = .06$] (see Figure 3).

In summary, analysis of mean performance on the synonym-judgment task revealed effects of session (i.e., training effects). Session effects interacted with context quality and intertrial spacing, as well as with vocabulary skill. These results demonstrate that the independent variables manipulated in the present study behaved as expected on a conventional test of word knowledge. The presence of these effects provides a useful framework for interpreting effects observed with MESA measures of definition accuracy. These effects are described in the following section.

MESA Results

To ground our interpretation of the MESA results, we first present some concrete examples of MESA scores for individual words and participants. These examples illustrate how MESA can be used to detect incremental changes in target-word knowledge across trials. Subsequently, we present analyses of variance, where the dependent measures consist of the average change in MESA scores across trials for each of the six experimental conditions.

Individual trial data. Example 1 provides a concrete example of how MESA scores may reflect successful learning of a target word (*abditive*) across multiple supportive contexts. Recall that scores can range from −1 (response is far from the correct definition) to 0 (ideal response). Note how the scores increase gradually across trials: The first four trials show an accumulation of knowledge that culminates with the responses on Trials 4–6, which show a correct understanding of the target-word meaning.

Example 1
Target Word, *Abditive* (Hidden);
Participant ID 1856; Narrow Intertrial Spacing

Trial No.	Response	MESA	Context Quality
1	avoidant	−1.00	GOOD
2	attitude	−.83	GOOD
3	sneaky	−.50	GOOD
4	secretive	−.35	GOOD
5	secretive	−.35	GOOD
6	hidden	−.34	GOOD

Example 2 illustrates how MESA scores may reflect a failure to learn the target word. More specifically, this participant (7476) correctly inferred the target-word meaning from the first two contexts, which were supportive, but was subsequently misled by the bad contexts (Trials 3–5). In this instance, the participant did not recover the original understanding on Trial 6, even though the final context was supportive. The response on the final trial ("ridiculous") may well be a metalinguistic commentary on the task itself, reflecting confusion about the meaning of the word, given the inconsistent cues that were provided across the six contexts.

Example 2
Target Word, *Abditive* (Hidden);
Participant ID 7476; Narrow Intertrial Spacing

Trial No.	Response	MESA	Context Quality
1	hidden	−.34	GOOD
2	secretive	−.35	GOOD
3	fake	−.91	BAD
4	accumulating	−1.00	BAD
5	cumulative	−1.00	BAD
6	ridiculous	−1.00	GOOD

Finally, Example 3 illustrates learning that is robust, even in the presence of misleading contexts. The first several trials yielded inconsistent responses to the target word (*bibulous*), reflecting the alternation of good and bad contexts. On Trial 4, the participant gave a correct response, and she continued to give this response on Trials 5 and 6, even though the fifth sentence was misleading. Indeed, on the sentence-congruity task, the participant correctly judged that the sentence on Trial 5 was incongruous. Having recognized that the sentence was misleading, the participant ignored this context and gave the correct response (i.e., the target, rather than the distractor, word definition), reflecting confidence in her knowledge of the target word.

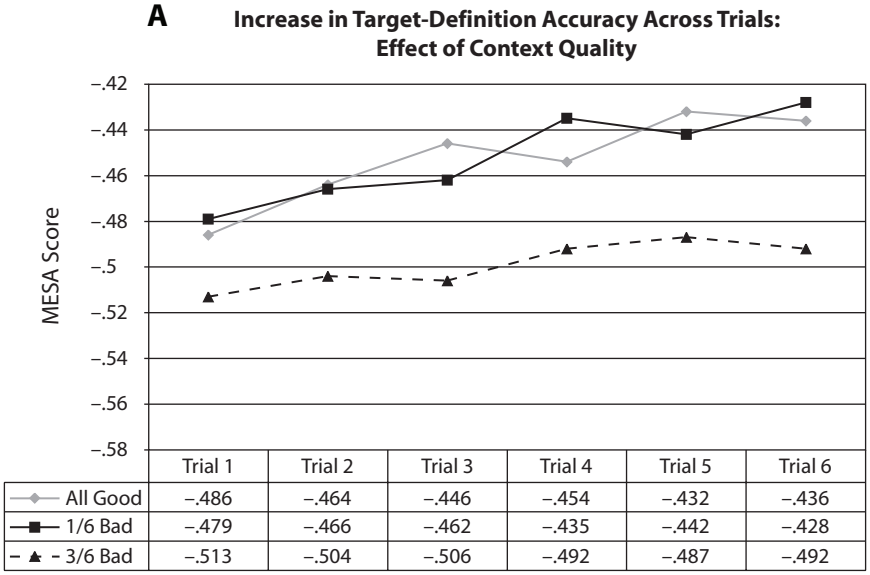
Example 3
Target Word, *Bibulous* (Drunken);
Participant ID 5545; Wide Intertrial Spacing

Trial No.	Response	MESA	Context Quality
1	spacey	−.72	GOOD
2	aggressive	−.67	BAD
3	NR	–	GOOD
4	drunk	−.24	GOOD
5	drunk	−.24	BAD
6	drunk	−.24	GOOD

