Article

Journal of Literacy Research
2015, Vol. 47(2) 153–185

DOI: 10.1177/1086296X15615363
jlr.sagepub.com
SAGE

Factors That Influence the Difficulty of Science Words

Gina N. Cervetti[1], Elfrieda H. Hiebert[2], P. David Pearson[3], and Nicola A. McClung[4]

Abstract

This study examines, within the domain of science, the characteristics of words that predict word knowledge and word learning. The authors identified a set of word characteristics—length, part of speech, polysemy, frequency, morphological frequency, domain specificity, and concreteness—that, based on earlier research, were prime candidates to explain variation in word knowledge and word learning. The outcome measures were the pretest (evidence of word knowledge) and posttest (evidence of word learning) vocabulary scores of second- through fourth-grade students who participated in one of several studies designed to evaluate the efficacy of science units that were part of a multiyear research and development program for an integrated science and literacy curriculum. The authors first examined individual predictors and then built multivariate models from the individually significant predictors of pretest score. Three characteristics were predictive of word knowledge (pretest score) at two or more grade levels; frequency, polysemy, and length predicted word difficulty independent of instruction. These three characteristics accounted for 39% of the variance in third graders' pretest scores. Polysemy and frequency alone accounted for 34% of the variance at second grade and 23% at fourth grade. In addition, frequency and polysemy explained students' vocabulary growth scores (posttest controlling for pretest) over the course of instruction at two of three grade levels. Understanding which characteristics of words are related to difficulty within the domain of science may provide a principled basis both for the selection of words for instruction and for differentiating instruction across categories of science words.

Keywords

vocabulary, science instruction, elementary

[1]University of Michigan, Ann Arbor, USA
[2]TextProject, Santa Cruz, CA, USA
[3]University of California, Berkeley, USA
[4]University of San Francisco, CA, USA

Corresponding Author:
Gina N. Cervetti, School of Education, University of Michigan, 610 E. University Avenue, Rm 4204D, Ann Arbor, MI 48109, USA.
Email: cervetti@umich.edu

This work begins with the assumption that one correlate of gaining new knowledge, especially in disciplines like science and history, is the acquisition of vocabulary—new words that we use to "name" our new knowledge in ways that are more precise and useful than can be accomplished with common words. It further assumes that not all words, and certainly not the ideas they stand for, are created equal—that, instead, some are harder to know and/or harder to learn than others. Any theory about word learning must take these differences into account. In this study, we set out to learn whether characteristics of words relate to the likelihood that they will be known by students and to the ease with which students will learn them. Specifically, we asked two related questions: (a) Within the domain of science, what word characteristics predict word difficulty before instruction? and (b) Do these characteristics also predict the likelihood that students will learn the words in the course of science instruction? Our focus, then, was on the features of words that predict both word knowledge and word learning.

To establish which word characteristics are the most salient, we drew on a conceptual framework identified by Nagy and Hiebert (2011) in a review of research focused on selecting words for instruction. We begin by laying out this conceptual frame of vocabulary knowledge, especially as it relates to the broader issue of acquiring knowledge and inquiry skill in science domains such as light, gravity, soil, and mixtures. Then, we turn to a description of the research context, the range of studies in which we gathered the data on word knowledge (operationalized as what students know without benefit of specific instruction in a domain like gravity) and word learning (operationalized as pre-post gains in word knowledge as a result of instruction). Finally, we examine the empirical evidence we collected from these studies to assess the power of word-level variables in predicting knowledge and/or learning.

Although this study focuses on word characteristics and word learning in science, we do not intend to suggest that word learning alone is a meaningful goal of science instruction. Instead, we view word knowledge and conceptual understanding as inextricably interwoven facets of word learning that are learned concurrently and iteratively. Existing research with young children has documented a bidirectional relationship between word learning and concept development (e.g., Kaefer & Neuman, 2013; Kemler Nelson, O'Neill, & Asher, 2008), and theories of conceptual development—most notably that of Vygotsky (1987)—have relied on the understanding that word knowledge and conceptual knowledge are codeveloped. Although Vygotsky cautions that word learning can be "mindless" and "empty" when it involves simply mapping a word onto an object, he also considered words integral to concept formation, because they focus attention and aid analysis (p. 170). In the instructional units in which students engaged as part of the studies included in the current research, words were treated as important but partial dimensions of conceptual understanding. Students were engaged in reading, writing, talk, and inquiries as a way of supporting them in codeveloping their conceptual understandings and the language (words and conventions of talk) that could facilitate both these understandings and their abilities to communicate them.

The Theoretical Context

The research on vocabulary acquisition offers well-tested ideas about general characteristics of instruction that result in vocabulary learning (National Institute of Child Health and Human Development [NICHD], 2000). Our understandings about what makes particular words easier or harder to learn, and how that knowledge might affect the selection and instruction of words, are less well developed, particularly in content-area instruction. The most prominent criteria for guiding teachers in selecting words for instruction emphasize (a) the usefulness or centrality of words in relation to particular texts and lessons and (b) the familiarity of the underlying concepts represented by words. These criteria are prominent in Beck, McKeown, and Kucan's (2013) three tiers: Tier 1 words—highly frequent, everyday labels for commonly known ideas (e.g., *water, person*); Tier 2 words—relatively novel words (e.g., *glamorous*) for common ideas (e.g., *pretty*) used by mature language users; and Tier 3 words—words used in specific disciplinary contexts (e.g., *transpiration, convection, gubernatorial*). Beck et al. recommend that teachers focus instruction on Tier 2 words, arguing that access to these words provides students with a kind of semiotic capital (i.e., knowledge of academically sophisticated words). Such a technique for selecting vocabulary words is likely to ensure that students accrue a sizable vocabulary of words typically found in literary texts, but a tiered approach that privileges instruction of Tier 2 words may not adequately address the content-specific words in the informational texts that are to be a central part of English Language Arts (ELA) instruction within the Common Core State Standards (CCSS; National Governors Association [NGA] and Council of Chief State School Officers [CCSSO], 2010) and that are an important feature of content-area learning. The analysis of Tier 2 and Tier 3 words from the text *Volcanoes* (Seymour, 2006)—one of the exemplar texts identified by CCSS writers as the type of science text to be taught in ELA instruction—illustrates the treatment of concept-critical words in informational texts within a tiered approach. With one exception (*eruption*), CCSS writers identified all concept-critical words as Tier 3 words—*crust, lava, molten, mantle, magma*, and *volcanoes* (NGA & CCSSO, 2010, Appendix A).

With an increased emphasis on informational text, the criteria for selecting vocabulary need to be reexamined. In a text such as *Volcanoes*, Tier 3 words such as *volcanoes* and *lava* are essential to comprehension of the text. Furthermore, these words are not exclusive to science informational texts; they can also be expected to be part of narratives, including in popular texts (e.g., *Vacation under the Volcano* in the Magic Tree House series). The wide-scale digitization of texts over the past decade has meant substantial growth in and insights from the field of corpora linguistics (Aijmer & Altenberg, 2013; Gries & Newman, 2013). The databases and measures resulting from this work have potential for increasing the conceptual foundation for selecting words in school instruction.

In a chapter in the fourth edition of the *Handbook of Reading Research*, Nagy and Hiebert (2011) drew on this work in corpora linguistics and from other disciplines to develop a framework for the selection of words for instruction. Nagy and Hiebert developed eight criteria for word selection and organized them into four groups based

on the role that words play in language, the lexicon, knowledge, and the lesson. These eight criteria, grouped by role, are provided in Table A1 in the online supplementary material (http://jlr.sagepub.com/supplemental). In the section that follows, we elucidate each of these factors, review existing literature on the learning of science vocabulary, and identify which of these variables were used to predict and explain students' word knowledge and learning in the current investigation.

Role in Language

Much of the meaning of words is neither accessed nor realized until words are encountered in contexts of language use. In written language, the focus of the present article, a small group of words account for the majority of the running words in text in all disciplines. By contrast, ideas—along with language and words—differ dramatically across disciplines and even topics within disciplines; because unique ideas are often expressed with unique words, we encounter wide variations in word use across domains. For example, domain specific words in an expository text on chemistry might include *acid, abrasive*, and *dissolve*, while an expository text on biology would unlikely include those words but might, instead, have words such as *prey* and *predator.* Taking both of these perspectives into account, Nagy and Hiebert (2011) identified two language role factors: frequency and domain specificity.

Frequency. Frequency has the longest history of any factor used to select words for instruction (Clifford, 1978). Its prominence as a criterion for selection is often justified by its predictive power in accounting for word learning. For example, Miller and Lee (1993) investigated a model of the variables underlying the acquisition order of word knowledge as indicated by their difficulty values on the Peabody Picture Vocabulary Test–Revised (PPVT-R), a test of receptive oral language capacity. Frequency accounted for most of the variance in the order in which the words are sequenced in this test (recall that serial position in this test corresponds to the order in which the words are typically acquired by students of different ages).

Frequency can be a nebulous variable, largely because it is confounded with several related variables, such as part of speech, polysemy, domain specificity, and even length. Typically, a focus on words in written texts means that variations in word meanings and parts of speech are seldom considered in determining a word's frequency. For example, a word such as *force* can function as a noun or verb, and within each form class, *force* has distinctive meanings that are associated with different disciplines or topics (e.g., physics, criminal law, philosophy). A word's predicted frequency in written text fails to take such diversity in meaning into account; this failure extends to homographs (e.g., *desert*) and words with distinctive etymological origins (e.g., *bank* of a river vs. *bank* for depositing money).

In a discipline such as science, the variable of frequency needs to be unpacked to a greater extent than can be captured by a typical overall index of frequency in written text (e.g., Zeno, Ivens, Millard, & Duvvuri, 1995), which is how frequency is measured in the current study. This shortcoming was part of the motivation for including,

in the present study, two variables usually confounded with frequency—part of speech and polysemy. We also included a third variable that has often been construed as an alias for frequency—word length as indexed by the number of syllables or letters in a word.

