Listening. Learning. Leading.®

Research Report
ETS RR–14-02

Creating Vocabulary Item Types That Measure Students' Depth of Semantic Knowledge

Paul Deane

René R. Lawless

Chen Li

John Sabatini

Isaac I. Bejar

Tenaha O'Reilly

June 2014

Discover this journal online at
Wiley Online Library
wileyonlinelibrary.com

ETS Research Report Series

EIGNOR EXECUTIVE EDITOR

James Carlson
Principal Psychometrician

ASSOCIATE EDITORS

Beata Beigman Klebanov
Research Scientist

Heather Buzick
Research Scientist

Brent Bridgeman
Distinguished Presidential Appointee

Keelan Evanini
Managing Research Scientist

Marna Golub-Smith
Principal Psychometrician

Shelby Haberman
Distinguished Presidential Appointee

Gary Ockey
Research Scientist

Donald Powers
Managing Principal Research Scientist

Gautam Puhan
Senior Psychometrician

John Sabatini
Managing Principal Research Scientist

Matthias von Davier
Director, Research

Rebecca Zwick
Distinguished Presidential Appointee

PRODUCTION EDITORS

Kim Fryer
Manager, Editing Services

Ayleen Stellhorn
Editor

Since its 1947 founding, ETS has conducted and disseminated scientific research to support its products and services, and to advance the measurement and education fields. In keeping with these goals, ETS is committed to making its research freely available to the professional community and to the general public. Published accounts of ETS research, including papers in the ETS Research Report series, undergo a formal peer-review process by ETS staff to ensure that they meet established scientific and professional standards. All such ETS-conducted peer reviews are in addition to any reviews that outside organizations may provide as part of their own publication processes. Peer review notwithstanding, the positions expressed in the ETS Research Report series and other published accounts of ETS research are those of the authors and not necessarily those of the Officers and Trustees of Educational Testing Service.

The Daniel Eignor Editorship is named in honor of Dr. Daniel R. Eignor, who from 2001 until 2011 served the Research and Development division as Editor for the ETS Research Report series. The Eignor Editorship has been created to recognize the pivotal leadership role that Dr. Eignor played in the research publication process at ETS.

ETS Research Report Series ISSN 2330-8516

RESEARCH REPORT

Creating Vocabulary Item Types That Measure Students' Depth of Semantic Knowledge

Paul Deane, René R. Lawless, Chen Li, John Sabatini, Isaac I. Bejar, & Tenaha O'Reilly

Educational Testing Service, Princeton, NJ

We expect that word knowledge accumulates gradually. This article draws on earlier approaches to assessing depth, but focuses on one dimension: richness of semantic knowledge. We present results from a study in which three distinct item types were developed at three levels of depth: knowledge of common usage patterns, knowledge of broad topical associations, and knowledge of specific conceptual relationships. We attempted to avoid common sources of variance across items (such as attractive distracters) and hypothesized that the item types that required greater depth of semantic knowledge would tend to show greater difficulty and discrimination after other sources of variance were accounted for. Our results, while still exploratory, support the conclusion that the item types measure different aspects of lexical knowledge, consistent with the hypothesis of increasing semantic depth.

Keywords Word knowledge; common usage patterns; topical associations; conceptual relationships

doi:10.1002/ets2.12001

Overview

The Need for More Sophisticated Measures of Vocabulary

Vocabulary is well recognized as an essential component of reading proficiency (Beck & McKeown, 1991; Carroll, 1993; Cunningham & Stanovich, 1997; Daneman, 1988; Hirsch, 2003; Perfetti, 1994) with correlations between vocabulary and reading comprehension assessments ranging from .6 to .7 (Anderson & Freebody, 1981). While the importance of vocabulary development is apparent to researchers and practitioners, the state of the art in vocabulary assessment tends to have a strong summative or clinical focus: Most reading vocabulary tests consist of a small sampling of words that vary in familiarity and a task that requires choice of a synonym or definition (e.g., Sheehan, Kostin, & Persky, 2006), and most clinical receptive/productive tests require the examinee to respond to a picture prompt with a verbal label (e.g., Peabody Picture Vocabulary [Dunn & Dunn, 1997]), or vice versa (e.g., Boston Naming Test [Kaplan, Goodglass, & Weintraub, 1978]). Both classes of vocabulary tests are broad measures that typically can be administered once or twice a year to estimate overall vocabulary growth. They tend not to be designed for many classroom-based, formative purposes. The cognitive and psycholinguistic literature, however, supports a richer set of vocabulary construct distinctions (e.g., estimating aspects of breadth, depth, and word learning skills [i.e., Chabot, Petros, & McCord, 1983; Daneman & Green, 1986; Dixon, LeFevre, & Twilley, 1988; Durso & Shore, 1991; Hogaboam & Perfetti, 1975; Hu & Nation, 2000; Sahhouse, 1993; Stanovich, West, & Harrison, 1995; Swanborn & De Glopper, 1999; Walczyk & Raska, 1992]), but these are rarely incorporated into existing standard instruments in a systematic way. For the purposes of the present study, we focus our attention on the *depth of knowledge about specific word meanings* as one of the most important (and clearly delimitable) aspects of vocabulary knowledge. Our goal is to develop a more sophisticated understanding of how to conceptualize and assess this particular aspect of word knowledge.

Word Learning and Depth of Vocabulary Knowledge

As word learning proceeds, meanings of words grow richer over time. Perfetti and Hart (2001) described word knowledge as a complex assemblage of representations that vary both in the information they contain and in the degree to which they have been fully specified (i.e., in terms of orthographic, phonemic, syntactic, and semantic quality), which Perfetti and

Corresponding author: P. Deane, E-mail: pdeane@ets.org

Hart referred to as the *lexical quality hypothesis*. Consistent with the lexical quality hypothesis, we expect that the normal course of development is one in which the meaning of a word is initially totally unknown and then gradually becomes more fully specified with continued experience.

Aligned with the topic of word learning are processes of partial word learning, driven in part by the effects of exposure to a variety of written texts. A number of theorists have outlined stages of word meaning that postulate differing degrees of depth of word knowledge acquired piecemeal. For example, Dale and O'Rourke (1986) postulated four stages of word learning: Stage I, the word is completely unknown; Stage II, implicit word knowledge; Stage III, partial knowledge but mastery in some contexts; and Stage IV, full mastery across a range of uses. Stahl (1986) outlined a similar (three-stage) theory, which is applied in Brown, Frishkoff, and Eskenazi (2005) to the task of automatically generating questions designed to probe different aspects of vocabulary depth. Brown et al. (2005) primarily used WordNet semantic relationships to generate definition, synonym, antonym, hypernym, hyponym, and cloze questions. In their discussion, they characterized these tasks as primarily providing evidence for the middle level of Stahl's (1986) hierarchy.

Designing Items to Measure Depth of Vocabulary Knowledge

If we focus on one aspect of the lexical quality hypothesis, we can identify a number of (somewhat separable) aspects of the representation of word meaning that correspond reasonably well to the kinds of inferences theorists have proposed for incremental word learning. The following list suggests the kinds of progressions we might see, if depth of semantic knowledge corresponds to the development of an increasingly rich and interconnected representation, driven by associative and inferential processes (cf., Beck, McKeown, & McCaslin, 1983; Carroll & White, 1973; Fukkink, Henk, & De Glopper, 2001; Graves, 1986; Nagy, Anderson, & Herman, 1987; Nagy & Scott, 2000; Schatz & Baldwin, 1986; Schwanenflugel, Stahl, & McFalls, 1997). That is, if we start with the assumption that the normal path to semantic knowledge starts with exposure to usage, involves inferential processes, and results in the gradual consolidation of a semantic/conceptual representation integrated with background knowledge, then we might be able to measure this process at the following levels:

1. *Familiarization with patterns of usage*. Experience with words, whether orally or in print, necessarily corresponds to some degree of familiarization with the contexts in which the word appears, and thus to characteristic patterns of usage. Psychologically this corresponds to the development of perceptual traces, whether or not the student has developed more truly semantic associations (and thus attention to some of the kinds of associations that form the focus of such models as Bates & MacWhinney, 1989).
2. *Development of appropriate semantic memory representations.* A somewhat richer form of semantic knowledge comes into play when we consider priming effects and other aspects of fast lexical access involving semantic memory. These kinds of processes embody some degree of generalization about words, in which similar words tend to be accessed together, without necessarily entailing full conceptual understanding (McKoon & Ratcliffe, 1998; Myers & O'Brien, 1998).
3. *Development of appropriate conceptual representations.* An even richer form of semantic knowledge involves what we might term definitional knowledge (i.e., being able to map from purely verbal to conceptual representations). Such knowledge is critical for various forms of reasoning, such as identifying broader or narrower categories (hypernyms and hyponyms) and drawing logical inferences (as in Fellbaum, 2010).
4. *Consolidation of conceptual representations with world knowledge.* Finally, a truly deep and complete understanding of what words mean may require reasoning that integrates purely semantic knowledge with encyclopedic understandings of the subject matter being addressed, involving the kind of processes that reach beyond purely linguistic representations discussed in Elman (2009).