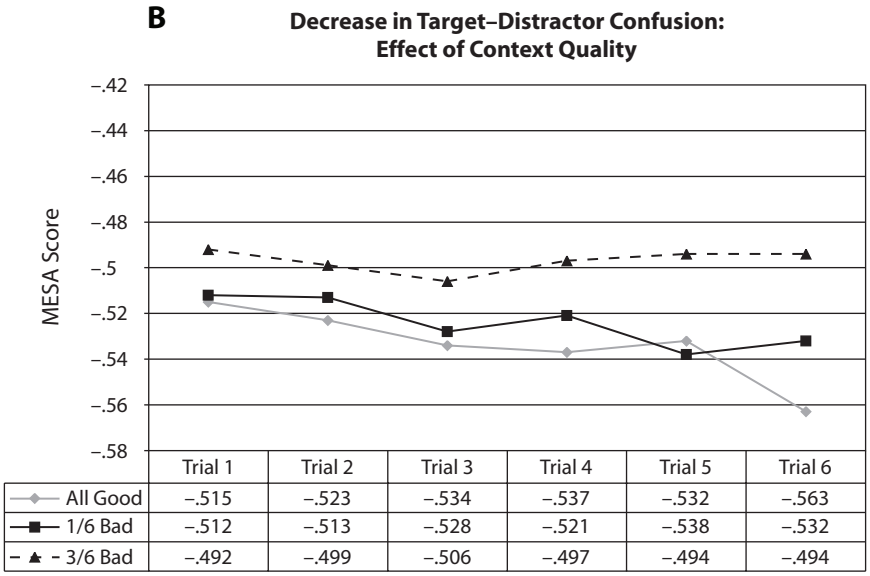
Analyses of variance. For statistical analysis of participant responses in the definition-generation task, we treated the MESA scores—indicating distance between the participant's response and the target-word meaning—as the dependent measure in a mixed ANOVA. Within-subjects variables were time (Trials 1–6), context quality (all good, 1/6 bad, 3/6 bad), and spacing (massed vs. spaced practice). Vocabulary skill was entered as the between-subjects variable. Results indicated a signifi-

cant increase in definition accuracy across trials [main effect of trial, $F(5,37) = 17.03$, $p < .001$]. In addition, the increase in definition accuracy was qualified by an interaction of trial × context quality [$F(10,70) = 1.97$, $p < .05$] (Figure 4A). We also examined effects of context quality and spacing on MESA scores that were computed to reflect the closeness of the target-word definition to the distractor meaning. This analysis revealed a main effect of trial [$F(5,37) = 4.40$, $p < .01$]. As illustrated in Figure 4B, target–distractor confusions tended to decrease over time. There was also an interaction between trial and context quality [$F(10,70) = 2.60$, $p < .01$], consistent with an increase in target–distractor confusion in the presence of misleading contexts (Figure 4B). This result suggests that as the number of misleading contexts increased, definition accuracy was systematically pulled toward the distractor-word meaning. No other effects were significant in the analyses of target–distractor confusion scores.

The change in MESA scores over time appeared to be strongly linear. To further examine effects of context quality, therefore, we computed robust measures of the slope (i.e., the change in MESA scores across trials) and treated the slope as the dependent measure. Our goal was to determine more precisely how word learning (as measured by the slope over MESA scores for Trials 1–6) was affected by the presence of one or more misleading contexts. A



	Trial 1	Trial 2	Trial 3	Trial 4	Trial 5	Trial 6
All Good	–.486	–.464	–.446	–.454	–.432	–.436
1/6 Bad	–.479	–.466	–.462	–.435	–.442	–.428
3/6 Bad	–.513	–.504	–.506	–.492	–.487	–.492



	Trial 1	Trial 2	Trial 3	Trial 4	Trial 5	Trial 6
All Good	–.515	–.523	–.534	–.537	–.532	–.563
1/6 Bad	–.512	–.513	–.528	–.521	–.538	–.532
3/6 Bad	–.492	–.499	–.506	–.497	–.494	–.494

Figure 4. (A) Change in target-definition accuracy (MESA scores that compare participant response with distractor-word meaning) across trials. (B) Change in target–distractor confusion (MESA scores that compare participant response to distractor-word meaning) across trials.

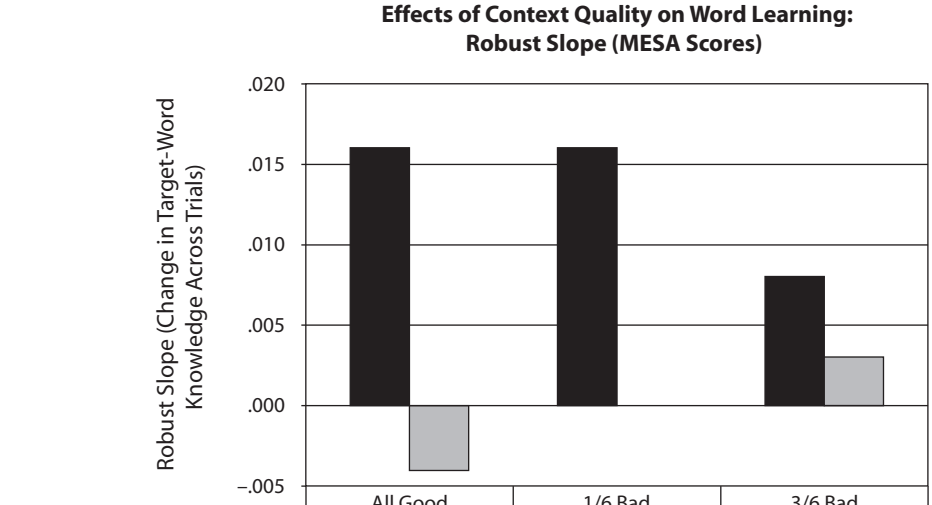


Figure 5. Effects of context quality on the rate of increase in word learning, as measured by robust slope. The black bars indicate slope over MESA scores for target-definition accuracy. The gray bars indicate slope over MESA scores for target–distractor confusion (comparing participant response to distractor-word meaning).

new ANOVA was conducted with slope as the dependent measure, context quality and spacing as within-subjects variables, and vocabulary skill as the between-subjects variable. Results showed an effect of context quality [$F(2,26) = 5.69, p < .01$] consistent with the interaction between trial and context quality in our earlier analysis. This effect is shown in Figure 5 (black bars).