With respect to *part of speech*, nouns have been identified as more accessible to young first and second language learners than other parts of speech (Pigada & Schmitt, 2006). Kweon and Kim (2008) found that nouns are learned more easily than verbs or adjectives in the incidental learning of words that occurs through extensive reading of text. In a study of young children's learning of science vocabulary, Dockrell, Braisby, and Best (2007) examined part of speech, concreteness of nouns (observable vs. non-observable phenomena), and domain specificity. Part of speech predicted learning, with growth on receptive and expressive tasks significantly lower for verbs than for nouns and adjectives. In addition, they found that domain-specific words (e.g., *migrate*) were more difficult than domain-general words (e.g., *reproduce*).

Polysemy is the characteristic of having multiple, related meanings. Nerlich and Clarke (2003) note that polysemy is a matter of degree as most words take on slightly nuanced meanings depending on their context of use. Consider, for example, the meanings associated with the word *model*: (a) three-dimensional representation of a person or thing (as a model of a dam), (b) a person who displays clothing or cosmetics, and (c) a tool used in science to predict and explain phenomena. For elementary-level students introduced to the word *model* in science, the everyday meaning of a human model may be salient. Educators, particularly those concerned with science vocabulary (e.g., Osborne, 2002), distinguish between the everyday and scientific or technical meanings of words as a critical aspect of polysemy.

A small but robust literature indicates that polysemous words influence students' word acquisition. In the Miller and Lee (1993) examination of difficulty values on the PPVT-R, polysemy accounted for a significant but smaller proportion of the variance than frequency in explaining order of word acquisition. For English as a Second Language (ESL) students, the salience of a particular meaning can be an obstacle in learning secondary meanings of a polysemous word; Bensoussan and Laufer (1984) found that ESL students clung to the first (typically primary) meaning of a word that they had been taught when encountering a word in a text with a secondary meaning, even if the primary meaning did not make sense in the context.

In a content area like science, where distinctions between everyday and scientific meanings of words may be particularly salient, Johnstone (Cassels & Johnstone, 1985; Johnstone, 1991) has shown that students consistently experience difficulty in recognizing the correct scientific usage of words that have distinct everyday and scientific meanings (e.g., *random*). The present study extends the work of Johnstone and colleagues by identifying a scale for measuring everyday and scientific meanings of words, one that considers the semantic relatedness of everyday and specialized scientific meanings.

Word length and frequency have long been recognized as inversely related (Zipf, 1935). More recently, Strauss, Grzybek, and Altmann (2005) established a strong relationship between length and frequency in 10 languages, including English. The most

frequent words in a language are short; less frequent words are longer. Length, as measured in number of syllables or letters, has been shown to be a reliable predictor of word recognition (Bergman, Martelli, Burani, Pelli, & Zoccolotti, 2006) and of recognition of word meaning (Miller & Lee, 1993). Although word length may seem rather prosaic when compared with some of the more nuanced variables in the Nagy and Hiebert (2011) framework, we included word length as number of syllables in our explanatory variables because we wanted the more complex variables (e.g., morphological frequency, domain specificity) to compete with this simpler, historically powerful rival.

Domain specificity. Domain specificity refers to the likelihood that a word appears in a single academic discipline (e.g., *habeas corpus* or *transpiration*) versus in multiple disciplines (e.g., *evidence* or *prediction*). The construct has connections to polysemy in that words that are used across multiple disciplines often take on nuanced differences in meanings across disciplines (e.g., *force* in physics and in civics). But often the same meaning travels from discipline to discipline, especially with a group of words that have come to be known as a part of academic language (Coxhead, 2000). The word *evidence* appears in numerous content areas (*D* index of .94 on a scale where 1 indicates all content areas and 0 indicates only one; Zeno et al., 1995). Evidence in law and evidence in science or history may connote a slightly nuanced difference in meaning (especially with respect to what kinds of phenomena count as evidence) but denote the common meaning of providing proof for a claim. By contrast, the word, *googol*, has a *D* index of .0, indicating its narrow dispersion across domains, that is, it appears in a single domain. We use the *D* index as a measure of domain specificity in the current study.

Factors Related to the Role in the Lexicon

The lexicon refers two distinct groups of words: (a) to the corpus of words within a language or (b) in the case of an individual, her personal collection of words that function as a richly interconnected network of word meanings. Two types of networks are especially relevant in describing word learning and knowledge: semantic and morphological.

Semantic relatedness. We alluded to one aspect of semantic relatedness in the discussion of domain specificity—words related to a topic are nearly always semantically intertwined. The nature of the relationships among words within individuals' lexicons can take a number of different forms. One form consists of category-like relations, such as the superordinate class to which a phenomenon belongs (e.g., dogs are canines), features of the thing (dogs have fur, make a barking sound, are often domesticated), examples or the subordinate classes of a phenomenon (Labradors and German Shepherds are kinds of dogs), and lexical cousins (dogs and cats are commonly domesticated and are both carnivores). Other common semantic classes include collocation of words that commonly are used together (e.g., *iron/will, horses/neigh*), synonyms (e.g., *car/automobile)*, part-whole (*toes/foot*), instrumental (*fire/burn*), and scriptal (i.e.,

people develop scripts from experiences to understand behavior and events in different settings: *school/teacher* or *principal*; Moss, Ostrin, Tyler, & Marslen-Wilson, 1995).

Semantic relatedness can influence the learning of new words. Jenkins and Dixon (1983) identified four potential relationships that can exist between known and new words/concepts: (a) unknown word but a known concept that can be expressed succinctly (*altercation/argument*)—this is the essence of Beck et al.'s (2013) Tier 2 construct—rare words for familiar concepts, (b) unknown word with a simple synonym but student is not likely to know the concept referred to by the synonym (*paen/hymn*), (c) unknown word that does not have a simple synonym but can be described through experience or explanation (e.g., *odometer*/the item on the automobile dashboard that tells how many miles you've gone), and (d) unknown word that does not have a simple synonym and for which students are not likely to have extensive experiences (e.g., *refraction*). *Refraction* is illustrative of a wide range of words, often typical of science, for which the meaning of a given word depends on other words that are often equally as unknown in the specific semantic network of the topic the students are studying. To understand refraction, students would have to grasp the meaning of light or sound waves, density, and velocity. For words like these, acquiring new knowledge and new vocabulary requires that students be involved in rich, robust activities that involve reading, writing, speaking, and direct experiences.

Nagy, Anderson, and Herman (1987) conducted one of the few empirical efforts to establish the complexity of the semantic networks of words and their relationship to student learning. They adapted Jenkins and Dixon's (1983) scheme to examine four semantic relationships: (a) known concept/simple synonym, (b) known concept/phrase definition, (c) unknown concept/definition with known concepts, and (d) unknown concept/semantically complex relationships in definition. Of numerous word features (length, part of speech, morphological complexity), only the classification of a word in the fourth category distinguished those words that were challenging for students to understand in the context of texts.

In a domain such as science where the majority of the words fall into the fourth category, this scheme appears to have less applicability. Even with a sample that included words in the first three categories (e.g., *desperate, frantically, headlamp, topple*), Nagy et al. (1987) achieved a low level of interrater agreement (57%) with respect to category membership. Furthermore, the Jenkins and Dixon (1983) framework, as with the Beck et al. (2013) tier system, appears to distinguish among the words typically in narrative texts rather than those characteristic of informational texts.

Morphological family size. As texts become more complex, a substantial percentage of the new words that students encounter are morphological relatives of already known words. Nagy and Anderson (1984) estimated that 60% of the new words encountered by middle-school students are derived forms of words with fairly transparent connections to root words.

Speed of accessing a word's meaning is consistently affected by the number of words that are morphologically related to it. Words in larger morphological families

are recognized more rapidly than words in smaller families (Dijkstra, Martín, Schulpen, Schreuder, & Baayen, 2007). A morphological family can be large for a variety of reasons. One morphological family can have many members but, when all have relative low frequencies, the overall frequency of the family may be low; for example, *accuse* has 19 derivatives (Oxford Dictionaries, 2010), but as a group, the family is expected to have 17 appearances per million words of text (Zeno et al., 1995). Another family may have a large overall frequency due almost entirely to a single word; for example, *average* has a predicted frequency of 80 occurrences per million words, but its 5 derivatives account only for another 5 appearances per million).

Carlisle and Katz (2006) set out to unpack the features of family size that predict students' word recognition of an unknown word in a family. They considered four features of morphological families: the frequency of the derived word (i.e., the unknown word), frequency of the base word, the number of members in a morphological family, and the average frequency of the members of the word family. A two-factor model explained fourth- and sixth-graders' knowledge of the meaning of new, derived words: The first factor consisted of the frequency of the base word and the average frequencies of family members, and the second involved the frequency of the derived word and the number of family members.

Related to morphological frequency is the more general construct of morphological awareness—the ability of students to use knowledge of morphemes to recognize words and their meanings. Morphological awareness contributes to school-age students' text comprehension (Kieffer & Lesaux, 2008; Nagy, Berninger, & Abbott, 2006) as well as reading and spelling of individual words (Carlisle & Stone, 2005). Morphological awareness develops as children learn language (Clark, 1978) but, once students learn to read and spell, their morphological awareness spikes upward rapidly (Anglin, 1993), and it continues to evolve. For example, sixth and eighth graders recognize morphologically complex words significantly better than fourth graders (Tyler & Nagy, 1989) and, in turn, high school students have higher levels of morphological awareness than elementary students (Nagy et al., 2006). In the current study, we assigned morphological frequency values to words based on the frequency with which the members of a word's morphological family are expected to appear per million words.

Factors Related to the Role in Students' Existing Knowledge

Familiarity and conceptual difficulty of new words, according to Nagy and Hiebert (2011), reflect students' existing knowledge. Familiarity (a reader phenomenon) is loosely coupled with frequency (a characteristic of words in use) but it typically gets operationalized as an outcome (i.e., the likelihood that students know a word and its underlying concept). For example, *shampoo* and *toothbrush* are expected to occur fewer than two times per million words of text but are likely familiar concepts to most students, even in Grade 1. But other things being equal, more frequently occurring words will be more familiar to students, though there are many discontinuities between the two scales. Databases are available for both familiarity in written language (e.g.,

Living Word Vocabulary; Dale & O'Rourke, 1981) and in oral language (e.g., age of acquisition norms; Kuperman, Stadthagen-Gonzalez, & Brysbaert, 2012). In the present study, familiarity is an outcome variable of the study, operationalized as the percentage of students in the various samples who were able to respond correctly to a given item measuring knowledge of the word's meaning.

