An Innovative Approach to Assessing Depth of Knowledge of Academic Words

Amy C. Crosson[1], Margaret G. McKeown[2] & Arthur K. Ward Jr.[3]
[1]Pennsylvania State University, University Park, USA; [2]University of Pittsburgh, Pittsburgh, USA; [3]Fluent Tutor, Pittsburgh, USA

To cite this article: Crosson, A. C., McKeown, M. G., & Ward Jr., A. K. (2019). An innovative approach to assessing depth of knowledge of academic words. *Language Assessment Quarterly*, 16(2), 196-216. doi:10.1080/15434303.2019.1612899

To link to this article: https://doi.org/10.1080/15434303.2019.1612899
Peer review process: https://www.tandf.co.uk/journals/authors/hlaqauth.pdf

The research reported here was supported by the Institute of Education Sciences, U.S. Department of Education, through Grant R305A100440 to the University of Pittsburgh. The opinions expressed are those of the authors and do not represent views of the Institute or the U.S. Department of Education. Correspondence regarding this article should be addressed to Amy C. Crosson, Curriculum and Instruction, Pennsylvania State University, 259 Chambers Bldg., University Park, PA 16802-1503, USA.

Table of Contents

ABSTRACT

The purpose of this research is to develop a vocabulary assessment, Evaluation of Academic Vocabulary or “EAV,” that gauges students’ depth of knowledge of academic words and is sensitive to vocabulary growth in the context of literacy interventions. We report results from two studies with native English-speaking middle school students designed to inform the development and pilot test the utility and technical qualities of an innovative vocabulary assessment that is sensitive to students’ level of understanding of general academic words. Across the two studies, we employed a mixed methods research approach. In the first study, we drew on classic psychometric assessment methods, Bayesian network methods, signal detection theory to evaluate the validity and technical qualities of the assessment. In the latter study, we conducted cognitive interviews to further validate the assessment. Results suggest that: 1) the assessment captures growth in knowledge of general academic words, i.e., words that are important for the comprehension of academic texts; and 2) it is sensitive to treatment effects academic vocabulary intervention for adolescents.

Keywords: vocabulary, intervention, academic language

An Innovative Approach to Assessing Depth of Knowledge of Academic Words

Assessing students' knowledge is a persistently thorny issue in vocabulary development, as word knowledge is a highly complex phenomenon. Not only is there a variety of aspects to know about any single word, such as its form, meaning, and conditions of use, but words interrelate in ways that influence how successfully and appropriately we employ them in understanding and expressing language (Nagy & Scott 2000; Nation, 2001). Surface level knowledge of word meaning components is not sufficient to support students' higher-level language skills such as comprehension (Baumann, Kame'enui, & Ash, 2003; Beck, McKeown, & Omanson, 1987; Pearson, Hiebert, & Kamil, 2007; Stahl & Fairbanks, 1986). This situation makes assessing students' knowledge a challenge. Clearly there is a need to assess facets of word knowledge, and do so in efficient ways.

The purpose of the research reported here was to develop a vocabulary assessment that gauges students' depth of knowledge of general academic words and is sensitive to learning over time. The assessment, Evaluation of Academic Vocabulary (EAV), was designed for use with adolescent learners who are native English speakers and for testing treatment effects of literacy interventions. We take a mixed methods approach to investigating the effectiveness and technical qualities of the assessment (Creswell, 2014). Our approach aligns with Creswell's Sequential Explanatory design strategy, which is characterized by collecting and analyzing quantitative data followed by collecting and analyzing qualitative data. Specifically, we administered our assessment in a conventional format followed by a second study using cognitive interviews to explain and interpret the quantitative results.

Introduction

Theoretical Framework

Learners acquire information about word meaning from cumulative encounters with words in informative contexts. From these encounters, a learner is able to generalize a word's meaning elements and features of its use such as collocations and connotations, which leads to establishing a flexible, nuanced representation of word meaning (Bolger, Balsas, Landen, & Perfetti, 2008; Nagy & Scott, 2000; Perfetti, 2007; Perfetti & Stafura, 2014). From such representations, learners build rich networks of connections among words, contexts in which words are typically found, and experiences that relate to word meanings. These rich lexical networks enable learners to bring relevant connections to bear to make sense of newly encountered contexts containing the words. Thus, words are not isolated pieces of information in memory.

Given the multidimensional nature of word knowledge, researchers have tried to conceptualize those dimensions and their implications. The general focus of such work has been the notion of vocabulary "depth," or how well one knows a word, in contrast to vocabulary "breadth" which is typically thought of as how many words one knows at a surface level of form-meaning associations.

Various conceptualizations for depth of knowledge have been offered by vocabulary scholars. Read (2000), distinguishes between two approaches to operationalizing and measuring depth of vocabulary knowledge. The first is a matter of degree, such that one might "know" a word along a continuum from no knowledge to mastery (i.e., appropriate use of the word in different contexts). The second is multidimensional, which suggests that there are many facets to word knowledge including, but not limited to, knowledge of collocations, syntactic information, register, and associations to other words. Of those scholars whose approach is multidimensional, perhaps the most comprehensive is Nation's (2001/2013. p 27), based on three broad categories of *form, meaning,* and *use*. These categories are further subdivided into areas such as, for the *meaning* category: *form and meaning*, *concepts and referents*, and *associations*. Each subcategory is paired with questions, such as "What meaning does this word form signal?" "What is included in the concept?" and "What words or types of words occur with this one?" Schmitt (2014) suggests that depth could be conceptualized as some degree of mastery over one or more of Nation's aspects.

Some scholars have questioned whether the distinction between vocabulary depth and breadth is useful. Indeed, the high correlation between measures of breadth (i.e., recognition of definitions) and depth has led many to assert that the two are not conceptually different (Schmitt, 2014; Vermeer, 2001). Yet other researchers have found that depth measures make important and distinct contributions to reading comprehension (Li & Kirby, 2012; Qian & Schedl, 2004). For learners who gain vocabulary knowledge as a consequence of numerous, meaningful interactions with language, the depth of knowledge of words grows in parallel with the size of their vocabulary. In such cases, performance on breadth measures may be a fair representation of depth of knowledge as well.

The distinctions between breadth and depth of vocabulary take on greater meaning, however, when we consider instruction as a vehicle for vocabulary growth. Different types of instruction can result in very different types of knowledge. If students are simply asked to practice definitions, their breadth of vocabulary increases, as measured by ability to recognize definitions. But familiarity with definitions is a superficial level of vocabulary knowledge, unlikely to support language comprehension or production (McKeown, Crosson, Moore, &

Beck, 2018; Nagy & Scott, 2000). To understand the success of instructional interventions we need to be able to characterize the kinds of knowledge that students have acquired.

Assessment in Vocabulary

Status of vocabulary assessment. The vocabulary field has long recognized the complex, multidimensional nature of word knowledge (Anderson & Nagy, 1991; Miller, 1978). Yet most vocabulary assessments used in US schooling contexts have not reflected this complexity (Pearson, Hiebert, & Kamil, 2007; 2012). In school-based intervention studies, assessing multiple dimensions of word knowledge has not been the norm (Elleman, Lindo, Morphy, & Compton, 2009). Although the motivation for vocabulary instruction is to enable students to apply that knowledge for comprehension and production, measuring what students can do with the words in their repertoires is often neglected. Rather, vocabulary assessments in schools have relied on assessing definitional knowledge, most commonly requiring students to select a definition for a word from choices (Scott, Lubliner, & Hiebert, 2006). Definition-based assessment implies an all or nothing picture of word knowledge, suggesting that recognizing a definition equates to "knowing" a word while failure to correctly identify a definition means lack of knowledge.

Newer large-scale standardized vocabulary assessments developed for schooling contexts in the US are beginning to change by requiring that students apply their knowledge of a word by integrating its meaning into sentence or passage contexts. Examples include vocabulary portions of the GRE for university application (Educational Testing Service, 2011) as well as NAEP (National Assessment Governing Board, 2008), the "nation's report card" that periodically tests knowledge and skills in language arts based on a cross-sectional national sample. As well, assessments developed to align with Common Core State Standards (PAARC, 2016; Sireci, 2012)—that is, the national effort toward defining grade-level competencies and knowledge—similarly require word-text integration when testing word knowledge. These represent advances over conventional assessments, yet they do not tap multiple aspects of vocabulary knowledge (Crosson, McKeown, Beck, & Ward, 2012; McKeown, Deane, Scott, Krovetz, & Lawless, 2017).