Ideally, we would want a vocabulary assessment to be able to support the kinds of qualitative judgments about the richness of vocabulary judgment that people are able to make.

There is not, necessarily, a guarantee that all people will develop semantic knowledge in the sequential order noted in the list above. That order is most plausible for vocabulary acquired implicitly, either as part of daily experience or through reading. Any vocabulary that is acquired explicitly — perhaps by memorizing dictionary definitions — might follow a different sequence. But it seems reasonable to assume — at least as a first approximation, all other things being equal — that we could expect people with implicit vocabulary knowledge would have less trouble making judgments about Item 1 in the list above than about Item 2, about Item 2 than about Item 3, and about Item 3 than about Item 4. This is a testable hypothesis,

although one that involves serious methodological issues, since the discrimination and difficulty of a vocabulary item can be driven by a wide range of factors (e.g., distractors employed).

Essentially, we are proposing the development of vocabulary items designed specifically to tap different aspects of depth of vocabulary knowledge. The approach we are following is analogous to the kind of trajectory described in Embretson (1998) and Embretson and Gorin (2001), in which the design is driven by cognitive psychological principles and in which features of the item design are chosen precisely to minimize the effects of construct-irrelevant factors. Prior studies have shown that various features tend in general to affect the difficulty of vocabulary items, including word frequency, abstractness and imageability of word meanings, and age of acquisition (Bird, Howard, & Franklin, 2003; Breland & Jenkins, 1997; Breland, Jones, & Jenkins, 1994; Carroll, 1970, 1971, 1976, 1993; Carroll & White, 1973; Coltheart, Laxon, & Keating, 1988; Gernsbacher, 1984; McFalls, Schwanenflugel, & Stahl, 1996; Paivio, 1971; Zevin & Seidenberg, 2002). A variety of factors beyond vocabulary have been implicated specifically in the difficulty and discrimination of vocabulary and reading items (Freedle & Kostin, 1992; Gao & Rogers, 2007; Gorin, 2005; Sheehan & Ginther, 2001; Sheehan, Ginther, & Schedl, 1999; Sheehan & Mislevy, 1990, 2001; Sheehan et al., 2006). This fact suggests the possibility of designing items—and controlling their discrimination and difficulty—by careful consideration of the aspects of knowledge about which information is to be obtained and by equally careful manipulation of variables that will have appropriate effects on the reasoning processes of people completing the resulting items. These kinds of considerations can have a significant impact on the validity of items (Rupp, Ferne, & Choi, 2006). But if successful, an item design that is carefully controlled may be able to yield valid evidence about different aspects—in this case—of the construct of vocabulary knowledge, and therefore enable the development of vocabulary measures that are much more sensitive to the presence of partial vocabulary knowledge. Scott et al. (2008) develop a similar research program, though they use control of distractor choices, rather than development of distinct item types, to measure different aspects of depth of vocabulary knowledge.

Thus, the current study can be placed in a larger context in which a cognitively motivated assessment design is proposed and then validated by examining whether items built to that design can be modeled using appropriate and cognitively motivated features. In this particular case, we expect that if we develop items specifically intended to measure a particular level of depth of vocabulary knowledge, and control for other factors, we will observe patterns of difficulty and discrimination that correspond to the level of depth of vocabulary knowledge that each item type requires.

The study reported here is intended primarily as a proof of concept, in which we went through each step of a cognitively motivated design and validation process: creating items to measure different levels of depth of vocabulary, observing item parameters in a field study, and then modeling those parameters in a cognitive model that uses observable features of the item (including word frequency, but going well beyond this to include other empirical information about the words used in each item) to predict item parameters.

Instrument Development

Three types of items were developed for and tested in this study—each with specific goals in mind to test different types, and hopefully depth, of vocabulary knowledge. These types included an idiomatic associates item type, intended to measure familiarity with patterns of usage; a topical associates item type, intended to measure the kinds of associations represented in semantic memory; and the hypernym item type, designed to measure access to conceptual representations and associated patterns of inference.[1] All three item types take the form of three-option multiple-choice questions containing one correct answer (key). The design of each item type was intended to avoid unnecessary sources of difficulty (such as attractive distractors) as much as possible, so that one could reasonably argue that success in answering the question demonstrated a control of the relevant kind of lexical knowledge about the targeted word. The format used and the specific words selected as part of each design were limited to meet specific design targets such as the overall frequency and probability of co-occurrence with the targeted word in a large corpus of edited English texts.

The Idiomatic Associates Item Type

The idiomatic associates item type is designed to test students' knowledge of the typical phrasal patterns characteristic of the targeted words. Because it is intended to test this level of knowledge, and no more, the design was motivated by the need to ensure that someone could answer the item correctly based only on implicit knowledge of common usage (e.g., co-occurrence). Figure 1 illustrates this item type.

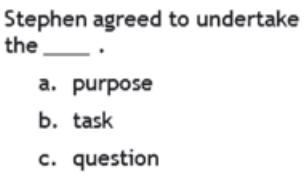

Figure 1 A sample idiomatic associates item (word being tested: undertake).

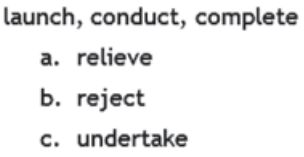

Figure 2 A sample topical associates item (word being tested: undertake).

The development of this item type was governed by the following design decisions:

- The prompt takes the form of a cloze sentence-completion, multiple-choice item. The word being tested is not one of the options. Instead it appears in the stem, just before or just after the blank, in order to make it possible to contrast judgments about the three different contexts expressed in the options. As much as possible, nothing in the stem *other than the targeted word* cues the correct answer.
- The key is a natural, idiomatic, and relatively frequent collocate of the targeted word in the context of the sentence presented in the prompt.
- There should be as little difference between the key and the other options as possible, except their plausibility in the context supplied. Thus, the key and the options should belong to the same part of speech and be approximately equal in frequency.
- More specifically, one of the distractors should be so unusual a usage such that it is ungrammatical, distinctly odd-sounding, or awkward. The other distractor should be plausible in context (if only meaning is taken into account) but should only occur rarely in that context.

We were able to enforce these design constraints by drawing upon corpus data about word frequency, co-occurrence, and patterns of co-occurrence, using the Lexile/SourceFinder corpus, a large (462 million word) corpus of edited English texts and word frequency information from the Touchstone Applied Science Associates (TASA) corpus (Zeno, Ivens, Millard, & Duvvuri, 1995). As long as these constraints are followed, our expectation is that anyone who has heard the targeted word frequently enough could recognize the key by recall from perceptual memory alone, without needing to access semantic information.

The Topical Associates Item Type

The topical associates item type is designed to test students' knowledge of the kinds of associations that reflect fast lexical access processes and thus support semantic priming. As a result, the major design constraint is the need to provide just enough information to strongly and unambiguously activate a single topic or concept. Once activated, the association between that concept and the targeted word should be obvious to anyone who has an appropriate representation of the targeted word in semantic memory. Figure 2 illustrates this item type.

In order to make sure that the intended association is clear, three stimulus words are presented. The three words are supposed to be strongly associated with one another and with the targeted word—but not with the other two options. Thus, if the intended associations are available in semantic memory, the student will be able to identify the correct answer, even if he or she does not have any deeper conceptual understanding of the target.

The development of this item type was governed by the following design decisions:

- The three stimulus words must belong to the same part of speech.
- The three options must belong to the same part of speech (but may be a different part of speech than the stimulus words).

To undertake something is to ____ it.

a. begin
b. continue
c. notice

Figure 3 Sample hypernym item (Word tested: undertake).

- There must be a relatively strong association between the stimulus words and the targeted word as measured by a mutual information statistic.
- The mutual information between the stimulus words and the other options must be much lower.
- The stimulus words must not be synonyms or hypernyms of the targeted word.
- The stimulus words must not be strong collocates of the targeted word.
- The stimulus words should not be less frequent than the targeted word.
- The distractors should be at least as frequent as the targeted word.

Once again we were able to enforce these design constraints by drawing upon the TASA word frequencies and Lexile/SourceFinder corpus data.