Conceptual difficulty refers to "the magnitude to which a concept or matter is hard to comprehend, due to the numerous theoretical ideas integrated and the precise manners in which they are linked" (Psychology Online Dictionary, n.d.). For example, the word *alveoli* has been found to be difficult for students to understand in the context of independent reading, whereas the word *dignified*, which can be defined with a single synonym (e.g., *noble* or *stately*), is more likely to be understood (Nagy et al., 1987). But it is not very satisfying to define a construct by its outcome; saying that conceptual difficulty is indexed by empirical difficulty (how many students at a given age level know the word?). It is akin to saying words are hard because they are hard. The last part of the definition in the *Psychology Online Dictionary* provides a hint about how it might be operationalized ("due to the numerous theoretical ideas integrated and the precise manners in which they are linked"), but to our knowledge, no one, including ourselves, has been able to accomplish it. Similar to familiarity, word difficulty is operationalized as the percentage of students in the various samples who were able to respond correctly to a given item measuring word learning.

The level of concreteness or abstractness of a word does not capture all possible dimensions of conceptual complexity; even so, concreteness might be construed as an indirect index of conceptual complexity because of the salience of concreteness in models of conceptual development (e.g., Piagetian stages) and its close link to materiality in models of cognitive processing (Varela, Thompson, & Rosch, 1991). Concreteness refers to how readily a word allows a typical reader to conjure up a "picture," a mental image, of the concept to which the word refers. And it has a rich history in cognitive examinations of word and text comprehension. The concreteness, or conversely abstractness, of a word has been shown to influence the speed with which words are processed, and the relative age at which words are acquired (Schwanenflugel, 1991). Many studies have found a significant effect for concreteness (an aspect of which has also been studied as imageability; see Paivio, Yuille, & Madigan, 1968) on speed of lexical processing (e.g., Altarriba, Bauer, & Benvenuto, 1999; Schwanenflugel, Akin, & Luh, 1992) and on vocabulary knowledge (De Groot & Keijzer, 2000).

Factors Related to the Role in the Lesson

The final cluster in the Nagy and Hiebert (2011) word selection framework addresses the centrality of words in particular texts and also in the larger curriculum in which a given set of texts are situated. This cluster represents a feature of the curriculum, rather than the words themselves. For narrative texts, where authors rarely repeat the same unique word when describing characters, contexts, and plots for a given story, defining this cluster is a critical one. But in science, where vocabulary, especially rare

vocabulary, is clustered around curricular topics (Marzano, 2004), the choices about vocabulary often precede the development of texts and materials; this follows from the fact that the words are often viewed as "following" the concepts they represent, and it is the concepts rather than the words that curriculum designers regard as central in planning curriculum. In the case of the current project, words were most often chosen for emphasis precisely because they represented the key concepts in the units. For this study, then, we did not need to identify criteria that capture this set of curricular factors in the word selection framework. Had we worked with narratives, these curricular considerations (i.e., figuring out which words to take the time to emphasize pedagogically) would have been a major task. And it is important to remember that it is not that the words for science text don't deserve instructional emphasis; it's just that selecting them is a simple task because choices are largely shaped by the conceptual structure of the content for a given text.

The Context of the Present Study

This work differs from most empirical studies in that it employs (more accurately repurposes) data on word knowledge and learning that was initially gathered for other purposes. Over a period of 10 years, we have been involved in an elaborate research and development effort to create, evaluate, revise, and test the efficacy of a range of integrated science-literacy units in which we have used reading, writing, and language as tools to enhance the acquisition of key ideas and inquiry processes in science. All of the measures of word learning used in our analyses came from studies designed to test the efficacy of the National Science Foundation (NSF)–funded integrated science-literacy program, *Seeds of Science/Roots of Reading* (*Seeds/Roots*), for which the units were developed, revised, and tested in field studies and efficacy studies. As we tested each of our units from Grades 2 to 5, we examined, quite unsurprisingly, vocabulary learning as an outcome variable. Our reasoning was that in the process of acquiring knowledge of science concepts through firsthand (doing science) and secondhand (reading and writing about science) inquiry activities, students were also acquiring the vocabulary words that index science concepts. (See also August, Branum-Martin, Cardenas-Hagan, & Francis, 2009.) So over several years, we compiled a significant database of information about students' word knowledge (from pretests before we taught the units) and word learning (from posttests after the units). But within those field studies and efficacy studies, we were not initially focused on word characteristics that might have predicted knowledge or learning.

As our interests in vocabulary learning developed, we wondered whether and how our approach fit the mold of secondary data analysis (Vartanian, 2011), in which researchers analyze data gathered for one purpose to meet another need or purpose. In these various field studies and efficacy studies, we did not attempt to examine the characteristics of individual words (e.g., frequency, polysemy, concreteness, and the like); instead, we were interested in an overall vocabulary score (or sometimes gain score) for each unit. By contrast, for our secondary analyses, the characteristics of individual words, examined across a range of units, became the precise focus of our

analysis. We believed that understanding more about the difficulty of words could help us make better decisions in the selection and instruction of words as we move ahead in our instructional design work.

Experimental Context of Curriculum Development

As the context in which we assess vocabulary knowledge and learning shapes the purposes and formats for assessing word knowledge, it is useful to share an example of how the work unfolds; as an illustration, we offer a portrait of the development process for our fourth-grade unit on light (Cervetti, Barber, Dorph, Pearson, & Goldschmidt, 2012).

Unit development within *Seeds/Roots* is a complex multiphase design and development process. It begins with a team of literacy and science educators who have a key curricular goal based on the prevailing science standards, either some amalgamation of state or national standards. That team prepares a unit that begins with science-related goals and standards (what we want students to know and be able to do), relevant literacy goals and standards (the sorts of routines and activities we know promote deeper reading, writing, and oral language activities in the context of science learning), and a plan for either an 8-week or a 4-week unit of activities that will help students and teachers meet the goals and standards.

Classroom tryouts. The model for classroom tryouts involves a three-person team of a teacher at a relevant grade level who collaborates with a science educator and a reading educator from the development team. This team is charged with the responsibility of smoke testing the unit in a single classroom. The purpose of the smoke test is to evaluate, revise, fine-tune and, in general, improve all aspects of the unit design—including the activities, the materials for students, and the pedagogical design embodied in the teacher guides. Revision and fine-tuning occurs on a daily basis and at the end of the initial phase of unit development. Included in this iterative design process are the various assessments, including the vocabulary assessments that were used in the current study to assess student learning over the course of the unit.

Of particular interest to this vocabulary investigation is the model of instruction that was used to teach words. The model emphasizes the need to highlight and deepen the conceptual relations among key domain-specific words. The focus is on learning new words/ideas in close association with other key words/ideas.

Field studies. The end result of the classroom tryout phase is a complete set of materials, including (a) student texts (for Grades 2-5, these have taken the form of science informational trade books); (b) student activity kits (for the hands-on, firsthand investigations); (c) student notebooks (for taking notes, recording observations, and creating reports of investigations to share with classmates); (d) pre- and posttests of science knowledge, vocabulary, reading comprehension, and writing; and (e) a two-part teacher guide that explains both the how (what to do at which point in the lesson) and the why and the what (why this approach or activity and not some other; what science

the teacher needs to know in more depth) of the pedagogy. These materials are sent out in the form of a unit-based kit to teachers from around the United States who have volunteered to be a part of what we call a field study. They teach the unit, administer the assessments, fill out instructional logs, questionnaires, and evaluations, and return all of the data to us. We, in turn, use the aggregated pre- posttest results as well as the teachers' qualitative feedback to revise features of the unit that have proved problematic, difficult, or unnecessary.

Efficacy studies. Revisions are made to the materials emerging from the field study phase to prepare a final, classroom-ready version of the units. A subset of the units are then selected as representative of the units at the grade level and tested in a full-scale experimental efficacy study. We find a district or area partner who is willing to work with us on an efficacy study in which classrooms and/or schools and/or districts volunteer to participate and are randomly assigned to a treatment or an appropriate comparison group. Often the comparison group is a business as usual control group, sometimes a cover-the-same-unit with regular nonintegrated units, and sometimes a text only version of *Seeds-Roots*. These groups can also become late-entry control groups who are given the *Seeds/Roots* materials at the end of the efficacy study phase.

Vocabulary data. The data harvested for the current investigation come from either Phase 2 field trials or Phase 3 efficacy studies, whichever was the last study conducted on a particular unit. What we realized when we looked across all of these studies that had been designed to evaluate the overall approach to curriculum delivery within the units is that we had, in the process, collected a massive amount of data on vocabulary across these units that varied by topic and the grade levels for which they were intended. Table A2 in the online supplementary material (http://jlr.sagepub.com/supplemental) shows the distribution of units across the grade levels of students for whom the units were intended.

The exact nature of the vocabulary instruction provided varied somewhat from study to study, but it was always guided by certain pedagogical principles, among them,

- Students should encounter new words in multiple language modalities—reading, writing, speaking, listening—and as frequently as possible (and never any singleton encounters).
- New words should be introduced explicitly (including the definition) and bound to hands-on scientific investigations: Students should use (i.e., speak and write) new vocabulary as they engage in inquiry.
- New words should be used when students are summarizing and reflecting on their learning within a science unit.
- To the degree possible, new words should be taught as parts of rich semantic networks of related ideas (e.g., habitat is taught alongside structural and behavioral accommodations and many biomes [forest, ocean, desert, plains]).

Though these underlying principles guided instruction in the units, the instructional activities varied across the units, and the degree of exposure to the words depended in part on the centrality of the word to the conceptual focus of the units. As is described below, we took these variations in exposure into account as we addressed our research questions. To summarize, the goal of this study was to learn, within the domain of science, what word characteristics predict word difficulty before instruction and whether these characteristics also influence the likelihood that students will learn the words over the course of science instruction. The framework identified by Nagy and Hiebert (2011) through a review of research was the basis for identifying seven variables—length, part of speech, polysemy, frequency, morphological frequency, dispersion (domain specificity), and concreteness.