The slope was positive and significantly different from 0 in the NoError condition [$t(27) = 6.67, p < .001$]. Furthermore, the slopes for the NoError and LoError conditions did not differ ($p > .9$), suggesting that there was no decrement in learning (slope) in the presence of a single misleading context. By contrast, the slope was significantly less positive in the HiError condition than it was in the LoError condition [$t(27) = 3.00, p < .01$], although the slope was still greater than 0 in the HiError condition [$t(27) = 4.40, p < .001$].

In a similar analysis, we entered the slope of the target–distractor confusion scores as the dependent variable (see Figure 5, gray bars). ANOVA results showed an effect of context quality [$F(2,26) = 3.31, p < .05$]. Post hoc comparisons indicated that the slope in the NoError condition was significantly less than 0 [$t(27) = -2.19, p < .05$], suggesting that multiple supportive contexts helped reduce confusion between the target and distractor words over time. The slopes in the other two conditions did not differ from 0.

In addition to effects of context quality, our analyses revealed one other effect, a significant four-way interaction of vocabulary skill × spacing × trial × context quality [$F(10,350) = 2.69, p < .05$]. As illustrated in Figures 6A–6D, the main difference occurred in the response of high-skilled readers to errors in the massed-practice condition: Although the HiError (3/6 bad context) condition caused a decrement in learning in the spaced-practice condition for both groups, only the low-skilled readers showed this same decrement when the trials were closely spaced.

Behavioral-Task Results for High- and Low-Skill Groups

Vocabulary skill was seen to affect the trajectory of word learning (the MESA scores), as well as measures of long-term retention (pre- vs. posttest scores). To help interpret these effects, we examined behavioral-task results with an eye to skill differences on each of the dependent measures. Table 3 shows the percent accuracy for the two skill subgroups on each task. As described above, the two groups differed in their knowledge of target-word meanings [$t(20) = 2.25, p < .05$]. In addition, their performance on the posttest measure of distractor-word knowledge revealed a significant difference [$t(20) = 3.29, p < .01$].

Table 3
Accuracy on Behavioral Task Measures for High- and Low-Skilled Participants (Collapsed Over Experimental Conditions)

	High Skilled		Low Skilled		
	M	*SD*	*M*	*SD*	*p*
Familiarity with target words (lexical-decision task)	.67	.14	.71	.14	n.s.
Knowledge of target-word meanings (synonym-judgment task)	.45	.09	.37	.05	<.05
Knowledge of distractor-word meanings (antonym-judgment task)	.60	.09	.49	.07	<.01
Sentence-judgment task (good sentences)	.60	.12	.55	.17	n.s.
Sentence-judgment task (detection of errors)	.34	.14	.32	.16	n.s.