Most of these constraints were intended to maximize the likelihood that the three stimulus words would prime the key but would not prime either of the other two options. A few of them were designed to rule out alternative paths to a correct answer. Our expectation is that as long as these constraints are upheld, anyone who can answer the questions correctly will have demonstrated that they have appropriate representations of the stimulus and targeted words in semantic memory.

The Hypernym Item Type

The hypernym item type is designed to test whether students have sufficient access to conceptual representations associated with a word to be able to make basic definitional inferences, primarily (but not exclusively) to what are sometimes referred to as hypernym relations (Fellbaum, 2010). In the case of nouns, this involves the ability to recognize the broad meaning or category to which the targeted words belong. In other cases (e.g., verbs), the item type as we defined it may involve the most prominent causal inferences or (in the case of adverbs) the broad category to which the nominal form of the word belongs. Figure 3 illustrates this item type.

The development of this item type was governed by the following design decisions:

- Like the idiomatic associates, instances of this item type take the form of cloze sentence-completion items.
- The targeted word is contained in the stem, in order to contrast judgments about the three possible hypernym relationships expressed in the options. Nothing in the stem *other than the targeted word* should provide a cue to the correct answer.
- The options should contain words that plausibly could fit in the blank, belong to the same part of speech, are approximately the same frequency as the targeted word, and are more or less at the same broad level of abstraction as the key.
- The key makes the sentence a true statement that partially defines the targeted word.
- The two distractors, when placed in the blank in the stem, should produce sentences that are not true as definitions, even if they might be contingently true in some situations.

We were able to enforce these constraints in part by setting thresholds on a number of natural language processing (NLP)-derived features: TASA and Lexile/SourceFinder word frequencies and WordFit similarities (Deane, 2003, 2005; Deane & Higgins, 2006) that indicate whether words tend to appear characteristically in the same *n*-gram contexts.

These constraints were intended to eliminate possible routes to a correct answer other than access to a definitional understanding of the targeted word. As the item is designed, it is not necessary to have a deep conceptual understanding of the word's meaning; all that is required is sufficient knowledge and understanding to make the correct inferences. Thus nothing in this item entails that someone who gets a hypernym item correct will be able to provide exact definitions of the word or produce the word only in semantically appropriate contexts.

Selection of Vocabulary

The main study required the selection of vocabulary targeted for instruction in the middle grades, so that many students would be likely to have partial vocabulary knowledge for these words and relatively few would have achieved the maximal level of depth (i.e., the full consolidation of lexical knowledge and its integration with appropriate, associated knowledge of the world). We therefore developed items for 50 general academic words selected from the vocabulary taught in middle school as part of the Word Generation vocabulary intervention (Snow, Lawrence, & White, 2009). We developed idiomatic associates, topical associates, and hypernym items for each type and a semantic associates item type intended to test the maximal depth of vocabulary knowledge.

These item types were field-tested in June 2009 by inclusion in the posttest for a study administered by the Word Generation research group. This test was administered to 2,825 students in 14 middle schools in an urban, New England school district. Test forms were assembled, each containing an anchor test of 50 Word Generation multiple-choice synonym items followed by one of 20 test forms comprised of 10–12 homogenous groups of the newly developed items that were randomly distributed among the students. Mean scores were calculated for each item type and by individual items to examine the estimated difficulty for this population. An item analysis was conducted and specific statistics were examined for inclusion/exclusion on the current study: the TASA standard frequency index (SFI),[2] item proportion correct (P+), point-biserial correlation (r_{pb}), and option-choice frequencies. These results enabled problematic items to be identified and revised prior to the main study and led to the rejection of the semantic associates item type as too difficult and unreliable for inclusion in this study. The use of the Word Generation items for the field test had the advantage that the words had been targeted for instruction, increasing the probability that students would have at least a partial knowledge of the words, which diminished the risk that a word might turn out almost universally unknown for the student population. Since we were using this data collection primarily to identify items in need of revision, the advantages of this venue outweighed the potential limitations.

For the main study, students were involved who had not been specifically targeted to learn these words. It was necessary to reduce the total number of words from 50 to 20, and we desired to do so in a way that would maintain a balanced set of items after word frequency and difficulty was taken into account. Words were therefore first sequenced in ascending order using the P+ values on the SERP Word Generation synonym items from the pilot study. They were then clustered into five ranges of difficulties: hard (less than .40), high-medium (.40–.49), medium (.50–.59), low medium (.60–.79), and easy (greater than .80). Next, the TASA SFIs[3] were looked up for each word. These word frequencies were computed from an analysis of more than 60,000 samples of printed text encountered by students in American schools, which is a good metric to use to ensure that the vocabulary clusters (based on P+) chosen for this study were approximately equivalent in terms of approximate difficulty and, based on an external measure, not influenced by the variability that can occur in student performances. To narrow down the list to a total of 20 words to be tested with the new item types, the point-biserial correlations were examined across item types for values of $r_{pb} < .15$, and option-choice frequencies were examined with the actual items to look for double-keys or distractors that may have been overly attractive. Preference was given to words for which all item types had acceptable point-biserial correlations and no evidence of double-keying. The final selection of words was made to sample across a wide range both of P+ values and word frequencies.

Study Design

The Experimental Instrument

The goal of the study was to explore whether the use of the three different types of multiple-choice items could reveal distinguishable levels of depth of vocabulary understanding from middle school students. A within-subjects design was incorporated in order to measure students' performances on all three item types for the same set of academic vocabulary words. Items were created for two sets of 10 vocabulary words to which students were exposed to one set of 10 or the other; hence, each student was exposed to each word within his or her assigned set three times in the context of the three item types. Because each item type required a different type of thinking, the tests were assembled such that students would be exposed to one item type at a time in an effort to reduce their cognitive load (i.e., the 10 homogenous items for each type were presented together). To detect any learning effects that may have been caused by multiple exposures of the 10 vocabulary words, the three sets of item types were assembled in six different sequences, as exemplified in Table 1.

Table 1 Test Assembly and Test Form Assignment of the Experimental Items in the Within-Subjects Study

Word set 1 form number	Word set 2 form number	Test Section 1	Test Section 2	Test Section 3
001	007	Item Type B	Item Type A	Item Type C
002	008	Item Type B	Item Type C	Item Type A
003	009	Item Type A	Item Type B	Item Type C
004	010	Item Type A	Item Type C	Item Type B
005	011	Item Type C	Item Type A	Item Type B
006	012	Item Type C	Item Type B	Item Type A

Note: Item type key: A = topical associations; B = idiomatic associates; C = hypernym.

Table 2 Tested Vocabulary and Item Sequencing for the Three Different Item Types

		Position in set		
Vocabulary word	Word set	Item type A	Item type B	Item type C
Adequate	1	10	1	2
Circumstances	1	5	9	9
Concept	1	3	2	3
Distribution	1	6	8	1
Eliminated	1	7	6	4
Explicit	1	1	5	7
Intrinsic	1	4	10	10
Invoked	1	2	7	8
Paralyzed	1	8	4	6
Undertake	1	9	3	5
Attained	2	3	5	1
Capacity	2	10	3	2
Outweigh	2	8	2	3
Generate	2	7	6	4
Compatible	2	2	8	5
Regime	2	4	10	6
Regulate	2	1	4	7
Acquired	2	6	1	8
Incentives	2	9	9	9
Enforced	2	5	7	10

Between each of the test sections, the placement of items (for the same words) was sequenced by random assignment, as can be seen in Table 2. This measure discouraged students from easily referring back to past items for information about the same word.

While we counterbalanced the test design to detect any ordering effects across the three depth item types, the most important feature of the design is the fact that each student was required to answer all three depth items for each tested word. This allowed us to place all three item types on a common scale and directly compare item characteristics across item types.

An important feature of the instrument was the inclusion of an anchor set of 20 items — to allow for the computation of a covariate for ability. The items in this section of the test were selected from the Word Generation synonym items for 20 of the words from the original set that were not tested using depth items. These items also were selected to maximize range of coverage and reliability of the individual items, using the field test data from the June 2009 study to estimate item characteristics. This section of the test was always administered as the final portion of the test.

Participants and Testing Conditions

The targeted population was comprised of students in Grades 7 and 8 from across the United States from six urban, six suburban, and eight rural schools from Alabama, Arkansas, Arizona, California, Connecticut, Georgia, Iowa, Idaho, Illinois, Indiana, Kentucky, Nevada, and Tennessee. A total of 1,449 seventh grade and 1,622 eighth grade students participated in this study. Parental consent was obtained for every student, and schools were paid $10 per completed test returned.