Method

To remind us of the research questions that guided this secondary analysis, we were interested in knowing what characteristics of words provided the best explanations of the variation in both word knowledge (in the absence of instruction) and word learning (in response to instruction). To examine how well characteristics of science words predict both knowledge and learning of the words from instruction, we relied on a large corpus of existing data on vocabulary knowledge and acquisition drawn from a series of field studies and efficacy studies of instructional units that integrate literacy learning and inquiry-based science. The studies were conducted between 2004 and 2009. In all, the studies took place across more than 20 states. In each study, treatment students participated in a field study or an efficacy study of an integrated science-literacy curriculum unit between the pretest and posttest. The curriculum units, which were 20 or 40 sessions in length, involved students in hands-on inquiry experiences, reading, writing, and discussions around standards-aligned topics in earth, life, and physical science. Table A2, as described earlier, describes basic information about the individual studies, including the conceptual focus of each unit, the number of participants at each grade level, and the unit length. In Table A3 in the online supplementary material (http://jlr.sagepub.com/supplemental), we include characteristics of participants across studies. Data on demographics were limited for the solar system unit, leading to the decision to exclude it from the analysis.

Data Sources and Collection

Collaborating teachers or members of the research staff administered pre-post assessments of vocabulary knowledge; testing took place 8 to 12 weeks apart to allow teachers sufficient time to complete the units. Detailed administration guides were provided to the administrators. All assessments were administered to groups of students in their regular classrooms. The assessment items were read aloud to the students, but the students responded individually by selecting a response from a set of four possible responses for each item on a test form. The vocabulary assessments were mainly focused on words that were central to the conceptual territory covered by the treatment unit, though a few

nonfocal words were included for the sake of comparison. Each paper-and-pencil vocabulary assessment tested 9 to 25 words using one or more multiple-choice items. Given that some words appeared on the assessment for more than one unit and that different item types were often used to assess a word within a particular unit, most of the words were assessed two or more times. The item types ranged from the commonly used definition matching items to more innovative types, such as providing a heading for a list of words. For examples of the full range of item types, see Appendix AA in the online supplementary material (http://jlr.sagepub.com/supplemental). Unfortunately, there was a high degree of variability and thus arbitrariness in the development of items and item formats across the assessments. Student responses were manually transferred from the paper assessments into an electronic database. Responses were entered twice by different individuals to check for consistency/accuracy. Errors were resolved through a check of the original paper assessment. Responses were scored as correct or incorrect using recoding from within the statistical program. These entries too were checked twice for accuracy.

In preparing the data for analysis, we excluded two-word phrases, such as *water vapor* and *phase change*.[1] Where a word appeared on a vocabulary assessment more than once, either within a unit or across units at a grade, we averaged the scores on either the pre- or the posttest. This left us with 170 word-grade-level combinations. There were a total of 93 unique words in the study, 75 of which were tested at more than one grade level. The number of student responses contributing to each item score ranged from 26 to 866. See Appendix AB in the online supplementary material (http://jlr.sagepub.com/supplemental) for a list of words and grade-level appearances along with mean pretest scores. For the growth score analysis, we maintained the embeddedness of words within instructional units to include variables related to instructional emphasis, or opportunity to learn. These analyses involved 209 word-grade-unit combinations.

Independent Variables

Next, we assigned values for the independent variables that we used to predict word knowledge and word learning. As shown in Table A1, the independent variables were frequency, part of speech, polysemy, length, domain specificity (dispersion), morphological family size, and concreteness-abstractness.

Frequency in written English. The frequency with which a word is predicted to appear per million words of text was obtained for each word, using the analyses of Zeno et al. (1995). The metric used for this analysis was the *U* function, which is the frequency of a word per million words weighted by the measure of dispersion (*D*; see below), which indicates how widely a word is used in different subject areas. The word *rotate*, for example, has a *U* function of 5, which indicates that it is estimated to appear 5 times within a million-word corpus.

Part of speech. Each word was identified by its part of speech. This particular corpus contained only nouns, verbs, and adjectives. In the analyses, part of speech was treated as a categorical variable.

Polysemy. A scoring scheme was developed for use in this study to distinguish words with science-specific meanings from everyday use. The process began with extensive sorting and discussion of a set of 22 words from the database by three of the principal investigators, culminating in a five-category system for polysemy specific to science content. A description of the polysemy categories appears in Table A4 in the online supplementary material (http://jlr.sagepub.com/supplemental). As is evident in this taxonomy, we are using the word polysemy in an educational sense, rather than a linguistic sense. That is, in this research, polysemy refers to words that are orthographically identical, but semantically different.

Three of the principal investigators then classified the remaining 68 words by placing each word into one of the 5 categories in Table A4. The three investigators disagreed on 14 of the 68 words. For 11 of those disagreements, the difference was between the coding of words as No. 1 or No. 2 (science meaning the same as everyday meaning or everyday and science meanings related but not quite as precise). In all 11 cases, two of the three raters had the same rating and one of the three differed (although not the same configuration of raters in each). The decision was to use the rating on which two of the three principal investigators agreed.

For the three remaining items (4% of the cases), the difference between raters spanned two or more categories. In two cases, the difference was between No. 1 and No. 3 (science meaning = everyday meaning and everyday meaning = independent of science). In the final case, the difference was substantial. One rater identified the meaning as an everyday meaning (No. 1), while the other two raters identified the meaning as scientific (No. 5). The inconsistencies among raters were resolved through discussion among the investigators.

In the end, we found that we had too few words at Levels 4 and 5 to analyze the data using this taxonomy. For example, only one word was classified as Type 4 at Grade 2, and only one was classified as Type 4 at Grade 4. As a result, we collapsed the scheme of five categories into two for analysis: nonpolysemous words (Types 1 and 5) and polysemous (Types 2, 3, and 4).

Length. Each word was assigned a value for length based on the number of syllables.

Domain specificity. We used Carroll, Davies, and Richman's (1971) dispersion index to assign values for domain specificity. Carroll et al. originally proposed the dispersion index (D) to report on how widely a word is used in different subject areas. Words that appear in only one content area have a *D* value of 0; words that appear in many content areas (e.g., social sciences, science, mathematics, fine arts, literature) could have a *D* value as large as 1.0. Just as for frequency of individual words and also morphological families of words, data on dispersion were obtained from the Zeno et al. (1995) database.

Morphological frequency. The frequency with which the members of a word's morphological family are predicted to appear per million words was established by combining the U functions from the Zeno et al. (1995) database for the morphological members, which had a transparent association with a word. According to Nagy and Anderson

(1984), transparent associations are ones that readers can make with knowledge of the target word. For example, readers who know the word rotate are likely to understand *rotation* ($U = 9$), *rotational* ($U = .60$), *rotates* ($U = 6$), *rotating* ($U = 4$), and *rotated* ($U = 1$). Morphological members were obtained only from words with frequencies of one or greater per million words (i.e., the 19,468 most frequency words within the Zeno et al., 1995, database). With the U function of .60 for *rotational* falling below the 1.0 threshold, the combined U function of morphological members transparently associated with rotate was 20, not 20.60.

Concreteness. Concreteness values were obtained from the Brysbaert, Warriner, and Kuperman (2013) database. This database uses a technique similar to that of the original database of concreteness (Paivio et al., 1968) where adults rate a word on a Likert-type scale from highly concrete (e.g., *seaweed*) to highly abstract (e.g., *the*). Even though this database is 10 times the size of previous databases (e.g., Coltheart, 1981), the Brysbaert et al. database did not provide a portion of the sample—18% of the words (e.g., *isopod*). However, the Brysbaert et al. database with ratings of the concreteness of 39,954 words meant that the analysis of concreteness was considerably more robust than previous analyses of this feature of science vocabulary.

Results

Results for Pretest Scores

Overall approach to analysis. In the first phase of analysis, we examined means, standard deviations, and correlations among the word characteristics that existing research led us to believe might be related to word knowledge. Findings from the correlational analyses revealed moderate and consistent relationships between morphological frequency and frequency in English ($r = .33$-$.53$) and domain specificity ($r = .49$ to.61) across all three grades. Moderate relationships were also systematically observed between length and frequency in English ($-.31$ to $-.33$). Table 1 includes the means and standard deviations for the vocabulary scores and the word characteristics. Table A5 in the online supplementary material (http://jlr.sagepub.com/supplemental) provides the correlations among the word characteristics.

We then employed regression analysis to predict pretest score using the seven word characteristics as independent variables. We first looked at individual predictors and then built models using several of the individually significant predictors of pretest score. Furthermore, to account for the possibility that the measurement of word knowledge was affected by item type, we also examined the relationship between item type and pretest scores; item type was significantly related to the pretest score at the second, $R^2 = .09$, $F(3, 76) = 2.77$, $p < .05$, and third grade, $R^2 = .08$, $F(4, 97) = 4.66$, $p < .01$, levels; thus, later analyses controlled for item type for these grades. The number of words included in each analysis varies, because some values were not available for all words using our chosen indices (concreteness, domain specificity etc.), but no analysis included fewer than 28 words.

Table 1. Means and Standard Deviations: Vocabulary Scores and Word Characteristics.

	Grade 2		Grade 3		Grade 4	
	M	*SD*	*M*	*SD*	*M*	*SD*
Pretest	0.50	0.16	0.58	0.23	0.61	0.20
Growth	0.26	0.13	0.18	0.15	0.15	0.13
Domain specificity	0.70	0.19	0.69	0.19	0.71	0.19
Polysemy	1.41	0.50	1.45	0.5	1.48	0.51
Part of speech	1.41	0.60	1.38	0.61	1.35	0.58
Length	2.32	0.90	2.55	1.07	2.77	1.19
Morphological frequency	46.78	51.31	53.97	60.46	69.86	74.56
Frequency in English	44.8	50.35	40.05	46.33	42.23	75.73
Concreteness	4.39	1.31	4.29	1.27	4.25	1.28

Note. Ranges: Pretest (0-.97), Growth (−.08-.67), Domain Specificity (.26-.96), Polysemy (1-2), Part of Speech (1-3), Length (1-5), Morphological Frequency (0-266), Frequency in English (0-464), Concreteness (2-7).