Toward assessing vocabulary depth. In the research arena, there have been several efforts to develop assessments to tap vocabulary depth that seem to reflect aspects of Nation's framework. Much of the theoretical work on multidimensional aspects of vocabulary knowledge has been led by scholars in second language (L2) acquisition and L2 scholarship informs the research with native English-speaking adolescents in the research reported here. However, it is important to keep in mind that word learning in one's native language (L1) is distinct from learning in L2. For example, native and non-native English speakers differ in lexical organization and acquisition of dimensions of word knowledge (Nation, 2013). L1 and L2 lexical systems are interrelated, such that L2 learners' word learning is influenced by L1 knowledge, the organization of such L1-L2 connections may change over time (Kroll, Dussias, Bice, & Perrotti, 2015). In L2 learning, dimensions such as collocational knowledge may be acquired later than conceptual word knowledge (Nation, 2013). These distinctions should be kept in mind in the following section as we address assessments that were designed for L2 learners, primarily adults. Insights from L2 assessment design—especially concerning dimensions of word knowledge—is used as a foundation of the current work with monolingual English speakers.

One example of efforts to tap vocabulary depth in L2 that seems to reflect aspects of Nation's framework is Read's test of vocabulary depth (Read, 1989; 1993; 1998) which was designed around lexical organization, or the view of the lexicon as an interrelated network of words and information about word use. In Read's Word Associates Format, learners choose from a list of words those that have either synonymous or collocational associations to target words. Another approach to assessing second language learners' depth was developed by Chui (2006) based on five aspects of word knowledge: recognize part of speech, recall meaning, recognize collocations, produce derivative form, write a sentence. Similarly, Qian (Qian & Schedl, 2004) incorporated syntactic information, meaning, collocational properties, and orthographic information in an operationalized definition of depth of word knowledge. A view of vocabulary depth as knowledge of polysemy was the basis of assessments developed by Schmitt (1998) and more recently by Crossley, Salsbury, & McNamara (2010). In Schmitt's (2014) review of research on depth of vocabulary knowledge, he concludes by saying that the most promising focuses on lexical organization, that is, the degree to which any item is integrated into the rest of the mental lexicon. This view makes contact with Meara and Wolter's (2004) claim that it is the *organization* of lexical knowledge that distinguishes stronger from weaker lexical representations for L2 learners, underscoring the importance of networks of semantic knowledge.

Native English speakers in elementary and high school have been the targets of two recent, large-scale projects to develop vocabulary depth measures (Deane, Lawless, Li, Sabatini, Bejar, & O'Reilly, 2014; Scott, Flinspach, Vevea, & Castaneda, 2015) and, like EAV, both make contact with the research base in L2 vocabulary assessment, as they are designed to assess multiple dimensions of vocabulary knowledge. Deane and colleagues (2014) developed a battery of item types to measure four aspects of vocabulary knowledge (collocational, definitional, topical, and categorical) for a set of academic, domain-specific words. Scott and colleagues (2015) addressed depth of knowledge using multiple-choice items to measure six dimensions of word knowledge (e.g., part of speech, morphological relations, semantically related words).

Both research programs show promise for assessments that examine multiple aspects of vocabulary knowledge. The EAV assessment described here shares a focus on multiple aspects of vocabulary and partial knowledge, but elements of EAV's design, purpose, and content distinguish it.

Design of EAV

EAV presents each word in a set of four sentences rather than 4-6 separate multiple choice items, which is practically advantageous as fewer items are necessary to test facets of word knowledge. Time requirements for assessments are of great concern to school-based educators; longer tests may be passed over simply because of the widespread perception that students already spend too much time taking tests. The purpose of EAV, unlike Deane et al. (2014) and Scott et al.'s (2015) assessments, is to assess growth in knowledge as a result of interventions. Finally, the content assessed on EAV is general academic words that appear with relatively high frequency in academic texts across subject areas, not restricted to specific disciplines. Domain-specific words such as those tested by Deane and colleagues, and words common in elementary grades curricula such as those tested by Scott and colleagues, are clearly important, yet general academic words have been targeted as key to language acquisition in academic contexts (Nagy & Townsend, 2012).

EAV presents four fill-in-the-blank (i.e., gapped or cloze) sentences per target word and requires students to decide whether the target word properly fits in the sentence. Target words assessed are 99 general academic words from the Academic Word List (AWL; Coxhead, 2000), for a total of 396 sentences. AWL comprises words that occur with frequency across content domains, identified from a 3.5 million-word corpus of academic texts across 28 domain areas.[1]

[1] Since publication of the AWL, newer word lists have become available based on rigorous methodologies and larger corpora such as the *Academic Vocabulary List* (AVL) (Gardner & Davies, 2013) and the *New Academic Word*

Sentence contexts, designed by the authors, were written so that specialized vocabulary and world knowledge were not necessary to comprehend them. Some items (e.g., see Martin Luther King, Jr. item for *rational* in Study One) included references to knowledge that would be widely accessible to students in US schooling systems.

Unlike multiple choice tests in which students can select the correct choice and move on, in the EAV test, students must attend to and make a decision about each sentence provided, asking themselves, "does this sentence make sense with [target word] in the blank?" We deliberately provided sentences with a blank, a choice motivated by a hypothesis that asking students to judge completed sentences would prompt a confirmation bias (Lockett & Shore, 2003), while sentences with a blank would prompt students to weigh whether the target word or some other word appropriately completed the sentence. Indeed, our data suggested that even with the fill-in-the-blank design, participants were biased to assume a fit in the face of uncertainty, as indicated in our D prime analysis where we obtained an average criterion score of -0.35. (See Results section for full description of D prime analysis.)

Two design elements of the EAV capture depth of word knowledge. First, it assesses knowledge of multiple senses. Second, foils (i.e., distractors) tap different facets of word knowledge that vary in difficulty, thus performance on different foil types is intended to indicate levels of word knowledge. We elaborate on these design elements below.

Conceptualizing and operationalizing multiple senses. General academic words that are the focus of EAV often carry multiple senses. In our development of an assessment of academic words, we began by recognizing the need to address the nature of academic words as polysemous (Crossley et al., 2010; Schmitt, 1998). By polysemous, we mean words that have a

---

List (NAWL) (Browne, Culligan & Phillips, 2013). For our purposes, AWL was sufficient for identifying target words that are dispersed with frequency across content domains in middle school texts.

single orthographic form but represent multiple, related senses (Pethö, 2001). For example, the *foundation* of a structure and the *foundation* of a theory, while corresponding to very different referents, are in fact related senses as they both refer to the “base on which something else can be built”—that is, they share a core meaning (Srinivasan & Snedeker, 2011). This relatedness is confirmed by their shared etymological roots. While polysemy is sometimes a literal connection in meaning between senses (e.g., referring to “fish” as the animal or as a dish), in the case of academic words, multiple senses are often characterized by metaphorical polysemy (e.g., “see” as “visually perceive” and “understand”). (For a review, see Eddington & Tokowicz, 2015).

Polysemy interacts with another characteristic of academic words – their abstract nature. Often academic words have both concrete and abstract senses. By abstract, we mean words with referents that are not concrete objects such as “hammer,” but instead are abstract conceptions such as “theory.” Schwanenflugel and colleagues (Schwanenflugel, Harnishfeger, & Stowe, 1988) view concrete words as linked to multiple associated memory contexts, and they argue that abstract words (often metaphorical senses) are more difficult to process because learners tend to retrieve less associated contextual information about abstract words from prior knowledge. The abstract conception of many academic words presents challenges for constructing item types that meaningfully capture word knowledge; presenting abstract words in context rather than in isolation may address this challenge by facilitating access to lexical representations of abstract words (Schwanenflugel & Shoben, 1983).

The EAV’s design departs from previous depth measures for young learners in that it explicitly addresses knowledge of multiple senses. To operationalize multiple senses for this study, we established a two-step process. First, two members of the research team coded all target academic words and compared outcomes. Coders were told that the senses must share a core meaning, grounded in the seminal Single Entry Model of semantic relatedness (Nunberg, 1979). Exact agreement was 82.29%, and disagreements were resolved by discussion of coding rationale and arriving at consensus. Second, multiple sense status of the target words was checked by using WordNet (Princeton University, 2010) synset relations; 96.86% of words identified as representing multiple senses by the coders were confirmed. As such, EAV is grounded in this theoretical and empirical framework related to measuring multiple senses.

Conceptualizing and operationalizing levels of word knowledge. Foil types were developed by considering the kind of initial knowledge or associations that learners may establish with a word upon having encountered it, but having not established a rich or precise set of connections to its meaning. The types of foils accounted for facets of word knowledge comprising syntactic, orthographic, and lexical association. We hypothesized that these dimensions mapped onto different levels of word knowledge, building on existing assessments that tap multidimensional knowledge (Chui, 2006; McKeown et al., 2017; Qian & Schedl, 2004).