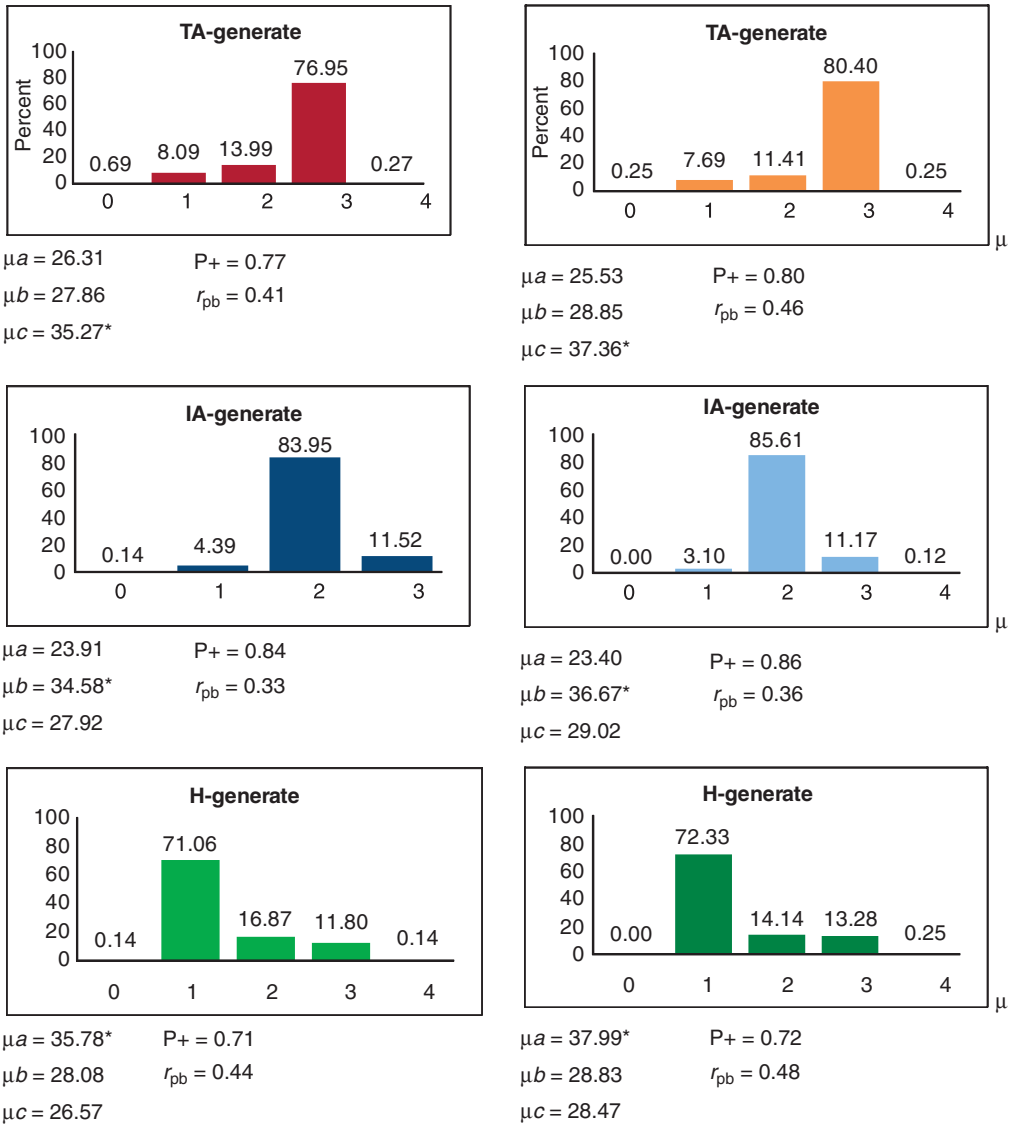

Figure 4 Option-choice frequencies for targeted word, *generate*.

Paper-and-pencil tests were administered to all students with accompanying Scantron answer sheets. Test forms were spiraled throughout each school, each grade, and each classroom to maintain the random design and ensure that all test forms were tested by approximately the same number of students. Teachers were given explicit administration instructions that included sample test questions and the relevant item instructions, which were to be reviewed with their students in advance. The teachers were also requested to accommodate any students who needed additional time to finish the test as it was not speeded, and they were reminded that their class would only earn the financial incentive for completed tests. They were also instructed to encourage students to provide their best guesses to questions they may find challenging, when the students were not absolutely certain of the answers.

Results and Initial Analysis

Initial Item Analyses

A routine item analysis was run on each item that was administered to examine each item's proportion correct (P+, point-biserial correlation (r_{pb}) and option-choice frequencies, for which histograms were generated and mean scores of students selecting each option were calculated (Figure 4).

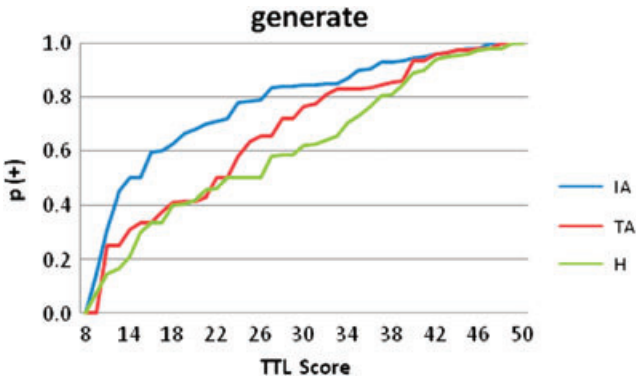

Figure 5 Sample plot of total test score by item (P+) for all students.

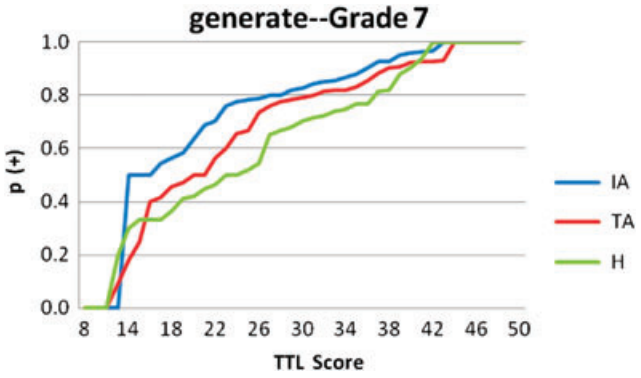

Figure 6 Sample plot of total test score by item (P+) for seventh grade students.

A close examination of the option-choice frequencies was performed to determine whether instances of attractive distractors or double-keys occurred. These items were coded as such. As a reminder, these item types were designed to assess whether or not students possessed a particular word sense within each item type. It was therefore imperative to make certain that the distractors were designed to purposely not be subtle; students either knew the word at a topical (i.e., superficial) level or they did not. This is why the distractors were designed to be obviously incorrect—in order to separate those students who have a good sense of the targeted word in the presented context from those who didn't. If 25% or more of the students selected a particular distractor, it was judged to be an attractive distractor. If the key was chosen less frequently than any given distractor, then it was judged to be a double-key.

For all students and for each grade, plots were created that superimposed line graphs of the three item types for each word of total test scores by item P+ values. As shown in Figures 5–7, plots were created to example the patterns created by the correct responses. The plots provide a visualization to roughly compare the item difficulty by item type and to allow for the identification of the middle scores where scores for students with partial knowledge could be identified for later, closer examination.

We hypothesized that for students possessing partial knowledge, the different item types may show different levels of depth of understanding. For example, an examination of the plots for the word *generate* all demonstrate a similar pattern: Hypernyms appear to be the hardest item type, followed by topical associates, with idiomatic associates appearing to be easiest. However, closer examination shows that although these three curves approximately track each other, there is a total score where the low-ability students can be separated out from the middle-ability students, and similarly, the middle-ability students from the high-ability students. It is these middle-ability students (*malleable middle*) that most interest us as we posit that the high-ability students will understand the targeted words regardless of the context of all three item types and the low-ability students may not understand or be familiar with the targeted words at all; so, while looking at the overall performance pattern for this word is helpful, it is the examination of Figures 6 and 7 that may be

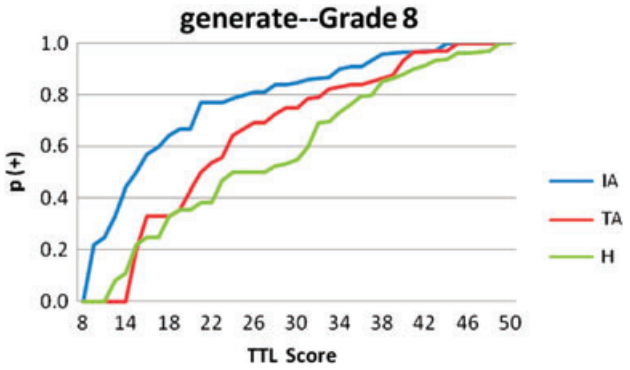

Figure 7 Sample plot of total test score by item (P+) for eighth grade students.