Results with individual predictors. Results using the word characteristics as individual predictors of pretest score are presented in Table 2. Polysemy and frequency in written English (henceforth referred to as "frequency") explained significant variance at all three grade levels (polysemy: 28% (Grade 2), 22% (Grade 3), and 15% (Grade 4); frequency: 7% (Grade 2), 12% (Grade 3), and 12% (Grade 4). Length explained 8% if the variance in pretest scores at Grades 2 and 16% at Grade 3 but did not account for significant variance at Grade 4.

Across grade levels, students had more knowledge of single-meaning and frequently occurring words, and shorter words were generally easier for younger (second- and third-grade) students. Furthermore, when item type was entered as a control variable at Grades 2 and 3, polysemy and frequency remained significant predictors of word knowledge. However, while length continued to be significant at second grade, it only approached significance at third grade, $R^2 = .12$, $F(5, 96)$ $p = .10$, with item type in the model. These mixed results suggest that, the relationship between length and word knowledge remains uncertain; however, it is possible that there are developmental differences in importance of length in word knowledge acquisition such that longer words tend to be the hardest for the youngest children. In contrast, polysemy and frequency appear to be important characteristics that explain differences in the level of word knowledge obtained, regardless of item format (e.g., definition or cloze) and grade level.

We also tested higher order (cubic and quadratic) terms for the three predictor variables. Findings indicated there were no significant quadratic or higher order effects, so the remaining regressions were run using only linear terms.

Models using multiple predictors. Stepwise forward and backward regression was used to examine the relationships among and variance explained by the individually

Table 2. Individual Predictors of Pretest Score by Grade.

	Grade 2				Grade 3				Grade 4			
Predictor	Beta	*t*-value	R^2	*F*	Beta	*t*-value	R^2	*F*	Beta	*t*-value	R^2	*F*
Domain specificity	−.03	−0.24	.00	(1, 52) = 0.06	.07	0.47	.00	(1, 70) = 0.22	−.15	−0.88	.02	(1, 37) = 0.77
Polysemy	−.17	−4.54***	.28	(1, 54) = 20.6	−.21	−4.44***	.22	(1, 72) = 19.74	−.15	−2.53*	.15	(1, 38) = 6.39
Part of speech	—	—	.05	(2, 53) = 1.44	—	—	.02	(2, 71) = 0.57	—	—	.09	(2, 37) = 1.79
verb	−.04	−0.88	—	—	.06	0.97	—	—	−.00	−0.04	—	—
adverb	−.14	−1.57	—	—	.06	0.58	—	—	2.66	1.87	—	—
Length	−.05	−2.15*	.08	(1, 54) = 4.61	−.09	−3.90***	.16	(1, 72) = 15.25	−.03	−1.06	.03	(1, 38) = 1.11
Morphological frequency	.00	1.65	.05	(1, 48) = 2.73	.00	0.66	.01	(1, 62) = 0.43	.00	1.32	.06	(1,27) = 1.74
Frequency in English	.00	2.02*	.07	(1, 53) = 4.06	.00	3.17**	.12	(1, 71) = 10.02	.00	2.30*	.12	(1, 38) = 5.26
Concreteness	.02	1.35	.04	(1, 49) = 1.82	.01	0.52	.00	(1, 67) = 0.27	−.00	−0.13	.00	(1, 28) = 0.02

Note. Part of speech was entered as a set of dummy variables with "noun" as the reference category.
*$p < .05$. **$p < .01$. ***$p < .001$.

significant predictors—polysemy, frequency, and length—and the outcome variable, pretest score. Across the various models, polysemy accounted for significant variance (between 10% and 28%) in the pretest regardless of which order it was entered in the model. Frequency explained between 8% and 12% of the variance at each grade level when entered after polysemy. However, when length was entered first, frequency was not significant at the second grade, and explained less additional variance at Grade 3 (3%) and Grade 4 (10%). Length only accounted for significant variance, 5%, when entered after polysemy and frequency at third grade. Table 3 provides the results from the stepwise regressions.

We reasoned that the impact of other word characteristics may be heightened if the words are also rare and therefore likely to be unknown to students. Therefore, we tested interactions between frequency and the other predictor variables. Our expectations notwithstanding, there were no significant interactions between frequency and either length or polysemy.

Across grade levels, the final models for the pretest point to a consistent and substantial role for polysemy in word knowledge. While there does appear to be a somewhat consistent yet relatively smaller effect of frequency, findings suggest that the relationship between length and vocabulary knowledge is probably not educationally meaningful, when controlling for polysemy and especially frequency. However, at all grades, frequency and length are moderately correlated (r is approximately −.32), so it may be the case that it is difficult to parse out the unique influence of these two-word characteristics on vocabulary knowledge.

Results for Growth Scores

Results with individual predictors. In the second phase of analysis, we looked at the explanatory relationships of the same set of word characteristics—polysemy, frequency, and length—with growth scores. To account for the degree to which instruction and exposure to the words would obviate the challenges presented by word characteristics at pretest, we controlled for number of appearances in student books (range = 1 to 293) and number of instructional sessions in which the word was directly instructed or heavily used by the teacher (range = 1 to 40). We removed from this analysis words with a zero for instructional emphasis.

Findings indicated that length did not predict growth score at any grade level. Polysemy explained significant variance at Grades 2 (13%) and 3 (7%), and was marginally significant at Grade 4 (explaining 8% of the variance). Frequency explained significant variance in growth scores at Grade 3 (12%) and was marginally significant at Grade 4 (explaining 8% of the variance in growth). Importantly, the pattern of results in the growth analysis was in the opposite direction from the pretest scores. At pretest, findings indicated that students tended to have significantly more knowledge of single-meaning and/or frequent words; by contrast, growth at posttest was positively related to polysemy (students made more growth on polysemous than single-meaning words) and negatively associated with frequency (students made greater growth on low frequency words). These directional changes suggest that science

Table 3. Summary of Stepwise Regression Analyses With Polysemy, Frequency, and Length as Predictors of Pretest Scores.

	Grade 2					Grade 3					Grade 4				
Predictor	Beta	*t*-value	R^2	R^2-change	*F*	Beta	*t*-value	R^2	R^2-change	*F*	Beta	*t*-value	R^2	R^2-change	*F*
1. Polysemy	−.17	−4.54***	.28	.28	(1, 54) = 20.60	−.21	−4.44***	.22	.22	(1, 72) = 19.74	−.15	−2.53*	.15	.15	(1, 38) = 6.39
2. Frequency	.00	2.52*	.34	.08	(2, 52) = 13.26	.00	3.80***	.34	.12	(2, 70) = 18.26	.00	2.07*	.23	.08	(2, 37) = 5.61
3. Length	−.02	−0.80	.35	.01	(3, 51) = 8.99	−.05	−2.37*	.39	.05	(3, 69) = 14.85	−.01	−0.17	.23	.00	(3, 36) = 3.65
1. Length	−.05	−2.15*	.08	.08	(1, 54) = 4.61	−.09	−3.90***	.16	.16	(1, 72) = 15.25	−.03	−1.06	.03	.03	(1, 38) = 1.11
2. Frequency	.00	−1.45	.10	.02	(2, 52) = 3.12	.00	2.17*	.22	.03	(2, 70) = 10.16	.00	2.03*	.13	.10	(2, 37) = 2.65
3. Polysemy	−.16	−4.31***	.35	.15	(3, 51) = 8.99	−.19	−4.36***	.39	.17	(3, 69) = 14.85	−.13	−2.25*	.23	.10	(3, 36) = 3.65

*$p < .05$. **$p < .01$. ***$p < .001$.

vocabulary growth, while shaped by existing knowledge, is more prominent among words that are not well known prior to instruction; that is, students' existing lexicons, which are comprised of higher frequency words with a tendency toward single meanings, may well serve as the foundation on which they are able to access the meanings of less common/more polysemous words. Table 4 displays the results for polysemy, frequency, and length as individual predictors of growth scores.

Because item type was significantly related to Grade 2 growth, $R^2 = .17$, $F(5, 71) = 2.83$, $p < .05$, we also examined the roles of polysemy and frequency, when controlling for item type at this grade level; however, both variables fell out of significance with item type in the model. Clearly, the frequency and polysemy effects are not consistent across item types at Grade 2, with some item types exhibiting sensitivity to these variables and others not. Ferreting out the specifics of these interactions will have to await a more deliberate effort, with planned variations across item types for various levels of polysemy and frequency, than we were able to muster in this opportunistic investigation.

Models using multiple predictors. Following the same analytic strategy as at the pretest, polysemy, frequency, and length were entered in a series of stepwise forward and backwards regressions. We report only the Grades 3 and 4 results here given the complicating interactions with item format at Grade 2.

Consistent with the results reported in for single predictors, length was not significant in any model. At Grade 3, polysemy explained approximately 8% of the variance in growth, regardless of which order it was entered; at Grade 4, it was borderline significant when entered first, and explained significant variance (11%), when controlling for length and frequency. Similarly, results indicated that frequency was an important word characteristic regardless of whether it was entered after polysemy or length at both grade levels. Albeit, consistent with the pretest results, frequency explained less variance when entered after length (Grade 3, 9% and Grade 4, 7%), which, again, drew attention to the overlap between length and frequency (i.e., longer words tend to be less frequent than shorter words). Table 5 provided the results from the stepwise analyses with growth as the outcome variable.