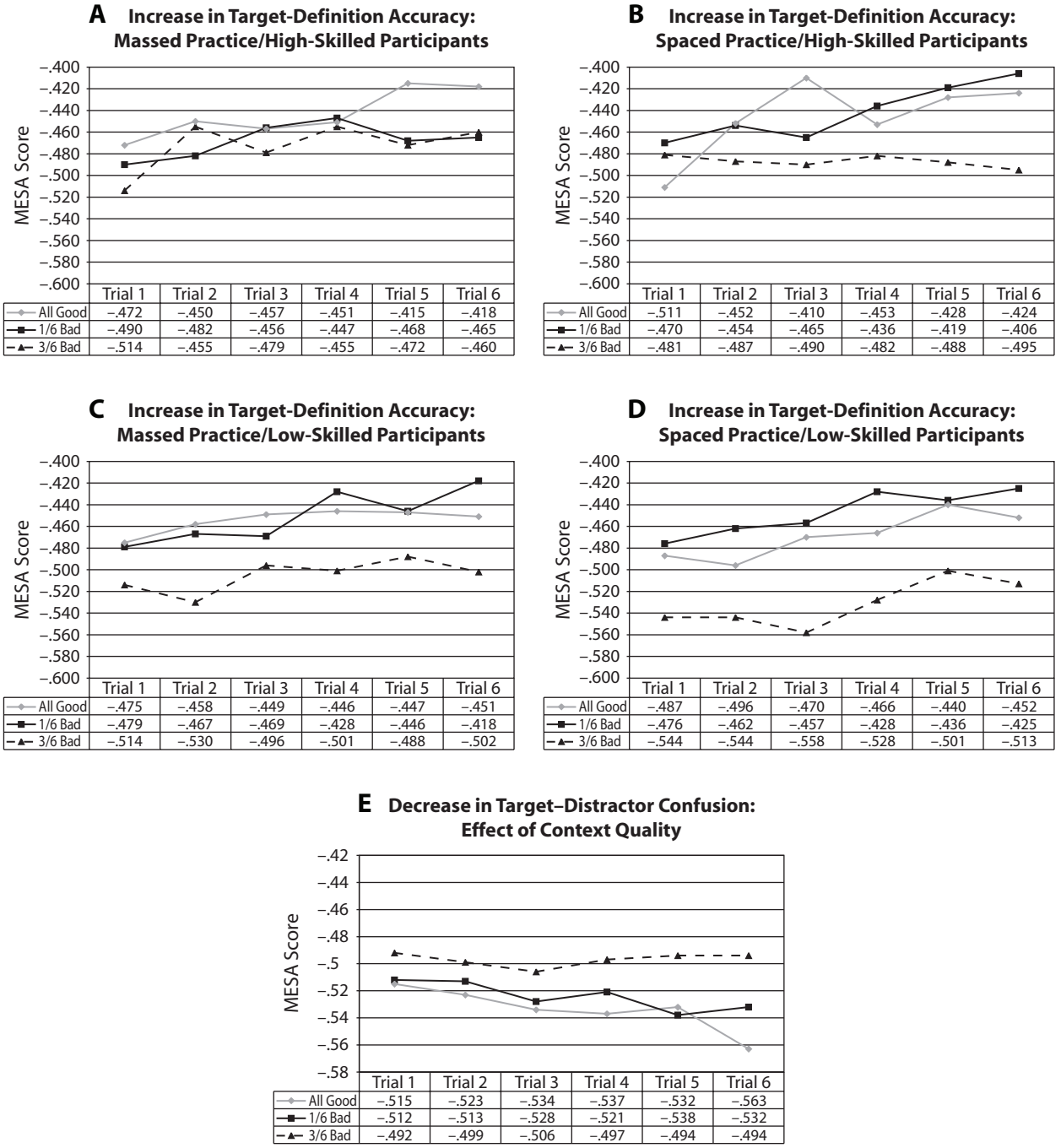


Figure 6. Change in target-definition accuracy (MESA scores that compare participant's response with distractor-word meaning) for (A) high-skilled participants in the massed-practice condition, (B) low-skilled participants in the massed-practice condition, (C) high-skilled participants in the spaced-practice condition, and (D) low-skilled participants in the spaced-practice condition. (E) Change in target/distractor word confusion (collapsed across high- and low-skilled participants).

Interpretation of this effect is complicated by the fact that we do not have pretest measures of distractor-word knowledge. This was a calculated trade-off, however: Exposing participants to both the target and the distractor words prior to training would have introduced another source of target–distractor confusion, making it harder to interpret the results of our context manipulations. Nonetheless, we cannot rule out the possibility that skill differences in knowledge of distractor-word meanings were altered by the training.

Interestingly, the two groups did not differ in mean accuracy on the lexical-decision task, although the meaning-judgment tasks did reveal differences in their knowledge of target and distractor words. There was a difference in accuracy on the sentence-judgment task, but it did not reach significance. There was no difference in mean

confidence for the two groups on any of the pre- or posttraining tasks.

Finally, given the strong correlation between vocabulary and comprehension ($r > .75$), we checked to see whether the group difference in knowledge of target- and distractor-word meanings would remain significant after we had accounted for variance that was due to comprehension skill. We ran a partial correlation of vocabulary skill with scores on the synonym- and antonym-judgment tasks. Results indicated that knowledge of target words was still correlated with vocabulary skill after we had accounted for the contribution of comprehension scores ($r = .55$, $p < .05$). The correlation of vocabulary skill with knowledge of distractor words also remained significant when we accounted for the variance that was due to comprehension scores ($r = .65$, $p < .01$).

Summary of Results

As predicted, the MESA scores showed a gradual increase in word knowledge across trials, a finding that is consistent with the idea that meaning is acquired incrementally. In addition, word learning was modulated by context quality—that is, by the presence of more informative or less informative cues to the target-word meaning. Not surprisingly, we found that the slope of the change in semantic knowledge was greatest in the NoError condition. Importantly, however, analyses suggested that there was no decrement in short-term learning in the presence of a single error (LoError condition). At the same time, there was a trend toward a difference between pre- and posttest accuracy on the synonym-judgment test. It may be premature, therefore, to conclude that a single misleading context has no discernible effect on learning. Learning was clearly affected in the HiError condition. Although the change in definition accuracy during the word-learning session was still positive, the slope of the MESA scores was not as steep as it was in the NoError and LoError conditions. Similarly, the increase in knowledge from pre- to posttest was smaller in the HiError condition than in the NoError condition.

In addition to the effects of context quality, there was an interaction between reading skill, context quality, and spacing of practice. High-skilled readers recovered more effectively from errors than did low-skilled readers, as evidenced by the increased quality of their responses across trials—but only in the narrow-spacing condition. Our tentative interpretation is that high-skilled readers were engaging in a "deeper" strategy when they detected inconsistencies or errors in word usage. Perhaps this led them to engage in an active comparison of the multiple contexts, which is only feasible when the contexts are closely spaced. If so, this strategy appeared to help them avoid the mistake of using cues from the misdirective contexts to identify the wrong target-word meaning. This idea is consistent with the observation that high-skilled readers knew more of the distractor words, as well as the target words, according to results from the synonym- and antonym-judgment tasks. This knowledge would have enabled them to detect word-to-text incongruities more readily and also to compare information on each trial with their prior knowledge.