Table 3 Sample Table of Item Response Theory (IRT) Parameters for the Targeted Word, Generate

	Grade 7			Grade 8		
Item type	*a*	*b*	*c*	*a*	*b*	*c*
Topical associate	1.0564	−1.3864	0.3447	0.9949	−1.2915	0.2668
Hypernym	0.8097	0.1761	0.3872	0.5103	−0.0688	0.2992
Idiomatic associate	0.8999	−0.3485	0.2969	0.8359	−0.6683	0.2699

Note. a = parameter for meaningful comparison of the discrimination, b = parameter for the item difficulty, c = the guessing parameter.

revealing. In this case, an examination of Figure 6 reveals that the seventh grade students with partial knowledge of the word, *generate* seem to have total test scores between 14 and 38; and in Figure 7, the curves begin to flatten out sooner as 80 percent of the eighth grade students appear to understand the idiomatic uses of the words and topical associations at lower total score points. This finding is important for the subsequent analyses that will be explained later in this section. One might also infer from these figures that in the case of the word *generate* the hypernym item may have elicited the deepest knowledge; the topical associate, midrange knowledge; and the idiomatic associate, the most superficial. It is for this reason that a closer examination of the performance of all of the items needed to be examined so hypotheses could be drawn and tested.

Item Response Theory Analyses

Because the sample size was adequately large and because we had an anchor test administered to all students, we were able to equate the tests between the two different sets of words (as shown in Table 2) that were administered and the two different grades and run a three parameter item response theory (3PL IRT) analysis. This action allowed for a meaningful comparison of the discrimination (*a* parameter), the item difficulty (*b* parameter), and guessing (*c* parameter) in standardized scales. The procedure for equating the items was similar to that used for the National Assessment for Educational Progress (NAEP), which uses the PARSCALE IRT software (Mislevy, Johnson, & Muraki, 1992).

First, the response and scored data were cleaned and sorted by targeted word and item type. Because there were two sets of targeted words for two grades, the data were divided into four groups: Grade 7 Set 1, Grade 7 Set 2, Grade 8 Set 1, and Grade 8 Set 2. Next, an item analysis program was used to create the input files for PARSCALE (NAEP Version 3.1) for each of the four datasets. PARSCALE was used to generate the item characteristic curves (ICCs) for the item responses and input files, and the software program TBLT (NAEP Version 2.30) was used to equate all datasets to the Grade 7 Set 1 parameters. The actual equating process used was as follows: when equating Grade 7 Set 2 to Grade 7 Set 1, the 20 common items were used as anchor items; when equating Grade 8 Set 1 to Grade 7 Set 1, all of the 50 items, including common items, were used as anchor items because all of the items were the same in the two tests; and when equating Grade 8 Set 2 to Grade 7 Set 1, the 20 common items were used as anchor items. The resulting parameters were calculated (Table 3) and ICCs were generated (Figure 8).

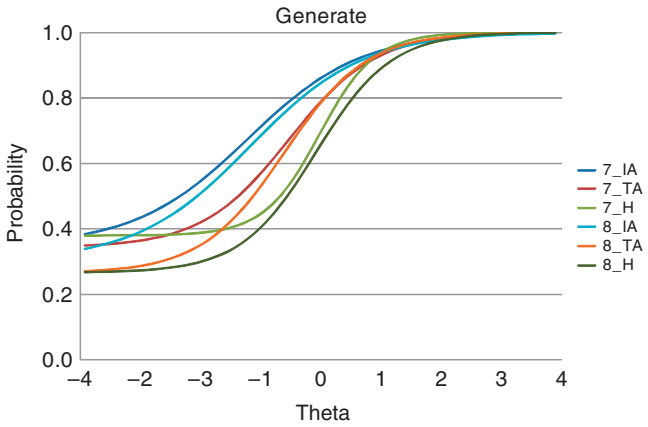

Figure 8 Sample three parameter (3PL) item characteristic curves for all item types and both grades.

Table 4 Means and Standard Deviations for Individual Responses to Each Item Type

Item type	Minimum	Maximum	Mean number correct	SD
Idiomatic associate	0	10	6.87	1.780
Topical associate	0	10	7.44	2.155
Hypernym	0	10	6.99	1.976

Validity Evidence

When this research commenced, hypotheses were formulated about the nature of the item types that were designed. The main idea was that different types of items could be developed to measure students' partial knowledge of specific words. However, we emphasize that the nature of this research is exploratory. In this section, we attempt to explain the relationships that we observed, but our conclusions come with the caveat that we were only able to test a relatively small number of words across item types; therefore, we should not generalize but only infer what was observed. With subsequent data collections, we hope that we will observe similar patterns that will add further evidence supporting our theories.

The hypotheses that motivated the development of the three item types imply certain relationships among them. All else being equal, we might expect the idiomatic associates item type to be easier to answer than the topical associates item type; that in turn might be expected to be easier to answer than the hypernym item type based upon the kinds of knowledge structures each item type is designed to draw upon. But all things are seldom equal, and with a small dataset, the expected underlying relationships (if present) may not be easy to tease out. Table 4 shows the surface-level performance of each item type, which does not show the order of difficulty among item types one might expect.

With only 20 cases to analyze, however, the influence of outliers can be large; for instance, two of the 20 idiomatic associates items have a P + below that which would be expected by chance, and their inclusion lowers the mean P+ below the level displayed by the other two item types. Yet there are other ways to evaluate the validity of the item design.

To begin with, we may consider that we are postulating a construct—the depth of semantic knowledge of word meaning—which is, ex hypothesis, more fully measured the deeper an item probes for the knowledge of a word's meaning. It follows that we would expect (all other things being equal) that the hypernym item type would correlate more strongly with an independent measure of this construct than the topical associates item type, and the topical associates item type to correlate more strongly with an independent measure of this construct than the idiomatic associates item type. This relationship is easily measured with our data, since we have an anchor set of 20 Word Generation synonym items, selected to have reasonably high reliabilities with total score, and to span a broad range of difficulty levels. Such synonym items require a high degree of semantic knowledge, since the respondent must differentiate between an exact synonym and nonsynonym distractors that may be plausibly similar and related to the targeted word. We correlated the total score of each item type with the total Word Generation score and obtained the following correlations ($N = 3{,}075$), which fall in the expected order, as shown in Table 5.

Table 5 Correlations Between the Three Item Types and a Separate Measure of the Construct

Item type	Correlation with word generation anchor set
Hypernym total correct	.61
Topical associate total correct	.58
Idiomatic associate total correct	.54

Note. $p < .001$.

Table 6 Discrimination Means and Standard Deviations for the Three Item Types

Statistic	Idiomatic associate	Topical associate	Hypernym
Mean *a*'s	.70	.89	.94
SD	.25	.29	.35

Note. a = parameter for meaningful comparison of the discrimination.

Another piece of validity evidence can be derived by considering the relationship between the mean 3PL IRT parameters. As already pointed out above, we would expect that there would be a relationship between item discrimination and difficulty and items that measure deeper aspects of semantic knowledge. The difficulty parameter (*b*), may be a somewhat less clear measure of depth of vocabulary knowledge, since it represents only the inflection point in the IRT model, which may be affected by a variety of factors not strongly associated with depth of vocabulary knowledge (e.g., word frequency), whereas the discrimination parameter (*a*) indicates how well an item separates individuals with high levels of vocabulary knowledge (who are therefore much more likely to have a deeper knowledge of a particular word) from those with much less knowledge (and are therefore likely to have much shallower knowledge of that word). And in fact, if we calculate the mean discriminations for each item type, they fall in exactly the expected relationship, as shown in Table 6. Higher values indicate the item type better discriminates between higher and lower-ability students (who were defined by their performances on the anchor set of Word Generation synonym items).

A two-way analysis of variance (ANOVA) comparing discrimination differences between item types and across grades yielded no main effects or interactions involving the grade in which students took the tests. However, there was a main effect with item type $F(2, 5) = 7.197$, $p < .01$, indicating that there were differences in the mean discriminations. A post hoc Tukey's honestly significant difference (HSD) test demonstrated that performances on the idiomatic associates item type were significantly less than performances on the topical associates ($p = .02$) and the hypernyms ($p < .001$). Although the means are different between the topical associates and the hypernyms, it is not a statistically significant difference and with this minimal number of items, the pattern can only be deemed as suggestive.