Syntax foils are contexts in which any word that could plausibly fit in the sentence is a different part of speech from the target word. For example, "I hurried to _____ the contest." for the target word, *integral*. Syntax foils distinguish between shallow levels of word knowledge, as they are based on the idea that learners who have had few encounters with a word may have little knowledge about the word's semantic properties, but have established sufficient memory traces of the word's syntactic function to reject the foil. In contrast, learners who have had no memory traces of a word's use or meaning will not know to reject the syntax foil.

Unrelated foils are contexts that contain no association with the target word, but any word that could plausibly fit in the fill-in-the-blank sentence is the same part of speech as the target word. For example, "I read an _____ fairytale" for the word *empirical*. Learners with shallow word knowledge about the target word – enough to recognize the word's syntactic role—but who are unaware of the word's semantic properties will not know to reject this type of foil.

Orthographic foils represent contexts in which a word that could plausibly complete the sentence is orthographically similar to the target word and is the same part of speech, such that students who do not have enough substantive knowledge of the word to distinguish it from another word that sounds or looks like it might believe that the target word makes sense. These are sentences in which a word that is orthographically similar to the target word, but not the target word itself, would make sense in the sentence. For example, an orthographic foil for *criteria* is constructed around the word *interior*: "During the storm, they gathered in the _____ of the building." This orthographic overlap inherently taps an element of phonological overlap for most items.

Semantic foils are contexts that contain a strong semantic association to the target word and construction similarity through collocations and common contexts of use. For example, a semantic foil for *criteria* is: "When Chris had to choose a college, he met many _____ to help him make a decision" based on *criteria* as something that one uses in the process of making a decision, and the phrase "met criteria" as a common construction in which the word is found. Research with native speakers shows dense networks of lexical associations (e.g., SLEEP, DREAM, PILLOW, BED) that establish semantic clustering of associations (Meara, 2009). Synset information from WordNet (Princeton University, 2010) was consulted to generate prototypical associations with the words drawing from gloss information and example sentences. Collocational associations were deliberately integrated into semantic foils. The intention of this item type was to test whether students were able to integrate word meaning with context beyond the local context of adjacent words to consider the more global context of the full sentence. For example, the phrase "met criteria" is a local context that makes sense on its own, but the broader sentence containing that phrase does not make sense. The similarity between semantic foils and unrelated foils is that both fit syntactically and elicited knowledge about conceptual meaning of target words. The distinction is that only semantic foils were deliberately designed to include semantic associations (e.g., *criteria*-decisions) and collocations.

Our intention in designing foils was that the degree of challenge for each foil type would fall along a continuum with *syntax* being easiest, testing surface-level word knowledge; *unrelated* and *orthographic* falling in the middle, and *semantic* being most difficult, testing deeper knowledge of word meaning and constraints of use.

Research Questions

Our research questions were as follows:

1. Does EAV measure growth in students' knowledge of academic words in the context of a vocabulary intervention?
2. Does performance on the foil types indicate levels of word knowledge as intended, such that their rank order of difficulty is as hypothesized?

In addition, we have explored possible approaches to addressing a third issue:

3. How can the impact of guessing behavior on precision of scores be minimized?

Below, we present results from two studies and explain how they have enabled us to address these issues. Study One addressed all three Research Questions. Study Two focused on our second research question: performance of the foil types. In particular, we explored qualitatively, through cognitive interviews, whether our rationale about the difficulty of item types and the kind of thinking that they elicited were correct.

Study One

Method

Intervention context. Study One was carried out in the context of an academic vocabulary intervention for native English-Speaking students in middle school, *Robust Academic Vocabulary Encounters* (RAVE; Crosson & McKeown, 2016; McKeown et al., 2018). The intervention, implemented in 11 instructional units over a 20-week period, was designed to teach 99 general academic words.

In RAVE, students engaged in analysis of multiple senses and word usage, including syntactic constraints, collocational knowledge, and register. To introduce each target word, students were presented with two non-fiction contexts (approximately 80 words each) illustrating prototypical uses of the word. In the case of multiple senses, different senses were illustrated in these contexts. For example, to introduce *confine*, students analyzed its meaning in a context about Thomas Edison who did not *confine* his creativity to invention of the light bulb and a context about an animal shelter that was forced to *confine* diseased kittens.

Throughout the instructional unit, students encountered and analyzed each target word in at least 15 different contexts. For example, for *extract*, students were exposed to contexts in which red dye was extracted from a cochineal bug, a splinter was extracted from a foot, a confession was extracted from a criminal, and the truth was extracted from a sister. Across multiple encounters, students were guided to interact with the target words both receptively and productively, and both orally and in writing. Teachers guided students to engage in active processing of word meanings through generating examples, considering nuances of meaning, and comparing words for semantic overlap and constraints around word use.

Participants. Participants were 105 sixth-grade students (61 RAVE and 43 control) from five sixth-grade classes (3 RAVE and 2 control) in a public middle school within a working class community in the northeastern US. There was no difference in reading between RAVE and control groups as measured by the total reading score of the Gates-MacGinitie Reading Test (MacGinitie, MacGinitie, Maria, & Dreyer, 2000) at pretest [$F(1, 102)=.008$, $p=.928$,]. About 25% of the students were African American and the rest European American; 55% received free or reduced-priced lunch. All were L1 English-speakers. In reading, 59% of sixth and seventh grade students scored at proficient levels on the state assessment in 2012. All students with informed consent were included in the study.

Measures.

Gates MacGinitie Reading Test. The Gates-MacGinitie Reading Test (GRMT) – Fourth Edition (MacGinitie et al., 2000) Level 6 is a group-administered standardized reading assessment with two subtests: Vocabulary and Comprehension. Vocabulary subtest measures ability to choose the word or phrase that means nearly the same as a target word. Comprehension subtest comprises short passages followed by multiple choice comprehension questions. Alternate, equated forms of Level 6 (Forms S and T) were administered at pre and posttest.

Evaluation of Academic Vocabulary (EAV). The version of EAV administered in Study One included the following foil types: unrelated, orthographic, and semantic. For words with one sense, there was one match and one of each foil type. For words with more than one sense, there were two matches, one orthographic foil, and one semantic foil. Items appeared to students as seen in following example:

<u>Rational</u>

He made a _________ decision.

The judge is a fair and _________ person.

Martin Luther King wanted ________ equality.

The ocean was calm and _________.

All sentences for a given word were presented simultaneously. As such, test takers were asked to judge the acceptability of the target word in each sentence. To calculate raw scores, students were awarded one point for each match correctly accepted and each foil correctly rejected.

For administration, students were given oral instructions and were presented with practice items for two familiar words. The administrator said, "your job is to read each sentence with the target word in the blank and to think about whether the sentence would make sense with that word in the blank. If the sentence would make sense, write the word in the blank. If the sentence would not make sense, put 'x' in the blank." Practice items illustrated all different item types tested. EAV was group-administered as a pre- and posttest. The posttest was administered following the 20-week intervention.

Procedures to analyze data. To provide support for concurrent validity, we first calculated Pearson's Correlation between the EAV and a standardized test of vocabulary, the GMRT pretest (MacGinitie, MacGinitie, Maria, & Dreyer, 2000), using SPSS 19 (IBM Corp., 2010). Correlations were examined with the GMRT vocabulary subtest instead of total reading score as this subtest is designed to measure knowledge of vocabulary and therefore is more closely aligned with the EAV. To address the question of whether EAV measures growth in knowledge of academic words, we employed one-way ANCOVA on the EAV gain score as a function of condition, adjusting for prior reading achievement using the GRMT total reading pretest score, using SPSS 19. Total reading score was used in this case as a general indicator of reading achievement, appropriate as a covariate. To address the question of rank order of difficulty of foil types we employed signal detection theory, computing d prime values for each foil type and comparing means.

Finally, we created a novel scoring method designed to use information from patterns of student responses to better estimate depth of word knowledge and to correct for guessing. This method uses Bayesian Belief Network (BBN) written in the R Statistical language Software package (R Core Team, 2012). To do this, we first constructed a network typology in which the probability of getting an item correct was influenced by both general word knowledge and the correctness on the next easiest item. This topology is shown in Figure 1.

Then for each sentence-type node, we estimated probabilities that a student could respond correctly given that the word was known, and that the neighboring node was answered correctly. The Bayesian network allowed us to estimate a relatively small number of probabilities in the understandable "forward" direction. The network then used Bayes' rule to calculate 'Likelihood of Mastery' scores in the reverse direction for each possible combination of responses.

