INTERNATIONAL LITERACY ASSOCIATION READING RESEARCH QUARTERLY

Reading Comprehension and Academic Vocabulary: Exploring Relations of Item Features and Reading Proficiency

Joshua F. Lawrence
Rebecca Knoph
University of Oslo, Norway

Autumn McIlraith
Texas Education Agency, Austin, USA

Paulina A. Kulesz
David J. Francis
University of Houston, Texas, USA

ABSTRACT

General academic words are those which are typically learned through exposure to school texts and occur across disciplines. We examined academic vocabulary assessment data from a group of English-speaking middle school students (N = 1,747). We tested how word frequency, complexity, proximity, polysemy, and diversity related to students' knowledge of target words across ability levels. Our results affirm the strong relation between vocabulary and reading at the individual level. Strong readers were more likely to know the meanings of words than struggling readers were, regardless of the features of the academic words tested. Words with more meanings were easier for all students, on average. The relation between word frequency and item difficulty was stronger among better readers, whereas the relation between word complexity and item difficulty was stronger among less proficient readers. Our examination of academic words' characteristics and how these characteristics relate to word difficulty across reading performance has implications for instruction.

General academic words are used across academic disciplines and more frequently in academic than nonacademic contexts (Nagy & Townsend, 2012). These words have been advanced as a promising target for instruction because of their importance for reading academic texts across disciplines (Townsend, Filippini, Collins, & Biancarosa, 2012). General academic words are particularly important for middle schoolers who encounter instructional texts that include higher proportions of lower frequency words and morphologically complex words (Hiebert, Goodwin, & Cervetti, 2018). There are many reasons these words may be difficult for adolescent readers. Unlike discipline-specific vocabulary, general academic words may not receive explicit instruction in content area classes (Hiebert & Lubliner, 2008). These words may be longer and harder to pronounce than words that students encounter in earlier grades. General academic words tend to be morphologically complex. They occur less frequently than many words learned in casual discussion. General academic words tend to have multiple related senses, some or all of which are abstract (Nagy & Townsend, 2012). In this article, we empirically examine what makes academic vocabulary difficult for middle school students. Using vocabulary and reading data from 1,747 English-speaking middle school students, in the present study, we examined which kinds of general academic words are hard for students and examined the relation between lexical features of items and item difficulty across the continuum of reading performance.

INTERNATIONAL LITERACY ASSOCIATION

Reading Research Quarterly, 0(0)
pp. 1–22 | doi:10.1002/rrq.434

Empirical Measures of Lexical Dimensions

Quantitative lexical measures have proliferated in the last decade. On the one hand, new measures have allowed researchers to test new models of how specific linguistic features relate to lexical processing and especially lexical access.[1] On the other hand, the proliferation of measures has made it difficult to generalize across studies using different word metrics that are believed to measure the same construct. As a practical matter, it is impossible to model all the competing lexical measures simultaneously or argue that one particular selection strategy is definitively better than another. Thus, as a preliminary step in studying factors that affect item performance on vocabulary tests, we made use of prior research to create a reduced feature set for inclusion in the models. This approach both reduced potential bias introduced by our measure selection process and helped us communicate our results to an audience who may be unfamiliar with (and potentially uninterested in) the details of the specific lexical measures. We began with 22 empirical word characteristics, each of which had clear documentation and had been used in earlier research. We excluded behavioral measures, such as age-of-acquisition and abstractness ratings, because we intended to use resulting factor scores as independent variables to model assessment and other behavioral data. Recent research (Knoph, Lawrence, & Francis, 2021) on these features using a set of high-frequency words (from the General Service List [GSL] developed by West, 1957) and the general academic words that are the focus of this article (from the Academic Word List [AWL] developed by Coxhead, 2000) identified five correlated factors: complexity, proximity, frequency, diversity, and polysemy. Next, we provide a brief overview of research related to each of these factors.