Spacing of practice was expected to have opposite effects on immediate performance versus delayed recall of word meanings. On the basis of prior studies (Karpicke & Roediger, 2007), we expected that widely spaced practice would lead to superior learning and retention on delayed tests of memory, but that massed practice would result in better immediate performance. This hypothesis was partially confirmed. Spaced practice did result in better learning and retention of words, according to the pre- versus posttest measures (synonym judgment). The main effect of spacing on immediate performance (MESA curves), however, did not approach significance. There are several possible explanations for our failure to observe this effect. First, to our knowledge the present study is the first to examine spacing of practice effects on word learning from context. Most prior studies of vocabulary have used associative-learning tasks (e.g., Karpicke & Roediger, 2007; Pavlik & Anderson, 2005). It is possible that word learning from multiple contexts elicits more complex processes, such as inferencing and integration (see, e.g., McKeown, 1985). This complexity could obscure effects of spacing that are due to simple forgetting (Pavlik & Anderson, 2005). Indeed, our explanation for the interaction of spacing, context quality, and vocabulary skill was based on an assumption that differences were due to an active process of comparison and integration of meaning across contexts, rather than passive decay of the memory trace.

DISCUSSION

In this final section, we discuss some implications of the present experimental results for theories of meaning acquisition and for the design of instructional tools to support reading and language development. We also point out features of the study design that may limit the generalizability of these results, which suggests the need for additional studies. We conclude by noting ongoing work, including computational modeling and designs for adaptive-vocabulary tutors that use MESA to detect specific dimensions of knowledge that are well established in memory or that may be missing or in need of additional practice.

Implications for Theories of Meaning Acquisition

The ability to measure partial word knowledge may open up new possibilities for understanding basic processes and stages of meaning acquisition. For example, MESA scores can be tuned to detect knowledge of specific meaning subcomponents. An interesting question for future applications is whether particular dimensions of meaning are more "basic," or more easily acquired. For example, a large body of work has suggested that word meaning is largely captured by three dimensions: evaluation (good–bad), potency (strong–weak), and activity (active–passive) (Osgood et al., 1957). It is therefore possible that these meaning dimensions are acquired before other, less salient, dimensions. In support of this idea, J. C. Brown et al. (2005) designed a variety of computerized assessments of word knowledge and found that par-

ticipants performed the fastest and the most accurately when they were asked to make speeded judgments about the Osgood semantic dimensions (e.g., good–bad, strong–weak). These results are consistent with our hypothesis that Osgood-derived measures align with early stages of word learning.

Implications for Instructional Design

We designed the randomization of our trials very carefully so that a range of these learning parameters were "sampled" to create each participant's set of trials. This design made it possible for us to use the data that were obtained in these experiments as training or validation data for some existing mathematical models of word learning (Pavlik & Anderson, 2005). In particular, given its ability to detect small changes in word knowledge, MESA could be modified and extended to support an adaptive framework for vocabulary training. Furthermore, because this method is largely automated, it may be possible to use it to develop an adaptive framework for instruction. In this context, *adaptive* refers to the ability to shape instructional events on each trial in response to learner performance, which can provide clues about the specific words (and word components) that the learner needs to practice at a given point in time. The automated nature of this method allows it to be effective, in principle, because it provides instructional materials that are selected to optimize learning by specifying constraints on selection and presentation of instructional contexts. For example, if we detect that a learner is confused about the connotation (positive or negative valence) of a word, then on subsequent learning trials we can provide feedback to the learner to highlight this particular feature of the word. In a paper that is currently in progress, we are building computational models to show how this method might work (Frishkoff, White, & Perfetti, in press).

Limitations and Future Directions

Additional studies are needed to test the generalizability of these results to other instructional contexts and to develop models of word learning that can account for stimulus-, learner-, and task-specific variables. In this section, we consider some specific questions for future research.

Generalizing results to other task contexts. Word learning is influenced not only by the frequency and the quality of practice, but also by task-specific variables, including task strategies and goals, instructional materials, and the type of word to be learned. Studies of word learning from context have tended to emphasize either incidental (nondirected) or intentional (directed) learning. Incidental learning, by definition, may recruit less active strategies. Recall that in the present experiment we observed a four-way interaction of vocabulary skill and word learning from context as a function of massed versus spaced practice. We conjectured that this effect was due to active metacognitive strategies in the narrow-spacing condition among high-skilled readers (see McKeown, 1985; van Daalen-Kapteijns & Elshout-Mohr, 1981). If this is true, then we might not expect to see this same pattern in a different task context, one in which metacognitive strategies play a lesser role.

Similarly, the nature of the training contexts themselves is likely to influence learning outcomes and task strategies. There has been a growing interest among reading researchers in identifying the properties of texts that lead to optimal word learning (Nagy et al., 1987; see Graesser, McNamara, Louwerse, & Cai, 2004, for a description of over 200 text-based properties that can be automatically computed), as well as in broader notions of "context" that recognize the important role of background knowledge and online inferencing in comprehension (McNamara, 2001; Rapaport, 2003, 2005). Among other things, this work has shown the importance of text attributes such as length and readability in predicting reading comprehension and vocabulary gains from "incidental" reading.