Likelihood of Mastery scores were calculated for pre- and post-test for each student in Experiment 1 (using two-sense words). Note that "mastery" as it is used here reflects terminology for describing and interpreting values generated by the network, distinct from the idea of "mastery" of vocabulary knowledge. Absolute mastery of word knowledge was neither a goal of the intervention nor of our assessment. Group means at pre- and posttest were then calculated for "Likelihood of Mastery" scores and raw scores, and effect sizes between the two approaches were compared.

Results

The bivariate correlation between EAV pretest and GMRT Vocabulary revealed a strong, positive relationship (r = .782, p <.001), providing some evidence of concurrent validity. To address the question as to whether EAV measures growth in knowledge of general academic words over time, a one-way ANCOVA was performed on the EAV gain score as a function of condition, adjusting for prior reading achievement using the GMRT total reading pretest score. Intervention classes had significantly higher gain scores than control classes, $F(1,98) = 55$, $p < .001$ with a large effect size of eta squared = .320, suggesting that the EAV measures change in word knowledge over time.

To address the question about the rank order of difficulty of foil types, we used signal detection theory. D prime (*d*') values [i.e., *Z*(hit rate) - *Z*(false alarm rate)] were computed to assess students' ability to distinguish between correct responses and types of foils. A higher *d'* indicates that the student is better at distinguishing foils from matches. Mean *d'* values demonstrated that students had some difficulty distinguishing between hits and foils in general (*d*' = .624). More important were the differences in *d'* values by type of foil (Table 1). Students were most successful at distinguishing unrelated foils from hits (*d*' = .809), next most successful at distinguishing orthographic foils from hits (*d*' = .694), and least successful at distinguishing semantic foils from hits (*d*' = .492). Thus, on average, performance on the foils supported our hypothesized continuum of difficulty. However, many target words did not follow this pattern. In particular, orthographic foils were highly variable.

As noted above, a subsidiary aim of this work was to investigate methods for reducing the impact of guessing on estimates of depth of word knowledge. We suspected that guessing may inflate raw scores because students have a 50% chance of guessing the correct response for each sentence. However, guessing behavior may show up in telltale patterns of item correctness. Table 2 illustrates this by contrasting two patterns of performance. We see that Student A succeeded on the more difficult foils (orthographic and semantic) but missed the easier trials (both matches). In contrast, Student B succeeded on both matches but missed the two more difficult items: the orthographic and semantic foils. The raw correctness score, shown in column 6, is identical in both cases. However, human experts don't commonly judge Students A and B to have identical word knowledge. A human expert will typically infer that it is less likely to know the foils but not the matches than it is to know the matches but not the foils. Therefore, human experts commonly judge Student A more likely to have guessed than Student B. In general, we judge that correct answers to more difficult items are more likely to be guesses, if answers to the easier items are wrong.[2]

We initialized this network by hand-estimating probabilities that students would be correct on each sentence given that they knew (or did not know) the target word and did (or did not) get the next easiest item correct. This involved estimating probabilities that they would be incorrect if they knew the word (i.e., "slip" probabilities), as well as probabilities that they would be correct even if they didn't know the word (i.e., "guess" probabilities). We estimated these probabilities for each node in the network. Note that these probabilities are estimated in the forward "causal" direction. After initialization, the network then is able to use Bayes' Law to reason in the backward direction, updating the probability of word mastery based on the set of responses for that word.

An example of the "Likelihood of Mastery" scores produced by the resulting network is shown in column 7 of Table 2. Note that, similar to human intuition, the network assigns a higher probability of mastery to Student B. This result encourages us to think that Bayesian networks

[2] While we believe that application of Bayes' rule produced more precise estimates of word knowledge, it is also possible that a student could have arrived at either of these patterns as a result of guessing by applying a blanket "accept all" or "reject all" policy. This possibility was one of the motivations for Study Two in which we carried out cognitive interviews to understand the strategies and reasoning employed by students to make decisions on this assessment.

may be useful in correcting for student guessing. Note also that a more conventional scoring approach, which assigned more credit for correctly answering harder items, would get the ranking of these two students exactly backwards.

Given that our hand-trained network was able to mimic human judgments about guessing, we were next interested in how correcting for guessing would affect the relative gain scores for our intervention. To do this, we compared pre- and posttest results from Study One using the Likelihood of Mastery scores generated by our Bayesian Belief Network for all words tested on the EAV that have two senses. First, a one-way ANCOVA was performed on the raw EAV gain score as a function of condition, adjusting for prior reading achievement using the GMRT total reading pretest score (Table 3). Second, a one-way ANOVA was performed on the Likelihood of Mastery gain score as a function of condition. In both cases, treatment classes had significantly higher gain scores than control classes, with the Likelihood of Mastery results yielding a slightly larger effect size (eta squared = .358).

Discussion of Study One

In Study One, we predicted that the foils would follow a hierarchy of difficulty, yielding information about depth of word knowledge. We confirmed that foils performed as expected on average. However, there was substantial variation from word to word, especially for ranking of the orthographic foil. Given the instability of its performance and the related challenge of selecting orthographic foils of the same ilk (e.g., similarity to target word in terms of orthographic similarity, word frequency, etc.) we decided to eliminate this foil type and replace it with the syntax foil in Study Two.

We also experimented with the Bayesian Belief Network approach to reanalyzing our data to generate scores that are more precise and less susceptible to guessing behavior. In Study

Two, we used qualitative data to assess whether Likelihood Mastery Scores were more closely associated with expert judgment about students' word knowledge, when compared to the association between expert judgment and raw scores.

Study Two

Method

Participants. A follow-up, small-scale study was conducted with one class of seventh graders in an urban charter school where 96.4% of students are African-American and approximately 60% scored proficient or above in reading on the state assessment in 2012. Students were 19 seventh graders, all African-American, and all native English-speakers. Participants did not participate in the RAVE intervention. This class was selected because the teacher was willing to permit the research team to administer EAV and interview students. A subsample of students (4 boys and 4 girls) was selected by their teacher to participate in 30-minute interviews with research team members after completing the EAV. The teacher was asked to select two high, four middle, and two low-achieving students.

Measures.

EAV. Students completed an abridged and revised version of EAV. Format and presentation were exactly the same as in Study One, but this version contained item sets for 35 words for a total of 140 sentences. In this revised version, orthographic foils were eliminated and in their place syntax foils were tested. Thus, the EAV that was administered in Study Two included the following foil types: syntax, unrelated, and semantic. For words with one sense, there was one match and one of each foil type. For polysemous words, there were two matches, one syntax foil, and one semantic foil. Instructions and procedures for EAV administration were exactly the same as Study One.

***Interviews*.** A cognitive interview protocol was used to interview the subsample of students about their decision-making processes on EAV items for nine target words. Cognitive interviews, grounded in cognitive psychology, are designed to probe participants' thinking around a subset of assessment items to prompt verbalization of reasoning about response choices (DeMaio & Rothgeb, 1996). The goal of the interviews was to assess whether each foil type was operating as intended. For example, we wanted to know whether a student's rejection of a syntax foil was attributable to knowledge about a given word's syntactic role, or if a student's rejection of a syntax foil was simply rejection of an incorrect scenario. Six words had only one sense (*capable*, *traditional*, *definitive*, *coherent*, *concept*, and *imply*); three words had two senses (*expose*, *interval*, and *minimize*). Target words were selected to represent a range of word knowledge in the absence of instruction, based on pretest results from Study One. For each word, researchers began by asking questions about students' knowledge of the target word: "Do you know the word, [target word]?" "Have you heard this word before?" "What would you say it means?" The interviewer then asked questions about the decision-making process for each of the fill-in-the-blank items for the target word asking, "Does that one make sense?" and "Why" or "Why not?"

Selected students were interviewed after they completed the EAV. Interviews were 25-30 minutes, conducted in quiet spaces in the school. Interviews were audio recorded.

Procedures to analyze data. Descriptive results from EAV were examined in light of performance on each item type. Then all eight interviews were transcribed and analyzed by four members of the research team.

Results

We addressed the question about the rank order of difficulty of foil types using two approaches. First, we examined performance on the different foil types for all the target words. Items for words with one sense and three foils (syntax, unrelated, semantic) were examined separately from items for words with two senses and two foils (syntax and semantic). Frequencies of correct responses for each type of foil per target word are presented in Figures 2 and 3. Our hypothesis was that the degree of challenge for each foil type would fall along a continuum with *syntax* being easiest and testing surface-level word knowledge, *unrelated* falling in the middle, and *semantic* being the most difficult.