A manual analysis of the design of items was conducted to investigate how well we adhered to the design specifications. Topical associates were constrained by the expectation that all three cue words would be associated with the key and that none of them would have strong associations with the distractors. After the initial item analyses were completed, the stimuli of 10 random topical associate items were manually coded for how well they met this constraint. This coding system was tested on the remaining ten items. The average *a* parameter for items that were judged to meet the constraint perfectly was 0.92. As Table 7 shows, items judged not to meet the constraint perfectly were far more likely to fall below the mean.

Table 7 Distribution of Topical Associate Items by Independent Judgment of Design Fit

Item type	Number of items where $a < .92$	Number of items with $a \geq .92$
Topical associate items judged not to fully complywith the intended design	9	1
Topical associate items judged to closely matchthe intended design	20	10

Note. $N = 40$ as parameters for seventh and eighth grades were calculated separately. a = parameter for meaningful comparison of the discrimination.

Table 8 Distribution of Idiomatic Associate Items with Attractive Distractors

Item type	Number of items where $a < 0.6$	Number of items where $a \geq 0.6$
Idiomatic associate items with attractive distractors	2	10
Idiomatic associate items without attractive distractors	12	16

Note. $N = 40$ as parameters for seventh and eighth grades were calculated separately. a = parameter for meaningful comparison of the discrimination.

Table 9 Distribution of Hypernym Items with Attractive Distractors

Item type	Number of items where $a < 1$	Number of items where $a \geq 1$
Hypernym items with attractive distractors	7	1
Hypernym items without attractive distractors	18	14

Note. $N = 4\,0$ as parameters for seventh and eighth grades were calculated separately. a = parameter for meaningful comparison of the discrimination.

This post hoc manual analysis of the item properties suggests that the differences between the topical associates and the hypernyms (shown in Table 6) might have been significantly different if the item design process had been more tightly controlled.

This type of analysis carried over into the other item types. For instance, one of the design constraints for all of the item types was to ensure that no attractive distractors were included in the three option choices. However, the initial item analysis (described previously) indicated a significant number of idiomatic associates items contained attractive distractors. The mean *a* parameter for the items without attractive distractors was 0.6, and as Table 8 shows, nearly all of the items with attractive distractors were above the mean, which suggests that if they had been more tightly constrained, the separation between these items could have been even stronger.

On the other hand, for the hypernym item type, the effects of attractive distractors went exactly in the opposite direction. Seven of the eight items with attractive distractors were below the mean for hypernym items without attractive distractors; thus, it is plausible that if such items had been avoided, the hypernym items would have been more strongly separated from the other two item types (see Table 9).

The net implication of these post-hoc analyses suggests that the differentiation of the item types by discrimination could have been even stronger than it was in this study and suggests the hypothesis that strict adherence to the item design specifications should replicate the observed effects even more clearly.

Further Validity Evidence Supported by Natural Language Processing Features

Additional analyses were conducted to ascertain whether certain NLP features could be used to predict the IRT parameters. NLP features are statistics that are calculated from corpus data using computational techniques that are used to classify different linguistic characteristics of text numerically.

Our items were constructed according to strict specifications designed to measure specific types and various degrees of semantic knowledge of words. This makes each item type a relatively simple, pure assessment of a specific type of knowledge. We determined that an attempt to predict the IRT parameters was far more likely to be successful if the vocabulary knowledge followed a gradient of depth and our item types succeeded in their design of measuring different levels of depth.

Since we had 20 items per item type, distributed over two grades, the sample sizes were just large enough to support a regression analysis. Given the relatively small number of NLP features we wanted to use in the regression, we avoided any model that had a large number of significant predictors. We treated this analysis as purely exploratory, to be tested by extension to additional items in subsequent studies.

We collected a series of NLP features designed to measure aspects of word knowledge, as follows:

- *Word frequencies*—These are from the TASA corpus (Zeno et al., 1995) and from the SourceFinder corpus, a 425-million-word collection of journal articles and readings appropriate to K–12.

- *Conditional probabilities*—These are based upon how frequently words co-occur in the SourceFinder corpus. This is a direct measure of the probability of one word appearing given another word's appearance within the same paragraph.
- *Association cosines*—These were calculated from a database developed by this project that identified clusters of topically associated vocabulary based upon the same corpus we used previously for word frequencies, which enabled us to identify how closely a given word was associated with specific *topics*. We were able to use this tool to determine how closely the words used in the topical associates items conformed to the intended item design specification. These cosines indicate whether the pattern of associations in which a word participates is similar to, or different from, the general pattern of associations found for the targeted words in our items.
- *Semantic vector measurements.* We had relatively easy access to two measures: latent semantic analysis (LSA; Landauer, Foltz, & Laham, 1998), and correlated occurrence analog to lexical semantics (COALS; Rohde, Gonnerman, & Plaut, 2005). These features measure the general latent tendency of any pair of words to appear in the same documents; thus a high cosine value between two words indicates that they tend to occur in similar topical, semantic, or syntactic contexts.
- *Word fit cosines* (Deane, 2003). These resemble LSA or COALS vectors, in that they are built using the same mathematical methods (singular value decomposition), but the underlying data is the association between words and phrasal contexts. As a result this measure provides estimates of how plausible a word sounds in a phrase, based upon corpus data.

We ran a stepwise regression to attempt to predict the IRT parameters of each item. We then ran separate regressions analysis for each item type. The results indicated that we could, in fact, predict item parameters, particularly difficulty and discrimination. In general, the models performed at a moderate level. The analyses showed interesting relationships among the item types, which will be discussed after we review the individual models.

Predicting Parameters for the Idiomatic Associates Item Type

A stepwise regression yielded a model in which discrimination for the idiomatic associates item type was predicted by one factor: the LSA cosine between the target word and the key. This model achieved an R of 0.52, an adjusted R^2 of 0.25, and a standard error of 0.22.

The model for the difficulty of the idiomatic associates item type was predicted by two NLP features, each of which was positively correlated with difficulty: (a) the WordFit cosine between the target word and the best distractor, indicating a distractor that also had some attraction to the sentential context, and (b) the LSA cosine between the target word and the best distractor. This model achieved an R of 0.61, an adjusted R^2 of 0.34, and a standard error of 1.02.

Predicting Parameters for the Topical Associates Item Type

For the topical associates, a stepwise regression yielded a model in which the discrimination was predicted by a number of NLP features: (a) the COALS cosine similarity between the key and the best distractor (indicating that at least one of the answers was easily confused with the key); (b) the word frequency; and (c) a feature that coded the pattern of association cosines between the three stimulus words and the key.

This last feature was intended to measure how well the topical associate item fit the design template we had specified for this item type, in which all three stimulus words should be associated with the key but with none of the options. If all three stimulus words were above a threshold association cosine with the key but none of the options, this feature received the value of 0. If only two stimulus words were above the threshold value, the feature value was 1. If only one stimulus word was above the threshold value, the feature value was 2. If any of the options were above threshold value for any of the stimulus words, the feature value was 3. This provided a stricter definition of our intended item design than was available when the items were first constructed, and it allowed us to identify which items deviated from the intended design and to quantify how serious that deviation was.

The model performed moderately well with an $R = 0.62$, adjusted $R^2 = 0.33$, and standard error of 0.24 on features (a) and (b), which strongly increased discrimination, and a negative coefficient, which represented the failure to match the item design as specified in the topic mapping tool.

Table 10 Constants in the Regression Equations Predicting Item Response Theory (IRT) Item Parameters for Each Item Type

Analysis	Idiomatic associate	Topical associate	Hypernym
Discrimination	0.43	0.64	0.92
Difficulty	−2.50	−0.63	1.23

A stepwise regression yielded a model in which difficulty for the topical associates item type was predicted by a one feature: how well the item fitted the ideal design specified using the topic tool (failure to do so makes items more difficult). This model achieved an R of 0.41, an R^2 of 0.14, and a standard error of 0.62.

Predicting Parameters for the Hypernym Item Type

For hypernyms, a stepwise regression yielded a model in which two NLP features predicted the discrimination of the hypernyms rendering an R of 0.65, an adjusted R^2 of 0.40, and a standard error of 0.27:

1. COALS cosine between the targeted word and the best distractor.
2. COALS cosine between the key and the best distractor.

In other words, the model predicts that hypernym items will discriminate best to the extent that distractors are reasonably similar to the word/hypernym pair that defines the item.