In sum, it appears that, even after students had received instruction that was intended, at least in part, to increase knowledge of word meanings, polysemy and frequency continued to exert substantial influences on word learning in science at Grades 3 and 4—with a relatively stronger effect for polysemy than frequency. Our best explanation is that because older children tend to have already developed well-specified representations of many relatively straightforward (i.e., more frequent and less polysemous) science words (as evidenced in the pretest), growth appears to involve the opportunity to expand the lexicon to include the meanings of more polysemous and less frequent words. However, as discussed, findings for younger, Grade 2 children are less clear, highlighting the possibly that they are in an in-between phase (making it difficult to measure and detect the influence of these word features), where they are simultaneously building on their prior knowledge to acquire the meaning of some polysemous and rare words, while at the same time they are also in the process of solidifying basic knowledge of many beginning (frequent and unambiguous) words.

Table 4. Individual Predictors of Growth Score by Grade.

	Grade 2					Grade 3					Grade 4				
Predictor	Beta	*t*-value	R^2	R^2-change	*F*	Beta	*t*-value	R^2	R^2-change	*F*	Beta	*t*-value	R^2	R^2-change	*F*
Polysemy	.10	2.91**	.19	.13	(3, 50) = 3.97	.08	2.27*	.10	.07	(3, 70) = 2.53	.08	1.90†	.15	.08	(3, 38) = 2.18
Frequency	−.00	−1.41	.09	.03	(3, 50) = 1.68	−.00	−3.10**	.15	.12	(3, 70) = 4.07	−.00	−1.95†	.15	.08	(3, 38) = 2.25
Length	.03	1.36	.09	.03	(3, 50) = 1.63	.03	1.47	.06	.03	(3, 70) = 1.51	.03	1.25	.10	.03	(3, 38) = 1.45

Note. R^2-change reflects the difference in R^2 between the model with only Book Appearances and Instructional Emphasis compared with the model with the addition of each individual predictor.
†$p < .10$. *$p < .05$. **$p < .01$. ***$p < .001$.

Table 5. Summary of Stepwise Regression Analyses With Polysemy, Frequency, and Length as Predictors of Growth Scores, Controlling for Book Appearances and Instructional Emphasis.

	Grade 2					Grade 3					Grade 4				
Predictor	Beta	*t*-value	R^2	R^2-change	*F*	Beta	*t*-value	R^2	R^2-change	*F*	Beta	*t*-value	R^2	R^2-change	*F*
1. Polysemy	.10	2.91**	.19	.13	(3, 50) = 3.97	.08	2.27*	.10	.07	(3, 70) = 2.53	.08	1.90†	.15	.08	(3, 38) = 2.18
2. Frequency	−.00	−2.03*	.26	.07	(4, 49) = 4.19	−.00	−3.51**	.23	.13	(4, 69) = 5.27	−.00	−2.46*	.27	.12	(4, 37) = 3.37
3. Length	.02	0.89	.27	.01	(5, 48) = 3.50	.01	.50	.24	.01	(5, 68) = 4.22	.01	0.69	.28	.01	(5, 36) = 2.75
1. Length	.03	1.36	.09	.03	(3, 50) = 1.63	.03	1.47	.06	.03	(3, 70) = 1.51	.03	1.25	.10	.03	(3, 38) = 1.45
2. Frequency	−.00	−1.20	.12	.02	(4, 49) = 1.60	−.00	−2.76**	.15	.09	(4,69) = 3.13	−.00	−1.78†	.17	.07	(4, 37) = 1.94
3. Polysemy	.11	3.15**	.26	.13	(5, 48) = 3.50	.10	2.72**	.24	.09	(5, 68) = 4.22	.09	2.56*	.28	.11	(5, 36) = 2.75

†$p < .10$. *$p < .05$. **$p < .01$. ***$p < .001$.

Discussion

This study was designed to identify whether particular characteristics of words predicted students' knowledge before instruction and the ease/difficulty with which students learn them during units of science instruction. We used a theoretical framework for identifying potential variables, with the Nagy and Hiebert (2011) framework as the foundation. We tested seven word characteristics—frequency, part of speech, polysemy, length, domain specificity (dispersion), morphological frequency, and concreteness.

Word Knowledge

Frequency, polysemy, and length were predictive of pretest scores at two or more grade levels. First, there appeared to be a small, yet consistent effect of frequency across the grades, even when controlling for polysemy, length, and item type. However, the frequency findings are somewhat limited by the overlap between frequency and length (e.g., shorter words tend to be more frequent than longer words), and it was thus difficult to disentangle these two characteristics. The results support the long-standing finding that frequency is an important feature of word knowledge, which is not a new finding; other things being equal, students are more likely to have already learned words that occur more frequently in the language. Based on previous research, we anticipated that morphological frequency would also be a factor (Carlisle & Katz, 2006). That did not prove to be the case—perhaps because a number of the science-specific words in our sample are not part of large morphological families.

The critical role of frequency in word knowledge does not always work to students' advantage. Many words are highly frequent because they have multiple meanings and thus serve multiple linguistic functions in ordinary and academic discourse. When the multiple meanings include everyday, colloquial use and scientific and technical use, word frequency can add a challenge to word learning precisely because students may apply the more familiar meaning rather than the most appropriate meaning in a particular context. Perhaps for this reason, polysemy exerted a greater influence on word knowledge and learning than did frequency. Polysemy predicted performances prior to instruction in all grade levels (even when controlling for frequency, length, and item type), and had the strongest effect on the youngest children. It is plausible that for younger students who are attempting to develop solid understandings of words, the idea that some words have many meanings, particularly precise meanings in a subject area like science, may be a difficult notion. Although the constraints of the present study, including the particular measures used and the nature of the sample, need to be considered before too much is made of this finding, the current evidence does support the idea that the everyday meanings of scientific terms are a potential source of interference in meaning-making in scientific discourse (Osborne, 2002). Moreover, the findings in the current study are consistent with previous studies (e.g., Bensoussan & Laufer, 1984). Given that many words in science have both a specialized scientific meaning and a more common everyday meaning (e.g., *property, model, energy, force*,

and *charge*), it may be useful to target such words for additional instruction, perhaps highlighting differences between everyday and scientific meanings.

The length of words was a factor at a particular level and point in time—third graders' knowledge prior to instruction. A possible explanation for the influence on third graders' knowledge but not on second or fourth graders' performances comes from an examination of the *p* values for words in Appendix AB. On the pretest, second graders achieved a *p* value of .75 on only three words—two of which were monosyllabic (*soil, root*) and one multisyllabic word with a consonant-vowel-consonant (CVC) pattern in the first syllable (*mixture*). By third grade, the ability to process multisyllabic words had increased substantially. In addition to the three words known by second graders, three quarters of the third graders knew 21 words. With two exceptions (*beach, prey*), all of these words were multisyllabic, including several three- and four-syllable words (e.g., *material, measurement*). The big jump in recognizing multisyllabic words appears to have occurred from second to third grade. By fourth grade, word recognition seems to have stabilized, and the factors of frequency and especially polysemy seem to have risen to a position of greater influence.

Word Learning

The second set of analyses explored the degree to which the important characteristics in word knowledge (frequency, polysemy, and length) were also associated with learning. Notably, findings suggested that, because students had already demonstrated their understanding of many frequently occurring and single-meaning words, knowledge growth tended to reflect the opportunity to build on this foundation to learn more challenging (less frequent and/or polysemous) words. As was true for the word knowledge analysis at pretest, across grade levels, polysemy had the most powerful effect (relative to frequency and length) and exerted the strongest influence on the youngest children, even when controlling for frequency and length. Frequency exhibited a consistent effect at Grades 3 and 4; less frequent words exhibited the greatest gains from pre- to posttest—that is, lower frequency words were less likely to be known in advance of instruction, but low frequency did not alone inhibit word learning

Whatever role length has in shaping word learning appears to be linked to its entanglement with frequency; recall that frequency changed from a significant to a nonsignificant predictor of word learning when the model controlled for length. There was no indication, however, that length continued to play an independent role in word learning—perhaps lending some support to the relationship between frequency and learning. In acknowledging the important role for frequency and polysemy, we need to remind ourselves that none of the key findings held at Grade 2 when taking into account the format (i.e., item type) in which learning was assessed; so the task of figuring out how the item format used to assess meaning interacts with word learning remains ahead of us.

Looking across the entire set of results for both knowledge and learning—and consistent with previous research (e.g., Bensoussan & Laufer, 1984)—our results provide evidence that frequency is an important word feature to consider as teachers select

texts for independent reading and plan instruction. Furthermore, our findings seem to provide even stronger support for the idea that polysemy is a core quality of science words related to knowledge and learning. Without benefit of instruction, as students encounter new domain-specific words, knowledge tends to be enhanced when students encounter words that occur more frequently and by the unambiguous nature of words with only one meaning. However, once this knowledge base has been well established, students are capable of building meaning representations for words that occur relatively less often in print and/or have several meanings. It therefore may be useful, for example, for teachers to scaffold the learning of science words that have both a specialized scientific meaning and a more common everyday meaning by first introducing a set of prerequisite unambiguous words.

Inconsistencies With Previous Research

We have already commented on the failure of the morphological frequency variable to predict either word knowledge or learning. Other variables that did not enter into the model were concreteness, part of speech, and domain specificity.

Concreteness. The concreteness finding stands in contrast to the studies in which this variable has been a critical contributor to word learning. One explanation for this finding may be that the majority of the words in our corpus were, on the whole, abstract. Even among the highly concrete words, only a handful of the words were truly concrete and likely known by elementary students (e.g., *beach, soil, ocean*). While adults may identify other words as concrete (e.g., *engineer, sphere*), many elementary students may not have encountered these concepts prior to introduction in academic settings.

Part of speech and domain specificity. In contrast to the current research, Dockrell et al. (2007) found that part of speech and domain specificity influenced acquisition of science vocabulary among 5- and 6-year-old students. Contextual differences between the studies may explain the differences. Not only were the students in the present study older than those in the Dockrell et al. study, but the definition of domain specificity appears to differ as well. For example, the use of dispersion values from Zeno et al. (1995), when applied to the Dockrell et al. data, reveals that several words that they had classified as domain specific had dispersion rates that indicated use across several domains (e.g., *lunar*, $D = .63$; *migrate*, $D = .59$), while several words classified as domain general were predicted to occur in only a single domain (e.g., *googol*, $D = .0$; *phylum*, $D = .02$).