Figures 2 and 3 by and large support the hypothesized rank order of difficulty. Performance on unrelated foils was more erratic than anticipated. However, in Figure 2, we see that syntax foils were nearly always easiest; semantic foils were most difficult for most target words. Foils performed more stably for two-sense words. Figure 3 shows that syntax foils were easier than semantic foils for all but four words.

The second approach to addressing the question about rank order of difficulty of foil types was analysis of students' think-alouds from the subsample of students interviewed. Students' explanations for why they believed the target word fit (or did not fit) different item types lent support to the claim that EAV item types tap different levels of word knowledge. Excerpts from interview data are presented below to illustrate how students explained their thinking about the three foil types and the match items.

Syntax foils. Syntax foils were designed to distinguish shallow levels of word knowledge, such that students who have virtually no memory traces of a word's use or meaning will not know to reject the syntax foil. The following example of Student 1's reasoning about the syntax foil suggested lack of familiarity with the target word. This student's responses to all interview questions for *interval* further indicated that he did not have associations with or knowledge of the target word.

We ______ many small animals on our hike through the woods. (interval)

Interviewer: Would that work do you think?

Student 1: No, because... to me it doesn't make sense. We inter...interval.

Interviewer: We interval.

Student 1: We interval many animals... on our... hike. Well... actually, that does make sense because like... we view... So, yeah it does.

Interviewer: Okay, so ... would go there? Okay. Alright.

As an aside, it is important to note that while in this case the student's response accurately conveyed his level of word knowledge (i.e., he incorrectly decided that the word did fit in foil), guessing could have easily enabled him to correctly reject the foil.

In contrast, in the following example, Student 2 states that he is familiar with the meaning of the target word, *definitive,* although he based that on an association between *definitive* and *definition.* He demonstrates surface-level knowledge when he correctly rejects the syntax foil, and is able to articulate why the foil cannot fit ("an object goes there"). His reasoning about the other sentence types, however, suggests that he is far from clear about the word's meaning elements.

My friend had a ______ when she got home from school. (definitive)

Interviewer: Definitive. Is that a new word for you... or that's a familiar word?

Student 2: It's familiar.

Interviewer: It is familiar. Okay. Do you know what definitive means?

Student 2: I have an idea.

Interviewer:: Oh, good. What's your idea?

Student 2: It has something to do with definition... .

Interviewer: Okay, so let's look at the first sentence. My friend had a definitive when she got home from school. What do you think about that one?

Student 2: I don't think that was right. Cause I think there should be... um... an object right there.

Interviewer: Okay. Okay. Very good. So you were thinking about the part of speech, too.

Unrelated foils. Unrelated foils were designed to capture shallow word knowledge about the target word. These foils may seem acceptable to a student who has only a sense of the word's syntactic role. Student 2 above, who correctly rejected the syntax foil for *definitive,* also correctly rejected the unrelated foil.

Sarah was so ______ when she couldn't find her homework. (definitive)

Interviewer: What did you think about that one?

Student 2: No.

Interviewer: Okay, why?

Student 2: Because definitive is not a feeling.

Interviewer: Okay

Student 2: And... if... and Sarah could have been sad or happy or something instead of definitive.

In Student 3's interview about the item for *coherent,* the student knew to reject the syntax foil but her knowledge of the semantic properties of *coherent* was unstable as seen in the excerpt below in discussing the unrelated foil. She seems to have an association with *incoherent*, but also believes that *coherent* could have multiple unrelated meanings.

Lucy felt so ______ in her new sweater. (coherent)

Student 3: Um... coherent means... like um... it's not straight on... it's um... like... your mind's in different directions....

Interviewer: Okay, alright. Good.

Student 3: Or like um... Or like some things are not going in the right direction. It's going the opposite way.

Interviewer: Okay... Lucy felt so coherent in her heavy wool sweater.

Student 3: Coherent... also um... it also deals with weight... and things like that.

Interviewer: Huh huh... So would it fit there?

Student 3: Yes.

*Semantic foil*s. Finally, *semantic foils* were designed to distinguish students with some word knowledge from those with stable and precise word knowledge. Students who have enough familiarity with the word's meaning or use to know the kinds of situations in which the word is used should correctly reject the foil. Students who have established some semantic associations and collocational associations but whose knowledge is still somewhat fragile are likely to incorrectly accept the foil. Note that in the excerpt below, Student 4 has knowledge about the word, *interval,* as seen in her correct rejection of the syntax foil and acceptance of the two match sentences. But some fragility of that knowledge is revealed in the way she incorrectly accepts the semantic foil:

I try to do my regular ______ between six and seven every morning. (interval)

Interviewer: So, what would you say the word interval means?

Student 4: Distance?

Interviewer: Alright … can you tell me about how you made decisions about whether or not to write in the word or an X for each of those? Okay, so the first one was, “We interval many small animals on our hike through the woods.” What did you decide for that one?

Student 4: X.

Interviewer: Okay, and can you say why you made that decision?

Student 4: Cause it didn't have anything to do with distance.

Interviewer: Doesn't have anything to do with distance... “There is an interval of one foot between desks in our classroom.” Did you write interval in that one?

Student 4: Yes.

Interviewer: How did you make that decision?

Student 4: Cause it had something to do with the distance.

Interviewer: Alrighty. “She started surfing again after an interval of three months.”

Student 4: Yes.

Interviewer: And why?

Student 4: Because it has something to do with distance.

Interviewer: “I try to do my regular interval between six and seven every morning.” How about that one?

Student 4: Yes.

Interviewer: And why did you write the word in there?

Student 4: Cause it has something to do time periods.

Interviewer: Cause it has to do with time periods. Okay. Okay. Super. Was that last one a hard decision for you or were you certain? Did you think... yeah, that definitely fits in there.

Student 4: Yeah.

Student 4 clearly has a strong, developing understanding of the meaning of *interval*. However, while she correctly holds strong associations between interval and the idea of distance and time periods, her understanding of interval is not precise or clear enough to enable the recognition that a single interval is not something you "do" even if it's during a certain time period and even if it appears with its collocation "regular." The pattern of responses seems to capture the student's level of word knowledge accurately. The student's correct rejection of the easier foils and correct acceptance of the hits all indicate some knowledge of the word's meaning elements. However, the student's failure to reject the semantic foil captures that the student's representation is not yet well established, stable, and precise.

In contrast, Student 5's rejection of the semantic foil for *minimize* seems to be a valid indicator of precise knowledge of word meaning, including understanding of constraints of word use even when the semantic foil includes a collocational association (i.e., "risk").

We saw signs along the trail to ______ *hikers about the risk of poisonous snakes.* (minimize)

Student 4: At first I thought minimize would go there, but then I realized that I don't think it can because "We saw signs along the trail to minimize hikers…" to minimize hikers about the risk. So pretty much what would go there is explain. But not minimize. At first I thought it would be minimize because like, minimize the risk of poisonous things.

Discussion of Study Two

Study Two provided additional evidence that semantic foils were most difficult and that syntax foils were easiest. Interview results allowed us to go further than confirming rankings of foil types, by providing insight into the reasoning students used to make their decisions about the sentences. These decision-making processes suggested that our rationale behind creation of the various foil types was accurate. Rejection of syntax foils often seemed to rest on understanding syntactic role, and general associations to a word's meaning allowed students to reject unrelated foils, for example, knowing that *definitive* was "not a feeling." Semantic foils seemed to draw attention to the associations around which those foils were designed, causing students to accept the foils if their knowledge was limited, or enabling them to reject the foils if their knowledge allowed them to reason around that association, as seen in the *minimize* example when the student recognized that it would be the risk, not the hikers, that would be minimized.

General Discussion

The new assessment reported in this study, Evaluation of Academic Vocabulary (EAV), is designed for use with native English speaking adolescent learners for testing treatment effects of literacy interventions. The EAV was designed to capture depth of knowledge of general academic words by systematically combining "match" and "foil" items that tap different dimensions of word knowledge. General academic words tend to be abstract and often carry multiple meanings, rendering assessment of word knowledge a challenge. Yet these words frequently carry important meaning in academic texts; we can learn valuable information by measuring the impact of intervention on adolescents' depth of knowledge of academic words.

EAV's design was informed by a recent surge of efforts to evaluate depth of vocabulary knowledge. The vast majority of these studies were designed for adult learners who are L2 learners, yet the theoretical underpinnings of many of these assessments and their operationalized definitions of word knowledge proved informative for our work.