A stepwise regression also yielded a model in which the two word frequency measures (the TASA SFI and the Log Sourcefinder frequency) combined to predict the difficulty of Hypernyms (with an R of 0.58, an adjusted R^2 of 0.32, and a standard error of 0.81). As one would expect, more frequent words were easier; less frequent words, more difficult. This result makes sense if we interpret the coefficients as creating a weighted average, using the Sourcefinder frequency to discount what appear to be slightly inflated estimates of word frequency in the TASA SFI measure.

Interpretation of the Regressions

A striking feature of the models that resulted is that the intercepts (i.e., constants) are consistent with the ordering of the three item types by depth in both difficulty and discrimination. On a theoretical basis, we would expect hypernyms to require the deepest semantic knowledge of targeted vocabulary words; idiomatic associates to require the least. This should correspond, in turn, to hypernyms tending to have the highest difficulties and discriminations, and idiomatic associates tending to have the least. This hypothesis is consistent with the regression analyses we have obtained, since the constants in the regression analyses (reported above) fall in the predicted order, as shown in Table 10. These results provide some confirmatory evidence that the three item types are, in fact, measuring knowledge at different levels of depth of knowledge, at least when predictable variance among items of the same type is factored out. What the regressions are doing is accounting for sources of error. In the case of the topical associates and idiomatic associates item types, the regressions also indicate that we are making these item types more difficult than intended when the design specifications are not followed closely.

Discussion

This study has a number of features that should to be taken into account before any inferences are drawn about its larger implications. First, only a limited number of words could be tested; since these were drawn from a set of academic vocabulary targeted for instruction in the middle school grades, caution should be exercised when making inferences about how and whether the results may extend to different types of vocabulary. Second, the population consisted entirely of seventh and eighth grade students in a convenience sample of U.S. schools. Additional studies will be needed to determine how and whether these results may vary by population. Finally, a key feature of the study design was that it required each student to make multiple judgments about the same word. This feature creates the possibility of priming between items, where (for instance) prior exposure to a word could produce cuing effects facilitating answers on subsequent items testing the same word. For the purposes of the analyses presented below, we pooled data across orders of presentation and treated any differences resulting from order as noise.[4]

An important implication of this study is that it strongly supports the feasibility of designing vocabulary items to fit a cognitive model. Each item type was built from the ground up to measure a different aspect or level of a theoretically motivated construct (depth of semantic knowledge). We defined and applied a consistent construction principle for each item type, informed by cognitive theory and took advantage of corpus resources to control potential sources of variation. The validity evidence suggests that this construction was successful, yielding differences among the item types consistent with the theoretical basis on which it was built.

Our results suggest that we did not achieve this goal perfectly. Some items appear to deviate from the intended model, but in ways that tend to confirm that the other items are functioning as intended. We intend to follow up this study with additional studies that replicate the results with different words, different populations, and closer control of variables that account for variations in item functioning. If these results confirm our initial findings, it may be possible to define precise construction principles for creating vocabulary items designed to measure the depth of vocabulary knowledge.

Given the design on which each item was based, we would also expect that the differences among item types might prove useful for discriminating among different patterns and levels in the acquisition of semantic knowledge. For instance, English language learners might acquire a large number of words purely from direct instruction, without the incremental development that reading large numbers of texts might provide. This might correspond to a shift in the pattern of performance across item types. Similarly, if some students had extensive reading experience but less facility in inferring conceptual meanings from text, there might be a shift in the relative difficulty of shallow versus deeper item types along the continuum we have begun to explore. Such possibilities go well beyond the conclusions that can safely be drawn from this study alone, but they suggest a line of research that might fruitfully be explored.

Acknowledgments

The work reported in this paper was supported under IES/PR Award Number R305A080647.

Notes

1 A *semantic associates* item type, intended to measure the fourth level of depth of vocabulary knowledge did not work well in pilot testing, and was excluded from the study.
2 "When interpreting SFI (Standard Frequency Index) values, note that the SFI statistics form a logarithmic scale, like the Richter scale used to evaluate the magnitude of earthquakes. As a result, arithmetic differences in SFI values correspond to geometric differences in word frequency" (Zeno et al., 1995, p. 12).
3 http://www.questarai.com/Products/WordFrequencyGuide/Pages/default.aspx 2-10-11
4 There do appear to be ordering effects; for two of the item types (topical associates and hypernyms) order was a significant predictor, with absolute position in the form accounting for half the variance in form means for the topical associates item type, and about one third of the variance in form means for the hypernym item type. These can be interpreted as cuing effects, with prior experience with a word facilitating later performance. It will be worthwhile to examine, in future studies, how such effects can be controlled or eliminated, perhaps by collecting data from different item types on different test sessions.

References

Anderson, R. C., & Freebody, P. (1981). Vocabulary knowledge. In J. Guthrie (Ed.), *Comprehension and teaching: Research reviews* (pp. 77–117). Newark, DE: International Reading Association.

Bates, E., & MacWhinney, B. (1989). Functionalism and the competition model. In B. MacWhinney & E. Bates (Eds.), *The crosslinguistic study of sentence processing* (pp. 3–73). New York, NY: Cambridge University Press.

Beck, I., & McKeown, M. (1991). Conditions of vocabulary acquisition. In R. Barr, M. Kamil, P. Mosenthal & P. D. Pearson (Eds.), *Handbook of reading research* (Vol. 2, pp. 789–814). New York, NY: Longman.

Beck, I. L., McKeown, M. G., & McCaslin, E. C. (1983). Vocabulary development: Not all contexts are created equal. *Elementary School Journal*, *83*, 177–181.

Bird, H., Howard, D., & Franklin, S. (2003). Verbs and nouns: The importance of being imageable. *Journal of Neurolinguistics*, *16*, 113–149.

Breland, H. M., & Jenkins, L. M. (1997). *English word frequency statistics: Analysis of a selected corpus of 14 million tokens*. New York, NY: College Entrance Examination Board.

Breland, H. M., Jones, R. J., & Jenkins, L. (1994). *The College Board vocabulary study* (Research Report No. RR-94-26). Princeton, NJ: Educational Testing Service.

Brown, J. C., Frishkoff, G. A., & Eskenazi, M. (2005, October). Automatic question generation for vocabulary assessment. In *Proceedings of human language technology conference and conference on empirical methods in natural language processing (HLT/EMNLP)* (pp. 819–826). Vancouver, Canada: Association for Computational Linguistics.

Carroll, J. B. (1970). An alternative to Juilland's usage coefficient for lexical frequencies, and a proposal for a standard frequency index (SFI). *Computer Studies in the Humanities and Verbal Behavior*, *3*, 61–65.

Carroll, J. B. (1971). Measurement properties of subjective magnitude estimates of word frequency. *Journal of Verbal Learning and Verbal Behavior*, *10*, 722–729.

Carroll, J. B. (1976). *Word retrieval latencies as a function of frequency and age-of-priming, repeated trials, and individual differences* (Research Report No. RR-76-07). Princeton, NJ: Educational Testing Service.

Carroll, J. B. (1993). *Human cognitive abilities: A survey of factor analytic studies*. New York, NY: Cambridge University.

Carroll, J. B., & White, M. N. (1973). Word frequency and age of acquisition as determiners of picture-naming latency. *Quarterly Journal of Experimental Psychology*, *25*, 85–95.

Chabot, R. J., Petros, T. V., & McCord, G. (1983). Developmental and reading ability differences in accessing information from semantic memory. *Journal of Experimental Child Psychology*, *35*, 128–142.

Coltheart, V., Laxon, V. J., & Keating, C. (1988). Effects of word imageability and age of acquisition on children's reading. *British Journal of Psychology*, *79*, 1–12.

Cunningham, A. E., & Stanovich, J. K. (1997). *Developmental Psychology*, *33*, 934–945.

Dale, E., & O'Rourke, J. (1986). *Vocabulary building*. Columbus, OH: Zaner-Bloser.

Daneman, M. (1988). Word knowledge and reading skill. In M. Daneman, G. MacKinnon & T. G. Waller (Eds.), *Reading research: Advances in theory and practice* (Vol. 6, pp. 145–175). San Diego, CA: Academic Press.

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

Deane, P. (2003). Co-occurrence and constructions. In L. Lagerwerf, W. Spooren & L. Desgand (Eds.), *Determination of information and tenor in texts: Multidisciplinary approaches to discourse* (pp. 277–304). Amsterdam, the Netherlands: Stichting Neerlandistiek VU.

Deane, P. (2005). A nonparametric method for extraction of candidate phrasal terms. In *Proceedings of the 43rd annual meeting of the Association for Computational Linguistics* (pp. 605–613). New Brunswick, NJ: Association for Computational Linguistics.