Vocabulary and Reading

Reading comprehension is the process of extracting and constructing meaning from print when a reader interacts with a text for a specific purpose or activity (RAND Reading Study Group, 2002). This process supports word learning by providing students with contextualized uses of new words but, at the same time, requires that readers have sufficiently developed orthographic, phonological, and semantic word knowledge (Perfetti & Hart, 2002). It is not surprising, then, that reading researchers have consistently found strong correlations between student performance on vocabulary and reading comprehension assessments (Cromley & Azevedo, 2007; Joshi, 2005; Joshi & Aaron, 2000; McKeown, Beck, Omanson, & Perfetti, 1983; Quinn, Wagner, Petscher, & Lopez, 2015; Tannenbaum, Torgesen, & Wagner, 2006; Wagner et al., 1997), across many language-learning contexts (Kieffer & Box, 2013; Qian, 2002; Rydland, Aukrust, & Fulland, 2013), and across age groups (Braze et al., 2016; Quinn et al., 2015; Snow, Porche, Tabors, & Harris, 2007). However, the relative importance of component skills used in reading change as students age.

Hoover and Gough (1990) showed that decoding skills are more related to reading comprehension in younger students but that verbal ability is more associated with reading ability in later grade levels. The simple view of reading also has implications for thinking about what might make a word difficult for students: The words that students find challenging to learn may vary in part as a function of their reading ability. For instance, less proficient readers who struggle with decoding skills may find orthographically complex words hard to master, even though this dimension may not relate to word difficulty as strongly among more skilled readers. As such, item performance may be jointly determined by reader ability and word features. To examine the joint influence of student and word features, in the current study, we examined item difficulty as a function of individual reading ability and word-level characteristics simultaneously. We also explored interactions to see if some words are more challenging or more manageable across ranges of reading ability.

Complexity

Word complexity is the orthographic and morphological complexity of a word. The word *feline* may be more challenging for some students to learn than the word *cat* simply because *feline* is longer and more complex. Complexity can be measured by the number of syllables, the number of letters, or the number of morphemes and is related to individual differences in vocabulary learning (Goodwin & Cho, 2016). In general, words with more letters take longer to process and are read more slowly than shorter words (for a review, see New, Ferrand, Pallier, & Brysbaert, 2006). However, there is a complicated relation between orthography and phonology in English, so the consistency and granularity of letter–sound mapping must also be considered (Ziegler & Goswami, 2005). The presence of clusters of consonants, for example, can slow down word reading in younger readers (Olson, Forsberg, Wise, & Rack, 1994), and clusters of vowels can result in less accurate decoding (Gilbert, Compton, & Kearns, 2011). In addition to phonological and orthographic considerations, the presence of multiple morphemes in a word can facilitate reading time and accuracy (Carlisle & Stone, 2005; Deacon, Whalen, & Kirby, 2011), especially if the base morpheme is higher in frequency than the derived word. These features not only affect word recognition but also impact access to meaning (Goodwin & Cho, 2016).

Proximity

The phonological or orthographic proximity of words can be measured by their overlap in letters or phonemes.

Similarly, word forms that share phonemic patterns or letter sequences with many others reside in denser neighborhoods than words with unusual forms. Both phonological and orthographic neighborhood density have facilitative effects on visual word recognition, lexical decision, and naming tasks. Coltheart's *N* (Coltheart, Davelaar, Jonasson, & Besner, 1977) is a measure of orthographic overlap, defined as the number of words that can be created by substituting a single letter in the original word (e.g., *rat*, *sat*, *car*, and *cab* are all neighbors of *cat*). Recently developed metrics have expanded this definition to include additions, subtractions, transpositions, and substitutions. Yarkoni, Balota, and Yap (2008) proposed a metric known as orthographic Levenshtein distance, defined as the number of operations (insertions, deletions, and substitutions) necessary to transform one word form to another. OLD20, the mean Levenshtein distance from a word to its 20 closest neighbors, then becomes another orthographic neighborhood density metric. It should be noted, however, that simpler words are often those with the densest neighborhoods. This is demonstrated by the fact that it is easy to think of near neighbors for the word *cat* but much harder to think of near neighbors for the word *necessarily*. As a result, proximity measures have loaded on the complexity factor rather than with other neighborhood relatedness measures in previous studies (see Brysbaert, Mandera, McCormick, & Keuleers, 2019; Yap, Balota, Sibley, & Ratcliff, 2012).