Specifically, Nation's (2001/2013) conceptualization of the multifaceted aspects of word meaning, and related work by Read (1989, 1993, 1998), Chui (2006), and Qian (1999; 2002; 2004) point to dimensions of word knowledge relevant to the population and context of the present work. Largely convergent across these studies, syntactic knowledge, meaning, collocational properties, and (to a lesser degree) orthographic information, were included as dimensions of word knowledge relevant to understanding and measuring depth. These dimensions informed development of item types on the EAV. Further, both Schmitt (1998) and Crossley and colleagues (2010) point to the importance of understanding polysemy as a critical dimension of depth. Building on this scholarship, EAV was designed to test students' development of general academic words' multiple meanings.

Assessments employed with adolescent learners typically fall short when it comes to measuring depth of word knowledge (McKeown et al., 2017). Recent assessment projects by Scott and colleagues (Scott et al., 2015) and Deane and colleagues (Deane et al., 2014) have made great strides in this direction. However, neither focuses explicitly on general academic words that students will encounter across disciplines in secondary and post-secondary schooling, Moreover, those assessments present students with multiple questions to assess dimensional knowledge of each target word. EAV, in contrast, focuses on academic words exclusively, including knowledge of multiple senses, and it assesses multiple aspects of word knowledge through presentation of only four fill-in-the-blank sentences per target word. Our data showed that students in an intervention group exhibited growth in target word knowledge, whereas students in a control group did not. EAV measures vocabulary and is efficient in ways valued by school district-based educators and administrators; for that reason, it may be a palatable choice for testing treatment effects of literacy interventions in schools. Use of such an assessment in schools also might promote educators' awareness of vocabulary as a multidimensional construct and of vocabulary learning as an incremental process. This could in turn engender more effective instructional strategies for vocabulary and more realistic expectations for students' learning.

In short, the EAV is at once rooted in the scholarship of existing depth measures and at the same time offers a new design and focus well-suited to the context of interventions with school-aged learners. The results of the two studies reported here provide preliminary evidence that: 1) EAV taps dimensions of knowledge of general academic words important for comprehension of academic texts; and 2) EAV is sensitive to treatment effects of a 20-week academic vocabulary intervention. Lack of precision in EAV scores continues to be a challenge given the potential for guessing. Bayesian Belief Networks might offer a promising approach to reduce the impact of guessing behavior.

The mixed methods design we employed, which included both quantitative analyses and interviewing students, was a useful approach for collecting information about the validity of the assessment. The current study sought to investigate whether performance on foil types indicates levels of word knowledge. We hypothesized that syntax items would capture a minimal level of knowledge about a lexical item, as reflected in Chui's findings (2006). We hypothesized that semantic foils would be most challenging and would capture more precise understanding of word meaning and constraints of word use. As such, semantic foils were designed to include collocations and common associations to words. Awareness of associations has been included in measures of word knowledge depth (e.g., Read,1998; Qian & Schedl, 2004). In our assessment, we used association as a lure to test whether students could reject a context in which a common association was used, but that did not reflect accurate use of the target word.

By combining methodological approaches in our investigation, we confirmed that foil types generally reflected the dimensions of word knowledge we intended to measure, and to some degree reflected the continuum of difficulty we anticipated. In Study One we compared means on D prime values of different foil types and found that, on average, these fell into the expected order, with semantic foils as the most challenging. In Study Two, we conducted cognitive interviews and confirmed that by and large students' reasoning as they made decisions about whether to accept or reject foils reflected thinking about dimensions of word knowledge we meant to assess.

Limitations and Future Directions

A notable limitation in the studies reported is that while item types capture a continuum of difficulty on average, for each item type on the EAV, the degree of difficulty varies from word to word. In future work, we intend to pilot many sentence candidates for each category so that we can collect information on the potential challenge of various individual sentences. This should provide multiple options for the final version of the assessment, thus allowing us to restrict the range of difficulty of sentences within categories.

A longitudinal study in which the EAV would be administered prior to intervention, at post-test, and at delayed post-test points would enable us to investigate how EAV can be used to measure incremental learning. In the present study, we were not able to administer a delayed post-test because of the district's goal of minimizing time students spend taking assessments.

Future work might address whether the innovative format we have developed could be used to assess word types such as polysemous words holding discipline-specific and everyday meanings. Another possibility would be the application of this format to testing function words such as connectives that are abstract, can only be understood in context, but are often critical for comprehension (Crosson & Lesaux, 2013).

Future work in this research program will also entail extending application of EAV to other populations beyond monolingual English-speaking adolescents. Validity of this measure for English Learners (ELs) is a particularly intriguing question. Given the prevalence of reading difficulties in this population (Kieffer, 2010; Mancilla-Martinez & Lesaux, 2010) and, in turn, the importance of assessing the effectiveness of academic vocabulary interventions with this population, we see a pressing need to move forward with this research agenda. However, as the EAV is tested with and adapted for EL adolescents, it will be necessary to confront challenges related to limitations in syntactic knowledge (Lipka & Siegel, 2007) and background knowledge (Garcia, 1991; Jiménez & Garcia, 1996).

Conclusion

Vocabulary knowledge is a multidimensional construct, and understanding of multiple aspects of word knowledge is needed to promote higher-level literacy activities such as reading comprehension (Baumann, Kame'enui, & Ash, 2003; McKeown et al., 1987; Pearson, Hiebert, & Kamil, 2007; Stahl & Fairbanks, 1986). A learner's constellation of multiple aspects of word knowledge connotes depth of knowledge. Measuring depth of word knowledge gained by vocabulary instruction is key to understanding the effectiveness of interventions. We designed EAV to measure dimensions that researchers have identified as key to word knowledge. EAV appears to be a useful instrument for assessing depth of knowledge among monolingual English-speaking populations and, in particular, for assessing treatment effects related to intervention work.

References

Beck, I. L., McKeown, M. G., & Omanson, R. C. (1987). The effects and uses of diverse vocabulary instructional techniques. In M. G. McKeown & M. E. Curtis (Eds.), *The nature of vocabulary acquisition* (pp. 147-163). Hillsdale, NJ, US: Lawrence Erlbaum Associates, Inc.

Anderson, R., & Nagy, W. (1991). Word meanings. In R. Barr, M. Kamil, P. Mosenthal, & P.D. Pearson, (Eds.), *Handbook of reading research*, Vol. 2, (pp. 690–724). New York: Longman.

Baumann, J.F., Kame'enui, E.J., & Ash, G.E. (2003). Research on vocabulary instruction: Voltaire redux. In J. Flood, D. Lapp, J.R. Squire, & J.M. Jensen, (Eds.), *Handbook of research on teaching the English language arts* (pp. 752-785). Mahwah NJ: Erlbaum & Associates.

Bolger, D.A., Balass, M., Landen, E., & Perfetti, C.A. (2008). Context variation and definitions in learning the meanings of words: An instance-based learning approach. *Discourse Processes, 45*, 122-159. DOI: 10.1080/01638530701792826

Browne, C., Culligan, B. & Phillips, J. (2013). *A new general service list*. Retrieved from http://www.newgeneralservicelist.org [Accessed on Oct 21 2017]

Chui, A.S.Y. (2006). A study of the English vocabulary knowledge of university students in Hong Kong. *Asian Journal of English Language Teaching, 16*, 1-23.

Creswell, J.W. (2014). *Research design: Qualitative, quantitative, and mixed methods approaches, 4th Edition*. Los Angeles: Sage Publications.

Coxhead, A. (2000). A new academic word list. *TESOL Quarterly, 34*(2), 213-238. DOI: 10.2307/3587951

Crossley, S., Salsbury, T., & McNamara, D. (2010). The development of polysemy and frequency use in English second language speakers. *Language Learning, 60*, 573-605. DOI: 10.1111/j.1467-9922.2010.00568.x

Crosson, A.C., McKeown, M.G., Beck, I.B., & Ward, A. (2012, July). Developing an assessment to measure depth of knowledge of academic vocabulary. Interactive paper presented at the 19th Annual Meeting of the Society for the Scientific Studies of Reading, Montreal, Quebec.

Crosson, A.C. & Lesaux, N.K. (2013). Pinpointing the challenging aspects of academic language: Does knowledge of connectives play a special role in the reading comprehension of English language learners and English-only students? *Journal of Research in Reading*, *36*, 241-260.

Crosson, A.C. & McKeown, M.G. (2016). How effectively do middle school learners use roots to infer the meaning of unfamiliar words? *Cognition and Instruction*, *34*, 148-171.