Deane, P., & Higgins, D. (2006). Using singular-value decomposition on local word contexts to derive a measure of constructional similarity. In E. Fitzpatrick (Ed.), *Corpus linguistics beyond the word: Corpus research from phrase to discourse*. Amsterdam, the Netherlands: Rodopi.

Dixon, P., LeFevre, J., & Twilley, L. C. (1988). Word knowledge and working memory as predictors of reading skill. *Journal of Educational Psychology*, *80*, 465–472.

Dunn, L. M., & Dunn, L. M. (1997). *Peabody picture vocabulary test* (3rd ed.). Circle Pines, MN: American Guidance Service.

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

Elman, J. L. (2009). On the meaning of words and dinosaur bones: Lexical knowledge without a lexicon. *Cognitive Science*, *33*, 547–582.

Embretson, S. E. (1998). A cognitive design system approach to generating valid tests: Applications to abstract reasoning. *Psychological Methods*, *3*, 300–396.

Embretson, S. E., & Gorin, J. (2001). Improving construct validity with cognitive psychology principles. *Journal of Educational Measurement*, *38*, 343–368.

Fellbaum, C. (2010). WordNet. In R. Poli, M. Healy & A. Kameas (Eds.), *Theory and applications of ontology: Computer applications* (pp. 231–243). Berlin, Germany: Springer.

Freedle, R., & Kostin, I. (1992). *The prediction of* GRE *reading comprehension item difficulty for expository prose passages for each of three item types: Main ideas, inferences, and explicit statements* (Research Report No. RR-91-59). Princeton, NJ: Educational Testing Service.

Fukkink, R. G., Henk, B., & De Glopper, K. (2001). Deriving word meaning from written context: A multicomponential skill. *Language Learning*, *51*, 477.

Gao, L., & Rogers, T. (2007, April). *Cognitive-psychometric modeling of the MELAB reading items.* Paper presented at the annual meeting of the National Council of Measurement in Education, Chicago, IL.

Gorin, J. (2005). Manipulating processing difficulty of reading comprehension questions: The feasibility of verbal item generation. *Journal of Educational Measurement*, *42*, 351–373.

Gernsbacher, M. A. (1984). Resolving 20 years of inconsistent interactions between lexical familiarity and orthography, concreteness, and polysemy. *Journal of Experimental Psychology: General*, *113*(2), 256–281.

Graves, M. F. (1986). Vocabulary learning and instruction. In E. Z. Rothkopf (Ed.), *Review of Research in Education, 13*, 49–89.

Hirsch, E. D. (2003). Reading comprehension requires knowledge of words and the world. *American Educator*, *27*(1), 10–31.

Hogaboam, T. W., & Perfetti, C. A. (1975). Lexical ambiguity and sentence comprehension. *Journal of Verbal Learning and Verbal Behavior*, *14*, 265–274.

Hu, M., & Nation, P. (2000). Unknown vocabulary density and reading comprehension. *Reading in a Foreign Language*, *13*, 1.

Kaplan, E. F., Goodglass, H., & Weintraub, S. (1978). *The Boston naming test*. Boston, MA: E. Kaplan & H. Goodglass.

Landauer, T. K., Foltz, P. W., & Laham, D. (1998). Introduction to latent semantic analysis. *Discourse Processes*, *25*, 259–284.

McFalls, E. L., Schwanenflugel, P. J., & Stahl, S. A. (1996). Influence of word meaning on the acquisition of a reading vocabulary in second-grade children. *Reading and Writing*, *8*, 235–250.

McKoon, G., & Ratcliffe, R. (1998). Memory based language processing: Psycholinguistic research in the 1990s. *Annual Review of Psychology*, *49*, 25–42.

Mislevy, R., Johnson, E., & Muraki, E. (1992). Scaling procedures in NAEP. *Journal of Educational and Behavioral Statistics*, *17*, 131–154.

Myers, J. L., & O'Brien, E. J. (1998). Accessing the discourse representation during reading. *Discourse Processes*, *26*, 131–157.

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.

Nagy, W., & Scott, J. (2000). Vocabulary processes. In M. Kamil, P. Mosenthal, P. D. Pearson & R. Barr (Eds.), *Handbook of reading research* (Vol. III, pp. 269–284). Mahwah, NJ: Erlbaum.

Paivio, A. (1971). *Imagery and verbal processes*. New York, NY: Holt, Rinehart and Winston.

Perfetti, C. A. (1994). Psycholinguistics and reading ability. In M. A. Gernsbacher (Ed.), *Handbook of psycholinguistics* (pp. 849–894). San Diego, CA: Academic Press.

Perfetti, C. A., & Hart, L. (2001). The lexical quality hypothesis. In L. Verhoeven, C. Elbro & P. Reitsma (Eds.), *Precursors of functional literacy* (Vol. 11, pp. 67–86). Amsterdam, the Netherlands: John Benjamins.

Rohde, D. L., Gonnerman, L. M., & Plaut, D. C. (2005). *An improved model of semantic similarity based on lexical co-occurrence*. Unpublished manuscript.

Rupp, A. A., Ferne, T., & Choi, H. (2006). How assessing reading comprehension with multiple-choice questions shapes the construct: A cognitive processing perspective. *Language Testing*, *23*, 441–474.

Sahhouse, T. A. (1993). Speed and knowledge as determinants of adult age differences in verbal tasks. *Journal of Gerontology: Psychological Sciences*, *48*, 29–36.

Schatz, E. K., & Baldwin, R. S. (1986). Contextual clues are unreliable predictors of word meanings. *Reading Research Quarterly*, *21*, 439–453.

Schwanenflugel, P. J., Stahl, S. A., & McFalls, E. L. (1997). Partial word knowledge and vocabulary growth during reading comprehension. *Journal of Literacy Research*, *29*, 531–553.

Scott, J. A., Hoover, M., Flinspach, S. L. & Vevea, J. L. (2008). A multiple-level vocabulary assessment tool: Measuring word knowledge based on grade level materials. In Y. Kim, V. J. Risko, D. L. Compton, D. K. Dickinson, M. K. Hundles, R. T. Jimenes, K. M. Leandor, & D. W. Rowe (Eds.), *57th annual yearbook of the National Reading Conference* (pp. 325–340). Oak Creek, WI: National Reading Conference.

Sheehan, K. M., & Ginther, A. (2001, April). *What do passage-based multiple-choice verbal reasoning items really measure? An analysis of the cognitive skills underlying performance on the current TOEFL reading section.* Paper presented at the annual meeting of the National Council on Measurement in Education, Seattle, WA.

Sheehan, K. M., Ginther, A., & Schedl, M. (1999, December). *The skills underlying performance on the current TOEFL reading comprehension section.* Paper presented at the National Reading Conference, Orlando, FL.

Sheehan, K. M., & Mislevy, R. J. (2001). *An inquiry into the nature of the sentence-completion task: Implications for item generation* (Research Report No. RR-01-13). Princeton, NJ: Educational Testing Service.

Sheehan, K. M., & Mislevy, R. J. (1990). Integrating cognitive and psychometric models to measure document literacy. *Journal of Educational Measurement*, *27*, 1–18.

Sheehan, K. M., Kostin, I., & Persky, H. (2006, April). *An examination of the reading skills underlying performance on the NAEP Grade 8 reading assessments*. Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco, CA.

Snow, C. E., Lawrence, J. F., & White, C. (2009). Generating knowledge of academic language among urban middle school students. *Journal of Research on Educational Effectiveness*, *2*, 325–344.

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

Stanovich, K. E., West, R. F., & Harrison, M. (1995). Knowledge growth and maintenance across the life span: The role of print exposure. *Developmental Psychology*, *31*, 811–826.

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

Walczyk, J. J., & Raska, L. J. (1992). The relation between low- and high-level reading skills in children. *Contemporary Educational Psychology*, *17*, 38 – 46.

Zeno, S. M., Ivens, S. H., Millard, R. T., & Duvvuri, R. (1995). *The educator's word frequency guide*. Brewster, NY: Touchstone Applied Science Associates.

Zevin, J. D., & Seidenberg, M. S. (2002). Age of acquisition effects in word reading and other tasks. *Journal of Memory and Language*, *47*, 1 – 29.

Action Editor: Joel Tetreault

Reviewers: Michael Heilman and Yoko Futagi

ETS, the ETS logo, and LISTENING. LEARNING. LEADING., are registered trademarks of Educational Testing Service (ETS). All other trademarks are property of their respective owners.

Find other ETS-published reports by searching the ETS ReSEARCHER database at http://search.ets.org/researcher/