Limitations

Using existing databases can create methodological challenges—as occurred in this study. But there are also benefits to relying on databases that others have built in that there is a large amount of data on which to draw—data which may be difficult, if not impossible, to gather in the conduct of a specific study. Repurposing of data sets is a

frequent strategy used to explore issues in educational research (Glass, 1976), including research on literacy issues (Guthrie, Schafer, & Huang, 2001). Issues surrounding differences in item types, vocabulary across grades, and the nature of instruction—all related to the use of a preexisting database—mean that we recognize the suggestive, rather than definitive, nature of our findings.

Secondary analyses. The fundamental premise of the study, that we could gain insights about factors that influence word knowledge and word learning by repurposing data gathered for distinctly different purposes, inevitably casts a shadow on the validity of such an effort. Why did we not just design an appropriate experimental study in which we manipulated the variables of interest in a highly systematic manner so that we could provide more definitive proof of their influence (or lack of influence)? As tempting as it is to move directly to controlled experiment, there are some advantages to mining data from a range of studies that vary in terms of the words selected, the item types employed, and the topics examined: If the phenomena of interest survive the journey through a range of variations in key factors, then it looks like they might be generalizable. Essentially, it boils down to the classic trade-off between internal validity, which nudges us toward more careful controls, and external validity, which inclines us to embrace natural variability in topic, item types, topics, and words chosen. In an early foray into a research space, we think that the type of data mining in which we engaged can give us a relatively efficient way of learning what is worth more careful analysis in later, more carefully designed efforts.

Item types. Particularly noteworthy, are the shortcomings of our items design. That is, secondary analysis involved the use of preexisting and thus somewhat arbitrary instruments to measure the vocabulary knowledge and learning constructs. We did not begin this journey across eight different unit evaluations with a well-developed a priori theory of item types. Instead, we learned as the journey progressed and as we discovered more about the functions that different item formats could serve. As a result, we cannot speak authoritatively about the specific affordances and constraints of particular item formats; all we can do is to point to the importance of taking item type into account and issue a promissory note for ourselves and others to take a more systematic approach in examining it in future studies of vocabulary learning. In the current study, we are unable to use our findings to improve or refine the measurement tools, or to use them to enhance our understanding of the constructs (Wilson, 2005).

Polysemy. One consequence of using existing data for these analyses is that we had insufficient variation across the various categories of polysemy that we had initially defined. This is a particularly significant limitation, because polysemy—even in the two-level analysis—is our most robust predictor of word knowledge and word learning. The current study leaves much to be learned about the role of polysemy in word learning in content-area instruction, but does highlight the fact that polysemy deserves further investigation and much more attention than it has received in schemes of vocabulary selection and instruction.

Conceptual complexity. Despite what seemed to us like some Sisyphean efforts on our part, we failed to find a plausible way to operationalize the construct of conceptual complexity. At one point, taking a clue from the work of Jenkins and Dixon (1983), we thought that the complexity of the definition of the scientific meaning of a word (as used in the unit) might provide a useful index of its conceptual complexity. So we looked at the ways in which these key words were defined in the glossaries provided to students, and we conducted a Kintsch-like propositional analysis (see Covington, 2007). We scaled all of the words according to the number and average depth (how deeply embedded is each proposition) of the propositions in its glossary definition. It did not work; neither the number of propositions nor their average depth predicted *p* value.

We considered other indicators but quickly rejected each. For example, semantic relatedness as indexed by the number of connections a word has to other words in the semantic network appropriate for the definition as used in the text. But what constitutes complexity—more or fewer connections? In one sense, the more connections a word has, the greater the likelihood a student might know it or the easier it ought to be to learn (or infer) its meaning. Does that mean that complexity facilitates word knowledge or learning? In the end, we concluded that conceptual complexity cannot be appropriately conceptualized as a characteristic of a word alone and instead resides as the intersection of reader, word, and task or context. That is, the conceptual complexity of a word resides in large part in the depth of understanding required by a text or task.

Looking Ahead

This study used the Nagy and Hiebert (2011) framework, which was intentionally built to serve as a working framework to guide word selection for instruction, not as a definitive model of meaning acquisition. Because it was focused on pedagogy rather than acquisition and/or learning, Nagy and Hiebert did not distinguish between word features as outcomes or predictors. The need to make this distinction became evident in the current work as we tried to identify factors that influence knowledge and learning. As we stated in noting our lingering concern about conceptual complexity, we believe that familiarity and conceptual complexity are precisely the areas that demand our attention in crafting an even richer model of acquisition.

Furthermore, the failure of particular variables, especially those that have demonstrated their efficacy in other studies, to predict word knowledge or growth in this study does not mean that they should be abandoned in future studies. Words are unique in the role that they have in different disciplines, perhaps even different topics. Within a model of word learning, we could anticipate that different variables influence students' word knowledge and learning at different points in time in students' development, as well as in different disciplines. A single word has many attributes and we need to recognize that these attributes may exert different sorts of influences across relevant contextual variables.

A perspective of this study that merits attention in the design of instruction is the distinction between existing (prior if you will) knowledge of a word and knowledge

acquired through instruction. Knowing which words to teach is critical, especially so in a content area such as science where the number of concepts (and associated words) are many. Some research (Gates, 1962; Stallman et al., 1989) suggests that many of the words chosen at least for basal reading (likely narrative) instruction are already known by students. Yet a one-size-fits-all perspective has dominated selecting vocabulary for instruction (as with many other aspects of literacy instruction!). For example, Spycher (2009) used a seven-step routine (e.g., repetition of the words, discussions of their meaning and use in context, active use of the words by students, and teacher-led questioning) to teach kindergartners Tier-2 and Tier-3 words. Students in the intervention classroom with this routine made greater growth on the 20 target words than students in an implicit instruction condition. But a closer look at the results reveals substantial variation in pre-post gains among those 20 words. For familiar words (e.g., *escape, amazing, hatch*), changes were insignificant but the mean growth was large for rare words (e.g., *metamorphosis, pupa*). The variability in change scores among the words begs the question of whether identical instruction was appropriate for all words. Just as we are advised to respond to individual differences among students, perhaps we need to be equally as sensitive to individual (more likely categorical) differences among words.

The approach that was taken in the current study (i.e., examining what is known by a large sample of students) could produce an alternative perspective for curriculum developers in choosing words for instruction than the frequent approach of devoting equivalent attention to all words for a topic. For example, words that are likely to be known could serve as anchors for an instructional unit, deployed as bridges to develop some of the more challenging concepts in the unit. Consider, for example, the vocabulary related to a unit on understanding the mixing of substances—words such as *ingredient, mixture, solution, dissolve, property, abrasive, acid*, and *soluble*. The *p* values in Appendix AB indicate that many third graders had an understanding of *mixture* ($p = .92$) and *ingredients* ($p = .86$) prior to instruction, while many did not know the meanings of *solution* ($p = .36$) and *soluble* ($p = .60$). This knowledge of mixtures could be the basis for developing an understanding of solutions and properties that make ingredients soluble in particular solvents.

For curriculum designers and publishers of guides for teachers, many databases are now available to help identify critical word-level variables that influence students' understanding of words; these include frequency, age of acquisition, word length, and part of speech. But for words that are unique to content areas such as science, the most powerful idea in shaping word learning may be knowledge networks, and the word learning data in this investigation certainly point in that direction. Recall, for example, the reversal in direction of the frequency and polysemy variables from the word knowledge (pretest) to the word learning (posttest) analyses. The idea of bridging from known to unknown words is an appealing instructional ploy, one we should attempt to evaluate carefully. Informed by evidence, instruction could focus on linking to words that are especially challenging and, in so doing, take a step closer to helping students develop the content knowledge they will need to be college-, career-, or citizen-ready. Investigations of this kind are particularly important in light of new standards that

explicitly link ELA with content-area learning. For example, both the aforementioned CCSS and the Next Generation Science Standards (NGSS; NGSS Lead States, 2013) include connections to the CCSS-ELA with each performance standard. With respect to vocabulary in particular, the NGSS and the framework upon which they were built recognize a strong role for literacy in learning and practicing science, including knowledge of the words that students need to communicate about science.

Authors' Note

Any opinions, findings, and conclusions or recommendations expressed in this materials are those of the authors and do not necessarily reflect the views of the National Science Foundation.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This material is based upon work supported by the National Science Foundation under Grants ESI-0242733 and ESI-0628272.

Note

1. These exclusions were necessary because we could not compute several item characteristic values for two-word phrases, such as word frequency and polysemy.

References

Aijmer, K., & Altenberg, B. (Eds.). (2013). *Advances in corpus-based contrastive linguistics: Studies in honour of Stig Johansson* (Vol. 54). Philadelphia, PA: John Benjamins.

Altarriba, J., Bauer, L. M., & Benvenuto, C. (1999). Concreteness, context-availability, and imageability ratings and word associations for abstract, concrete, and emotion words. *Behavior Research Methods, Instruments, & Computers*, *31*, 578-602.

Anglin, J. M. (1993). Vocabulary development: A morphological analysis. *Monographs of the Society for Research in Child Development*, *58* (Serial No. 238) , 1-166.

August, D., Branum-Martin, L., Cardenas-Hagan, E., & Francis, D. J. (2009). The impact of an instructional intervention on the science and language learning of middle grade English language learners. *Journal of Research on Educational Effectiveness*, *2*, 345-376.

Beck, I., McKeown, M. G., & Kucan, L. (2013). *Bringing words to life: Robust vocabulary instruction* (2nd ed.). New York, NY: Guilford Press.

Bensoussan, M., & Laufer, B. (1984). Lexical guessing in context in EFL reading comprehension. *Journal of Research in Reading*, *7*, 15-32.

Bergman, C., Martelli, M., Burani, C., Pelli, D. G., & Zoccolotti, P. (2006). How the word length effect develops with age. *Journal of Vision*, *6*, 999.

Brysbaert, M., Warriner, A. B., & Kuperman, V. (2013). Concreteness ratings for 40 thousand generally known English word lemmas. *Behavior Research Methods*, *46*, 904-911. doi:10.3758/s13428-013-0403-5

Carlisle, J. F., & Katz, L. A. (2006). Effects of word and morpheme familiarity on reading of derived words. *Reading and Writing: An Interdisciplinary Journal*, *19*, 669-693.