Deane, P., Lawless, R., Li, C., Sabatini, J.S., Bejar, I.I., & O'Reilly, T. (2014). *Creating Vocabulary Item Types That Measure Students' Depth of Semantic Knowledge*. ETS Research Report no. RR-14- 02. Published online at http://dx.doi.org/10.1002/ets2.12001

DeMaio, T. J., & Rothgeb, J. M. (1996). Cognitive interviewing techniques: In the lab and in the field. In N. Schwarz & S. Sudman (Eds.), *Answering questions: Methodology for determining cognitive and communicative processes in survey research* (pp. 177-195). San Francisco: Jossey-Bass.

Eddington, C.M. & Tokowicz, N. (2015). How meaning similarity influences ambiguous word processing: the current state of the literature. *Psychological Bulletin Review, 22*(1), 13-37. DOI: 10.3758/s13423-014-0665-7

Educational Testing Service. (2011). *GRE General Test*. Princeton, NJ.

Elleman, A.M., Lindo, E.J., Morphy, P. & Compton, D.L. (2009). The impact of vocabulary instruction on passage-level comprehension of school-age children: A meta-analysis'. *Journal of Research on Educational Effectiveness, 2*(1), 1-44.

Gardner, D. & Davies, M. (2013). A new academic vocabulary list. *Applied Linguistics, 35*, 305-327. DOI: 10.1093/applin/amt015

Garcia, G.E. (1991). Factors influencing the reading test performance of Spanish-speaking Hispanic children. *Reading Research Quarterly, 26*, 371-392. DOI: 10.2307/747894

Kroll, J. F. , Dussias, P. E., Bice, K., & Perrotti, L. (2015). Bilingualism, mind, and brain. In M. Liberman & B. H. Partee (Eds.), *Annual review of linguistics*, *1*, 377-394.

IBM Corp. (2010). *IBM SPSS Statistics for Windows, Version 19.0*. Armonk, NY: IBM Corp.

Jiménez, R.T. & Garcia. G.E. (1996). The reading strategies of bilingual Latina/o students who are successful English readers: Opportunities and obstacles. *Reading Research Quarterly, 31*(1), 90-112. DOI: 10.1598/RRQ.31.1.5

Kieffer, M.J. (2010). Socioeconomic status, English proficiency, and late-emerging reading difficulties. *Educational Researcher, 39*, 484-486. DOI: 10.3102/0013189X10378400

Li, M., & Kirby, J.R. (2012). Breadth and depth of vocabulary knowledge in second language reading. *Language Learning, 4,* 79-103.

Lipka, O. & Siegel, L.S. (2007). The development of reading skills in children with English as a second language. *Scientific Studies of Reading, 11*(2), 105-131. DOI: 10.1080/10888430709336555

Lockett, J.N. & Shore, W.J. (2003). A narwal is an animal: Partial word knowledge biases adults' decisions. *Journal of Psycholinguistic Research, 32*, 477-496.

MacGinitie, W.H., MacGinitie, R.K., Maria, K., & Dreyer, L.G. (2000). *Gates-MacGinitie Reading Tests* (4th Edition). Rolling Meadows, IL: Riverside Publishing.

Mancilla-Martinez, J., & Lesaux, N.K. (2010). Predictors of reading comprehension for struggling readers: The case of Spanish-speaking language minority learners. *Journal of Educational Psychology, 102,* 701-711.

McKeown, M.G., Crosson, A.C., Beck, I.B. & Moore, D.W. (2018). Word knowledge and comprehension effects of an academic vocabulary intervention for middle school students. *American Educational Research Journal, 55*, 572-616.

McKeown, M. G., Deane, P. D., Scott, J. A., Krovetz, R., & Lawless, R. R. (2017). *Vocabulary assessment to support instruction: Building rich word-learning experiences.* Guilford Press: New York, NY.

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

Meara, P. & Wolter, B. (2004). V_links: Beyond vocabulary breadth. In D. Albrechtsen, K. Haastrup, & B. Henriksen (Eds.), *Angles on the English speaking world,* 4 (pp. 85-96). Copenhagen: Museum Tusculanum Press.

Nagy, W.E. & Scott, J.A. (2000). Vocabulary processes. In M.L. Kamil, P.B. Mosenthal, P. David Pearson, & R. Barr, (Eds.), *Handbook of reading research* (Vol. III, pp. 69-284). Mahwah, NJ: Erlbaum.

Miller, G.A. (1978). Semantic relations among words. In M. Halle, J. Bresnan, & G.A. Miller, (Eds.), *Linguistic theory and psychological reality* (pp. 61-118). Cambridge, MA: MIT Press.

Nation, I.S.P. (2001). *Learning vocabulary in another language*. Cambridge, UK: Cambridge University Press.

Nation, I.S.P. (2013). *Learning vocabulary in another language* (2nd Edition). Cambridge, UK: Cambridge University Press.

National Assessment Governing Board (2008). *Reading framework for the 2009 National Assessment of Educational Progress*. Developed for the National Assessment Governing Board under contract number ED–02–R–0007 by the American Institutes for Research. U.S. Department of Education, Washington, D.C.

Nagy, W.E., & Scott, J.A. (2000). Vocabulary processes. In M.L. Kamil, P.B. Mosenthal, P. David Pearson, & R. Barr, (Eds.), *Handbook of reading research, Vol. III*. (pp. 69-284). Mahwah, NJ: Erlbaum.

Nagy, W.E., & Townsend, D. (2012). Words as tools: Learning academic vocabulary as language acquisition. *Reading Research Quarterly*, *47*(1), 91–108.

Nunberg, G. (1979). The non-uniqueness of semantic solutions: Polysemy. *Linguistics and Philosophy, 3*(2), 143-184.

Partnership for Assessment of Readiness for College and Careers (2016). PAARC Assessment. New York: Pearson.

Pearson, P.D., Hiebert, E.H., & Kamil, M.L. (2007). Vocabulary assessment: What we know and what we need to learn. *Reading Research Quarterly*, *42*(2), 282-296. DOI: 10.1598/RRQ.42.2.4

Pearson, P.D., Hiebert, E.H., & Kamil, M.L. (2012). Vocabulary assessment. Making do with what we have while we create the tools we need. In J. Baumann and E. Kame'enui, (Eds.), *Vocabulary Instruction: Research to Practice* (2nd Ed.). (pp.231-255) New York, NY: Guilford Press.

Perfetti, C.A. (2007). Reading Ability: Lexical Quality to Comprehension. *Scientific Studies of Reading, 11*(4), 357-383. DOI:10.1080/10888430701530730

Perfetti, C., & Stafura, J. (2014). Word knowledge in a theory of reading comprehension. *Scientific Studies of Reading*, *18*, 22-37. doi:10.1080/10888438.2013.827687

Pethö, G. (2001). What is polysemy? – A survey of current research and results. In: E.T. Németh & K. Bibok, (Eds.), *Pragmatics and flexibility of word meaning* (pp.175-224). Amsterdam: Elsevier.

Princeton University (2010). *WordNet: A lexical database for English*. New Jersey: Princeton. <http://wordnet.princeton.edu>

Qian, D.D. (1999). Assessing the roles of depth and breadth of vocabulary knowledge in reading comprehension. *Canadian Modern Language Review 56*, 282-308.

Qian, D. D. (2002). Investigating the relationship between vocabulary knowledge and academic reading performance: An assessment perspective. *Language Learning. 52*, 513-536.

Qian, D.D., & Schedl, M. (2004). Evaluation of an in-depth vocabulary knowledge measure for assessing reading performance. *Language Testing*, *21*(1), 28-52.
DOI: 10.1191/0265532204lt273oa

R Core Team (2012). *A language and environment for statistical computing*. R Foundation for Statistical Computing. Vienna, Austria ISBN 3-900051-07-0, URL http://www.R-project.org/

Read, J. (1989). *Towards a deeper assessment of vocabulary knowledge*. ERIC Document Reproduction Service ED 654 321. Washington, DC: ERIC Clearinghouse on Languages and Linguistics.

Read, J. (1993). The development of a new measure of L2 vocabulary knowledge. *Language Testing 10*(3), 355-371.

Read, J. (1998). Validating a test to measure depth of vocabulary knowledge. In A. Kunnan (Ed.), *Validation in language assessment* (pp. 41-60). Mahwah, NJ: Erlbaum.

Read, J. (2000). *Assessing vocabulary*. Cambridge: Cambridge University Press.

Schmitt, N. (1998). Tracking the incidental acquisition of second language vocabulary: A longitudinal study. *Language Learning, 48,* 281-317.