In the present study, it is important to note that we used single-sentence contexts to promote word learning. We used brief, as opposed to more extended, contexts for purely practical reasons: Our goal was to provide materials to support word learning over multiple contexts within a single, 2-h session. This necessitated that the contexts themselves be brief. Future work will be required to test the generalizability of our findings to contexts that are longer and that vary in other ways from the contexts used in the present study. Moreover, it will be important to conduct additional studies comparing the efficacy of shorter versus larger stretches of text, particularly given that—somewhat surprisingly—shorter passages have been linked to better word-learning outcomes in some prior work (Graham & Watts, 1990). Cain et al. (2004) have suggested that it might be harder to learn words in context when the relevant cues are spatially separated from the target, particularly for younger and less-skilled readers; they have further linked this effect to working memory limitations. This result might explain why longer passages, although they provide additional cues to word meaning, might be more difficult to process (and may therefore be less optimal for word learning). It also underscores the importance of considering how variables such as readability, informativeness, and the number, distribution, and types of semantic cues might affect word learning in different contexts. The sheer number of textual variables to consider suggests a strong need for systematic and programmatic research in this area.

It is also important to note the nature of the word stimuli that were selected for the present experiment. Our first requirement was that the trained words be very low frequency, or "rare." Words such as *irenic* and *discinct* clearly meet this criterion. In other work (Frishkoff & Perfetti, in press), we have described the differences between word-learning effects for very rare words versus those for more familiar, or "frontier," words. We believe that these differences may exist because very rare words are novel in form, as well as in meaning. The words that were used in the present study are also somewhat unusual, in this sense, because they possess a lexical "form-mate"—that is, a word that is similar in form—which we used to create distractor

contexts. In this respect, the target words that were used in our study are both novel and disturbingly familiar. For this reason, learning these words may require additional effort in order to suppress the more familiar (distractor) lexeme (see Vitevitch, 1997, for evidence that malapropisms often involve the use of low-frequency words that are slightly more frequent—i.e., more strongly represented in memory—than the intended word).

In addition, the words that were used in this study were appropriate to the task constraints: We asked participants to limit their response to a simple one- or two-word synonym or definition. This procedure may work well for words that have close paraphrases, but possibly less so for words that encode more novel concepts (see Nagy et al., 1987). This issue can be addressed in future experiments. In addition, with future modifications of the MESA algorithm, it may be less critical to impose artificial constraints on learner responses, such as limiting responses to a single word.

Ongoing and future research. One shortcoming of the present experiment is the absence of a pretraining baseline for the MESA scores. It is possible that some effects—particularly those that are due to context quality—would have been stronger if we had acquired definition-accuracy measures before the first learning trial. Indeed, Figure 6 shows that there was an effect of context quality on the very first trial. This result is consistent with data from studies of retrieval practice, where the first trial has been shown to have a disproportionate effect on learning outcomes (Roediger & Karpicke, 2006). In a follow-up study, we addressed this issue by including a pretraining block, in which participants were asked to generate meanings for words that are presented in isolation (Frishkoff & Perfetti, in press).

It is also important to note that we have not accounted for order effects in the present analyses. It seems likely that the presence of an error on the first trial has a greater impact on learning outcomes than an error in the middle or near the end of the practice session. We are examining this question in a follow-up study with additional analyses. Another factor that may be important to consider is the degree of form-based (orthographic and/or phonological) similarity between target and distractor words. Words that are more similar are also more likely to be confused, particularly if both the target and the distractor word are unfamiliar (Vitevitch, 1997). Similarly, the semantic "distance" or similarity between target and distractor words is likely to mediate learning of target words.

Another important area for research concerns the role of prior knowledge and skills, and of online "strategies," in word learning from context. As mentioned earlier, prior work has described different word-learning outcomes for readers with high reading comprehension and vocabulary skills versus those with low reading comprehension and vocabulary skills. These effects are not consistently found across studies with comparable paradigms, however (for reviews and meta-analyses, see Stahl & Fairbanks, 1986; Swanborn & de Glopper, 2002). Additional, systematic studies are needed to reconcile this body of results and to link these patterns to an explicit model of word learning that accounts for the roles of conceptual and linguistic knowledge and participant-specific strategies. These studies may benefit from the use of new computational measures (Graesser et al., 2004) and from technologies such as ERP (Frishkoff & Perfetti, in press) and eyetracking (Reichle & Perfetti, 2003), which reveal dimensions of behavior and cognitive processing that may not emerge from behavioral methods alone.

Conclusion

In conclusion, we have demonstrated that the MESA algorithm can be used to capture incremental changes in word knowledge. Using this measure, we found that word-learning trajectories were sensitive to context quality, the use of spaced versus massed practice, and skill level. In general, this method may be a useful research tool, opening up new lines of inquiry and new experimental paradigms for research on word learning.

AUTHOR NOTE

This research was supported by an American Psychological Association/Institute of Education Sciences (IES) Postdoctoral Education Research Training fellowship under the Department of Education, Grant R305U030004 to G.A.F. and C.A.P., and by an IES Postdoctoral Research Fellowship, Grant R305B050022 to C.A.P. We thank Erika Taylor for assistance with data archiving and preprocessing. Correspondence concerning this article should be addressed to G. A. Frishkoff, LRDC Rm. 642, 3939 O'Hara Street, University of Pittsburgh, Pittsburgh, PA 15260 (e-mail: gwenf@pitt.edu).

REFERENCES

BAUMANN, J. F., KAME'ENUI, E. J., & ASH, G. E. (1998). Research on vocabulary instruction: Voltaire redux. In D. C. Simmons & E. J. Kame'enui (Eds.), *What reading research tells us about children with diverse learning needs* (pp. 183-218). Mahwah, NJ: Erlbaum.