Carlisle, J. F., & Stone, C. (2005). Exploring the role of morphemes in word reading. *Reading Research Quarterly*, *40*, 428-449.

Carroll, J. B., Davies, P., & Richman, B. (1971). *The American heritage word frequency book*. Boston, MA: Houghton Mifflin.

Cassels, J. R. T., & Johnstone, A. H. (1985). *Words that matter in science: A report of a research exercise*. London, UK: Royal Society of Chemistry.

Cervetti, G. N., Barber, J., Dorph, R., Pearson, P. D., & Goldschmidt, P. (2012). The impact of an integrated approach to science and literacy in elementary school classrooms. *Journal of Research in Science Teaching*, *49*, 631-658.

Clark, E. V. (1978). Awareness of language: Some evidence from what children say and do. In A. Sinclair, R. Jarvella, & W. J. M. Levelt (Eds.), *The child's conception of language* (pp. 17-43). New York, NY: Springer-Verlag.

Clifford, G. J. (1978). Words for schools: The applications in education of the vocabulary researches of Edward L. Thorndike. In P. Suppes (Ed.), *Impact of research on education: Some case studies* (pp. 107-198). Washington, DC: National Academy of Education.

Coltheart, M. (1981). The MRC psycholinguistic database. *The Quarterly Journal of Experimental Psychology*, *33*, 497-505.

Covington, M. A. (2007). *CPIDR 3 user manual* (CASPR Research Report 2007-03). Athens: Artificial Intelligence Center, the University of Georgia. Retrieved from http://www.ai.uga.edu/caspr/

Coxhead, A. (2000). A new academic word list. *TESOL Quarterly*, *34*, 213-238.

Dale, E., & O'Rourke, J. (1981). *Living word vocabulary*. Chicago, IL: World Book/Childcraft.

De Groot, A., & Keijzer, R. (2000). What is hard to learn is easy to forget: The roles of word concreteness, cognate status, and word frequency in foreign language vocabulary learning and forgetting. *Language Learning*, *50*, 1-56.

Dijkstra, T., Martín, F. M., Schulpen, B., Schreuder, R., & Baayen, R. H. (2007). A roommate in cream: Morphological family size effects on interlingual homograph recognition. *Language and Cognitive Processes*, *20*, 7-41.

Dockrell, J. E., Braisby, N., & Best, R. M. (2007). Children's acquisition of science terms: Simple exposure is insufficient. *Learning and Instruction*, *17*, 577-594.

Gates, A. I. (1962). The word recognition ability and the reading vocabulary of second-and third-grade children. *The Reading Teacher*, *15*, 443-448.

Glass, G. V. (1976). Primary, secondary, and meta-analysis of research. *Educational Researcher*, *5*(10), 3-8.

Gries, S. T., & Newman, J. (2013). Creating and using corpora. In R. J. Podesva & D. Sharma (Eds.), *Research methods in linguistics* (pp. 257-287). New York, NY: Cambridge University Press.

Guthrie, J. T., Schafer, W. D., & Huang, C. W. (2001). Benefits of opportunity to read and balanced instruction on the NAEP. *The Journal of Educational Research*, *94*, 145-162.

Jenkins, J. R., & Dixon, R. (1983). Vocabulary learning. *Contemporary Educational Psychology*, *8*, 237-260.

Johnstone, A. H. (1991). Why is science difficult to learn? Things are seldom what they seem. *Journal of Computer Assisted Learning*, *7*, 75-83.

Kaefer, T., & Neuman, S. B. (2013). A bidirectional relationship between conceptual organization and word learning. *Child Development Research, 2013*, Article 298603. doi:10.1155/2013/298603

Kemler Nelson, D. G., O'Neill, K. A., & Asher, Y. M. (2008). A mutually facilitative relationship between learning names and learning concepts in preschool children: The case of artifacts. *Journal of Cognition and Development*, *9*, 171-193.

Kieffer, M. J., & Lesaux, N. K. (2008). The role of derivational morphology in the reading comprehension of Spanish-speaking English language learners. *Reading and Writing: An Interdisciplinary Journal*, *21*, 783-804.

Kuperman, V., Stadthagen-Gonzalez, H., & Brysbaert, M. (2012). Age-of-acquisition ratings for 30,000 English words. *Behavior Research Methods*, *44*, 978-990.

Kweon, S., & Kim, H. (2008). Beyond raw frequency: Incidental vocabulary acquisition in extensive reading. *Reading in a Foreign Language*, *20*, 191-215.

Marzano, R. J. (2004). *Building background knowledge for academic achievement*. Alexandria, VA: Association for Supervision and Curriculum Development.

Miller, L. T., & Lee, C. J. (1993). Construct validation of the Peabody Picture Vocabulary Test—Revised: A structural equation model of the acquisition order of words. *Psychological Assessment*, *5*, 438-441.

Moss, H. E., Ostrin, R. K., Tyler, L. K., & Marslen-Wilson, W. D. (1995). Accessing different types of lexical semantic information: Evidence from priming. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *21*, 863-883.

Nagy, W. E., & Anderson, R. C. (1984). How many words are there in printed school English? *Reading Research Quarterly*, *19*, 304-330.

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. E., Berninger, V. W., & Abbott, R. D. (2006). Contributions of morphology beyond phonology to literacy outcomes of upper elementary and middle-school students. *Journal of Educational Psychology*, *98*, 134-147.

Nagy, W. E., & Hiebert, E. H. (2011). Toward a theory of word selection. In M. L. Kamil, P. D. Pearson, E. B. Moje, & P. P. Afflerbach (Eds.), *Handbook of reading research* (Vol. 4, pp. 388-404). New York, NY: Longman.

National Governors Association and Council of Chief State School Officers. (2010). *Common Core State Standards: English language arts & literacy in history/social studies, science, and technical subjects—Appendix A: Research supporting key elements of the standards*. Washington, DC: Author. Available from www.corestandards.org

National Institute of Child Health and Human Development. (2000). *Report of the National Reading Panel: Teaching children to read: An evidence-based assessment of the scientific research literature on reading and its implications for reading instruction* (NIH Publication No. 00-4769). Washington, DC: U.S. Government Printing Office.

Nerlich, B., & Clarke, D. D. (2003). *Polysemy: Flexible patterns of meaning in mind and language*. Boston, MA: Walter de Gruyter.

Next Generation Science Standards Lead States. (2013). *Next generation science standards: For states, by states*. Washington, DC: The National Academies Press.

Osborne, J. (2002). Science without literacy: A ship without a sail? *Cambridge Journal of Education*, *32*, 203-218.

Oxford Dictionaries. (2010). *The Oxford dictionary of English*. New York, NY: Oxford University Press.

Paivio, A., Yuille, J., & Madigan, S. A. (1968). Concreteness, imagery and meaningfulness values for 925 nouns. *Journal of Experimental Psychology, 76*(1, Pt. 2), 1-25.

Pigada, M., & Schmitt, N. (2006). Vocabulary acquisition from extensive reading: A case study. *Reading in a Foreign Language*, *18*, 1-28.
Psychology Online Dictionary. (n.d.). Retrieved from http://psychologydictionary.org
Schwanenflugel, P. J. (1991). Why are abstract concepts hard to understand? In P. Schwanenflugel (Ed.), *The psychology of word meanings* (pp. 223-250). Mahwah, NJ: Lawrence Erlbaum.
Schwanenflugel, P. J., Akin, C., & Luh, W. (1992). Context availability and the recall of abstract and concrete words. *Memory & Cognition*, *20*, 96-104.
Seymour, S. (2006). *Volcanoes*. New York, NY: HarperCollins.
Spycher, P. (2009). Learning academic language through science in two linguistically diverse kindergarten classes. *Elementary School Journal*, *109*, 359-379.
Stallman, A. C., Commeyras, M., Kerr, B., Reimer, K., Jimenez, R., Hartman, D. K., & Pearson, P. D. (1989). Are "new" words really new? *Literacy Research and Instruction*, *29*(2), 12-29.
Strauss, U., Grzybek, P., & Altmann, G. (2005). Word length and word frequency. In P. Grzybek (Ed.), *Word length studies and related issues* (pp. 255-272). Boston, MA: Kluwer.
Tyler, A., & Nagy, W. (1989). The acquisition of English derivational morphology. *Journal of Memory and Language*, *28*, 649-667.
Varela, F. J., Thompson, E. T., & Rosch, E. (1991). *The embodied mind: Cognitive science and human experience*. Cambridge, MA: The MIT Press.
Vartanian, T. P. (2011). *Secondary data analysis*. Cary, NC: Oxford University Press.
Vygotsky, L. S. (1987). Thinking and speech. In R. Rieber & A. Carton (Eds.), *The collected works L. S. Vygotsky* (Vol. 1, pp. 39-285). New York, NY: Plenum.
Wilson, M. (2005). *Constructing measures: An item response modeling approach*. Mahwah, NJ: Lawrence Erlbaum.
Zeno, S., Ivens, S., Millard, R., & Duvvuri, R. (1995). *The educator's word frequency guide*. Brewster, NY: Touchstone Applied Science Associates.
Zipf, G. K. (1935). *The psycho-biology of language*. Boston, MA: Houghton Mifflin.

Author Biographies

Gina N. Cervetti is an associate professor in the School of Education at the University of Michigan. She studies science as a context for elementary and middle-school students' language and literacy learning.

Elfrieda H. Hiebert is president and CEO of TextProject, Inc., a not-for-profit aimed at increasing student-reading levels through appropriate texts and is also a research associate at the University of California, Santa Cruz. Her research focuses on features of text that support struggling readers and English learners.

P. David Pearson is a professor in the Language, Literacy, and Culture area within the Graduate School of Education at University of California (UC) Berkeley where he works on issues of theory, practice, and assessment of reading processes in educational settings.

Nicola A. McClung is an assistant professor in learning and instruction at University of San Francisco. Her work focuses on illuminating various factors in the environment that mitigate or exacerbate reading difficulties.