Schmitt, N. (2014). Conceptual review article. Size and depth of vocabulary knowledge: What the research shows. *Language Learning, 64,* 913-951. DOI: 10.1111/lang.12077

Schwanenflugel, P.J., Harnishfeger, K.K., & Stowe, R.W. (1988). Context availability and lexical decisions for abstract and concrete words. *Journal of Memory and Language*, *27*(5), 499–520. DOI:10.1016/0749-596X(88)90022-8

Schwanenflugel, P.J., & Shoben, E.J. (1983). Differential context effects in the comprehension of abstract and concrete verbal materials. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 9*, 82-102 DOI: 10.1037/0278-7393.9.1.82

Scott, J.A., Flinspach, S.L., Vevea, J.L., & Castaneda, R. (2015). Vocabulary Knowledge as a Multidimensional Concept: A Six Factor Model. Presented at the annual meeting of the Society for Scientific Study of Reading. Hapuna Beach, HI.

Scott, J., Lubliner, S. & Hiebert, E.H. (2006). Constructs underlying word selection and assessments tasks in the archival research on vocabulary instruction. In C.M. Fairbanks,

J. Worthy, B. Maloch, J. Hoffman, & D. Schallert, (Eds.), *National Reading Conference yearbook*. Oak Creek, WI: National Reading Conference.

Sireci, S.G. (2012). *Smarter Balanced Assessment Consortium: Comprehensive research agenda*. Report Prepared for the Smarter Balanced Assessment Consortium. Retrieved from https://www.smarterbalanced.org/wp-content/uploads/2016/05/Comprehensive_Research_Agenda.pdf

Srinvasan, M. & Snedeker, J. (2011). Judging a book by its cover and contents: The representation of polysemous and homophonous meanings in four-year-old children. *Cognitive Psychology, 62,* 245-272. DOI:10.1016/j.cogpsych.2011.03.002

Stahl, S.A., & Fairbanks, M.M. (1986). The effects of vocabulary instruction: A model-based meta-analysis. *Review of Educational Research*, *56*, 72–110. DOI: 10.3102/00346543056001072

Vermeer, A. (2001). Breadth and depth of vocabulary in relation to L1/L2 acquisition and frequency of input. *Applied Psycholinguistics, 22*, 217-234. DOI: 10.1017/S0142716401002041

Table 1

D prime Values Comparing Students' Sensitivity to Different Types of Foils

Foil type	d prime
Unrelated foils	.809
Orthographic foils	.694
Semantic foils	.492

Table 2

Example Probabilities that Students Know a Target Word with Two Senses Based on Patterns of EAV Performance

Sample student	Match one	Match two	Orthographic foil	Semantic foil	Raw score	Likelihood of mastery
Student A	0	0	1	1	2	.28
Student B	1	1	0	0	2	.64

Table 3
Comparison of Scores Using Raw Scores vs Bayesian Probabilities

	Raw Scores (std dev)				Likelihood of Mastery Scores (std dev)			
Group	Pre	Post	Gain	Effect size	Pre	Post	Gain	Effect size
Control (n=43)	35.29 (5.21)	37.20 (6.13)	1.85 (3.00)	.320	32.49 (6.22)	34.51 (7.27)	1.87 (4.29)	.358
Treatment (n=61)	34.71 (4.38)	41.46 (6.05)	6.50 (3.84)		31.10 (5.29)	39.52 (7.13)	8.23 (4.59)	

Note. effect sizes reported are eta squared.

Figure 1

Proposed Network Typology for Bayesian Belief Network

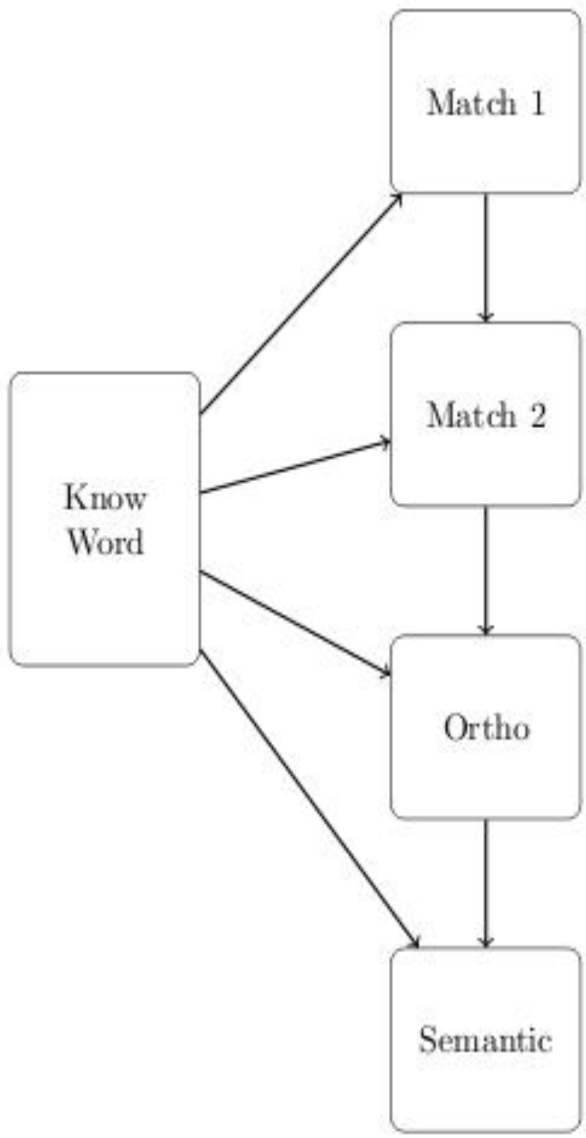

Figure 2

Frequencies of Correct Responses to Different Foil Types for Cloe Target Words with One Sense (n=19 seventh grade students)

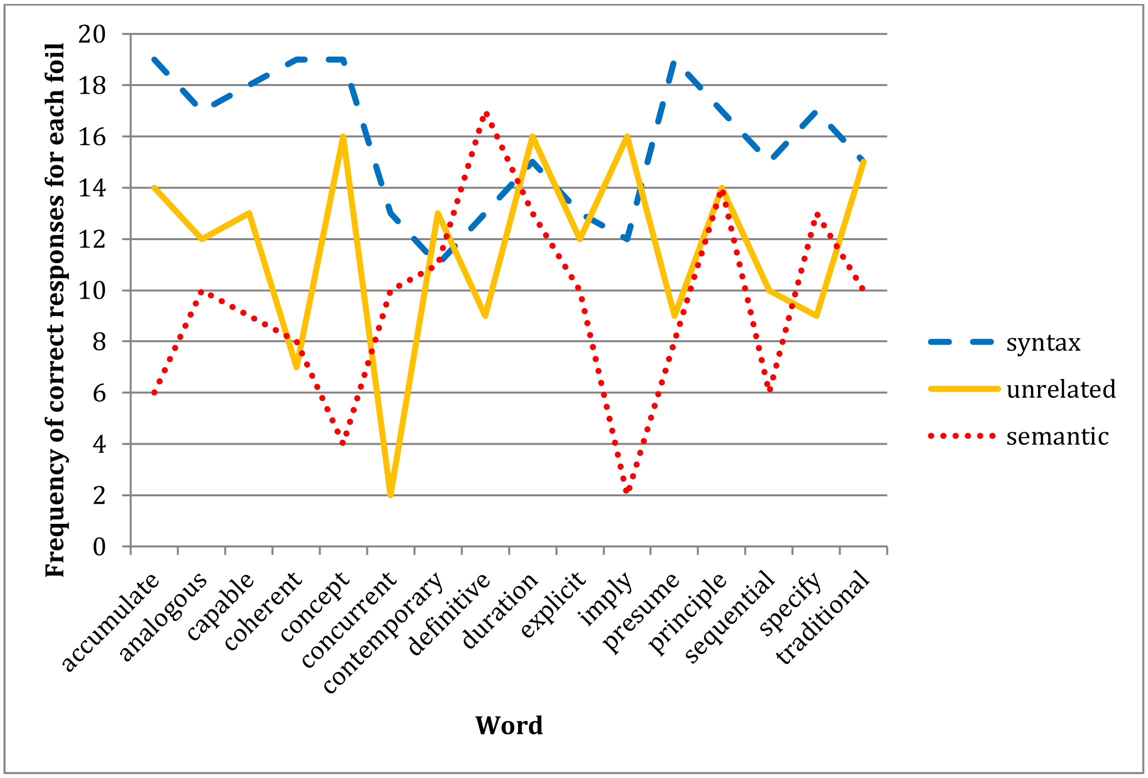

Figure 3

Frequencies of Correct Responses to Different Foil Types for Cloe Target Words with Two Senses (n=19 seventh grade students)