BIEMILLER, A. (2004). Teaching vocabulary in the primary grades: Vocabulary instruction needed. In J. F. Baumann & E. J. Kame'enui (Eds.), *Vocabulary instruction: Research to practice* (pp. 28-40). New York: Guilford.

BINDER, J. R., MCKIERNAN, K. A., PARSONS, M. E., WESTBURY, C. F., POSSING, E. T., KAUFMAN, J. N., & BUCHANAN, L. (2003). Neural correlates of lexical access during visual word recognition. *Journal of Cognitive Neuroscience*, **15**, 372-393.

BJORK, R. A., & LINN, M. C. (2006, March). The science of learning and the learning of science: Introducing desirable difficulties. *APS Observer*, **19**. Retrieved March 1, 2008, from www.psychologicalscience.org/observer/getArticle.cfm?id=1952.

BROWN, J. C., FRISHKOFF, G. A., & ESKENAZI, M. (2005). Automatic question generation for vocabulary assessment. In *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing* (pp. 819-826). Vancouver, BC: Association for Computational Linguistics.

BROWN, J. I., NELSON, M. J., & DENNY, E. C. (1973). *The Nelson–Denny reading test: Forms C and D for high schools and colleges.* Boston: Houghton Mifflin.

CAIN, K., OAKHILL, J., & 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. (2004). Information retrieval for language tutoring: An overview of the REAP project. In *Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 544-545). Sheffield, U.K.: Association for Computing Machinery.

COLLINS-THOMPSON, K., & CALLAN, J. (2007). Automatic and human scoring of word definition responses. In *Proceedings of the NAACL-*

HLT 2007 Conference (pp. 476-483). Rochester, NY: Association for Computational Linguistics.

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

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

FEINGOLD, A. (1983). The validity of the Information and Vocabulary subtests of the WAIS for predicting college achievement. *Educational & Psychological Measurement*, **43**, 1127-1131.

FRISHKOFF, G. A., & PERFETTI, C. A. (in press). Lexical quality in the brain: ERP evidence for robust word learning from context. *Developmental Neuropsychology*.

FRISHKOFF, G. A., PERFETTI, C. A., & WESTBURY, C. (in press). ERP measures of partial semantic knowledge: Left temporal indices of skill differences and lexical quality. *Biological Psychology*.

FRISHKOFF, G. A., WHITE, G., & PERFETTI, C. A. (in press). In vivo testing of learning and instructional principles: The design and implementation of school-based experimentation. In L. Dinella (Ed.), *Conducting high-quality psychological research in school-based settings: A practical guide*. Washington, DC: American Psychological Association.

GRAESSER, A. C., MCNAMARA, D. S., LOUWERSE, M. M., & CAI, Z. (2004). Coh-Metrix: Analysis of text on cohesion and language. *Behavior Research Methods, Instruments, & Computers*, **36**, 193-202.

GRAHAM, D. M., & WATTS, S. M. (1990, November). *Contextual analysis in naturally occurring prose: Effects of passage length, word frequency, and grade*. Paper presented at the annual meeting of the Northeastern Educational Research Association, Ellenville, NY.

INCE, E., & CHRISTMAN, S. D. (2002). Semantic representations of word meanings by the cerebral hemispheres. *Brain & Language*, **80**, 393-420.

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, & Cognition*, **33**, 704-719.

KOEDINGER, K., & ALEVEN, V. (2007). Exploring the assistance dilemma in experiments with cognitive tutors. *Educational Psychology Review*, **19**, 239-264.

LANDAUER, T. K., FOLTZ, P. W., & LAHAM, D. (1998). An introduction to latent semantic analysis. *Discourse Processes*, **25**, 259-284.

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

MCKEOWN, M. G., BECK, I. L., OMANSON, R. C., & POPLE, M. T. (1985). Some effects of the nature and frequency of vocabulary instruction on the knowledge and use of words. *Reading Research Quarterly*, **20**, 522-535.

MCNAMARA, D. S. (2001). Reading both high-coherence and low-coherence texts: Effects of text sequence and prior knowledge. *Canadian Journal of Experimental Psychology*, **55**, 51-62.

MILLER, G. A. (1995). WordNet: A lexical database for English. *Communications of the ACM*, **38**(11), 39-41.

NAGY, W. E., ANDERSON, R. C., & HERMAN, P. A. (1987). Learning word meanings from context during normal reading. *American Educational Research Journal*, **24**, 237-270.

NATIONAL READING PANEL (NRP) (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.

OSGOOD, C. E., SUCI, G. J., & TANNENBAUM, P. H. (1957). *The measurement of meaning*. Urbana: University of Illinois Press.

PAVLIK, P. I., JR., & 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., & HART, L. (2001). The lexical basis of comprehension skill. In D. S. Gorfein (Ed.), *On the consequences of meaning selection: Perspectives on resolving lexical ambiguity* (pp. 67-86). Washington, DC: American Psychological Association.

PERFETTI, C. A., & HART, L. (2002). The lexical quality hypothesis. In L. Verhoeven, C. Elbro, & P. Reitsma (Eds.), *Precursors of functional literacy* (pp. 189-213). Amsterdam: John Benjamins.

RAPAPORT, W. J. (2003). What is the "context" for contextual vocabulary acquisition? In P. P. Slezak (Ed.), *Proceedings of the 4th Joint International Conference on Cognitive Science/7th Australasian Society for Cognitive Science Conference* (Vol. 2, pp. 547-552). Sydney: University of New South Wales.

RAPAPORT, W. J. (2005). In defense of contextual vocabulary acquisition: How to do things with words in context. In A. Dey et al. (Eds.), *Proceedings of the 5th International and Interdisciplinary Conference on Modeling and Using Context (Context-05)* (pp. 396-409). Berlin: Springer.

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**, 219-237.

ROEDIGER, H. L., III, & 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. (2003). Words are learned incrementally over multiple exposures. *American Educator*, **27**, 18-19.

STAHL, S. A., & FAIRBANKS, M. M. (1986). The effects of vocabulary instruction: A model-based meta-analysis. *Review of Educational 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. (2002). Impact of reading purpose on incidental word learning from context. *Language Learning*, **52**, 95-117.

TOUTANOVA, K., MANNING, C. D., & NG, A. Y. (2004). Learning random walk models for inducing word dependency distributions. In *Proceedings of the Twenty-First International Conference on Machine Learning* (p. 103). New York: Association for Computing Machinery.

VAN DAALEN-KAPTEIJNS, M. M., & ELSHOUT-MOHR, M. (1981). The acquisition of word meanings as a cognitive learning process. *Journal of Verbal Learning & Verbal Behavior*, **20**, 386-399.

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.

VITEVITCH, M. S. (1997). The neighborhood characteristics of malapropisms. *Language & Speech*, **40**, 211-228.

WALBERG, H. J., & TSAI, S.-L. (1983). Matthew effects in education. *American Educational Research Journal*, **20**, 359-373.

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

APPENDIX A
Target–Distractor Word Pairs

Target	Distractor	Target	Distractor
fractious	factious	meretricious	meritorious
baneful	baleful	auspicious	suspicious
abrogate	arrogate	horrent	horrific
belie	betray	zealous	jealous
reticent	reluctant	bibulous	bilious
tortuous	torturous	calvous	callous
reboant	redolant	osculate	oscillate
irenic	ironic	imminent	eminent
turbid	turgid	elude	allude
callow	shallow	volable	voluble
choleric	caloric	ingenuous	ingenious
ablution	absolution	discinct	distinct
ameliorate	alleviate	elusory	illusory
impertinent	irrelevant	vociferous	voracious
corroborate	collaborate	perquisite	prerequisite
acidulous	assiduous	maunder	meander
invidious	insidious	immure	inure
sapid	vapid	gambit	gamut
odious	odorous	noisome	noisy
incisive	decisive	antidote	anecdote
inchoate	incoherent	affinity	infinity
derogate	delegate	circumspect	circumscribed
apposite	opposite	ulterior	anterior
ascetic	aesthetic	laudable	laughable
mitigate	militate	indelible	infallable
flagrant	fragrant	abditive	additive
enervate	energize	incondite	recondite
venal	venial	conticent	complacent
salubrious	salacious	censure	censor
impacable	impeccable	veracious	voracious

APPENDIX B
Lexical-Knowledge Battery

1. *Pseudohomophone (Orthographic Knowledge) Test.* Participants are asked to judge which letter strings, when pronounced aloud, sound like real English words. This test is scored according to the number of items that are selected correctly (e.g., "hits") and incorrectly (e.g., "false alarms"). This test provides a measure of participants' decoding ability.

2. *Spelling Test.* This is a spelling discrimination task in which participants are presented with one correct and four incorrect spellings of irregular, easily misspelled words (e.g., *nuisance*, *nuisence*, *newsance*, *newcense*, *newsince*). Items are from the Baroff spelling test, with additional items added for increased difficulty.

3. *Nelson–Denny Comprehension Test* (J. I. Brown et al., 1973). This is a text-comprehension test, in which eight paragraphs are followed by 4 to 5 questions, for a total of 36 questions. Participants were given 15 min to complete the test, instead of the usual 20 min. Both speed (number of items attempted) and accuracy (on only those items attempted) were recorded.

4. *The Nelson–Denny Vocabulary Test* (J. I. Brown et al., 1973). Twenty items were chosen, spanning the range of difficulty of the task. Participants were given 2 min to circle the correct definitions of these 20 words, in a multiple-choice format. Both speed (number of items attempted) and accuracy (on only those items attempted) were recorded.

5. *The Raven's Progressive Matrices* is a measure of nonverbal, adaptive intelligence. Participants are presented with 12 test items; each test item is a 3×3 array of nine patterns, with the ninth pattern omitted. Participants are asked to correctly identify the missing pattern that is required to complete the array from six potential choices. The items on the test progressively increase in difficulty; thus, each item requires greater cognitive processing than does the previous item. The test is scored according to how many items are answered correctly.

6. *Phonological Awareness Task*. The Phonological Awareness Task (PhAT) asks a participant to remove a sound within a word and then replace it with another sound. For example, if one removes the /p/ from "speak," the result is "seek." Add /l/, and the result is "sleek." The scores are analyzed on the basis of spelling of the word, correct intention with the word, and phonological removal or correction. A perfect score is given for a word only when it is spelled correctly and is the intended word. Partial credit is given for misspelling of the intended word, or if a real word is presented.

(Manuscript received October 7, 2007;
revision accepted for publication March 12, 2008.)