Frequency

Kučera and Francis (1967) used punch cards to tabulate word frequency using IBM computers in creating what has become known as the Brown University Standard Corpus of Present-Day American English (or just Brown Corpus). Because word frequency measures from sufficiently large and diverse samples generalize well, these measures can be used as a proxy for the relative number of encounters a learner may have had to specific English words. Kuperman, Stadthagen-Gonzalez, and Brysbaert (2012) found that Living Word Vocabulary levels, which indicate the grade level at which a word is widely known (Dale & O'Rourke, 1981), correlate strongly with item frequency as estimated with the Brown Corpus ($r = -.69$; Kučera & Francis, 1967). Biemiller and Slonim (2001) tested 100 words from each Living Word Vocabulary level and found a strong relation between word frequency and the grade level at which 50% of students knew a word ($r = -.57$). Age of acquisition is similar to difficulty in that it estimates the age at which a learner first masters a word. Kuperman et al. found that age-of-acquisition estimates based on adult self-reports correlated ($r = -.64$) with the word frequency in the Brown Corpus (see also Breland, 1996; L.T. Miller & Lee, 1993). Findings like these motivated Coxhead (2000) to exclude the high-frequency words from her AWL; high-frequency words are likely already known or can be learned independently.

Diversity

Whereas it is relatively easy to count the number of occurrences of a word, it is harder to quantify the diversity of its usages within and across texts. Researchers have used latent semantic analysis within texts to create the semantic diversity measure, which estimates how distinct word usages are at the local level (Hoffman, Lambon Ralph, & Rogers, 2013). This measure quantifies the diversity of words that occur adjacent to or near a target word. For example, the word *aquarium* has a low semantic diversity rating, indicating that it appears next to a stable set of collocates (e.g., *fish*). A related measure, contextual diversity, is a measure of the number of times a word appears across text selections that make up a text corpus, regardless of the document-level features (Adelman, Brown, & Quesada, 2006; Brysbaert & New, 2009), although contextual diversity could alternatively be considered a way of measuring frequency.

Educational researchers have taken the additional step of categorizing the documents that make up a corpus by academic discipline and analyzing the occurrence of words in documents across categories. Zeno, Ivens, Millard, and Duvvuri (1995) counted word occurrences across text selections classified by the academic category of the texts in which they appear to create dispersion estimates. Coxhead (2000) analyzed a 3.5 million–word corpus containing over 400 texts that fell into the categories of arts, commerce, law, and science. After refining target words based on frequency, she excluded word families that did not occur in each of the four disciplinary areas at least 10 times. The resulting list of 570 word families provides much better coverage of academic texts than an alternative list based only on frequency. The resulting AWL has been touted in influential instructional books (Beck, McKeown, & Kucan, 2013) and has been referenced in creating vocabulary interventions for middle school students (Lawrence, Crosson, Paré-Blagoev, & Snow, 2015; Lesaux, Kieffer, Kelley, & Harris, 2014).

Polysemy

We say a word is polysemous when it has several related senses. A recent analysis of 13,783 nouns and 8,998 verbs using results from WordNet found that the nouns average 2.9 senses ($SD = 2.4$) each and that the verbs average 4.3 senses ($SD = 4.5$) each (Lawrence et al., 2021). General academic words tend to have many senses. For instance, according to WordNet, the word *retain* has four meanings, and the word *obtain* has three. In contrast, *disproportionately* only has two meanings, and *controversy* has one.

In English, word forms with more senses are more frequent than word forms with fewer senses ($r = .53$; Hoffman et al., 2013). A good deal of evidence demonstrates that polysemous words are accessed more rapidly than words with single senses (Azuma & Van Orden, 1997; Borowsky & Masson, 1996; Hino & Lupker, 1996). Homophones, in contrast, are word forms that have two or more distinct meanings (e.g., *bank* meaning the side of a river vs. a place for money). These words are much less frequent in English and are processed less efficiently in speeded lexical decision tasks (Beretta, Fiorentino, & Poeppel, 2005; Rodd, Gaskell, & Marslen-Wilson, 2002) and semantic categorization tasks (Hino, Lupker, & Pexman, 2002). Given the ubiquity of polysemy in English and that sense disambiguation is essential to skilled reading, it is surprising that research into polysemy with educationally relevant outcomes has been rare. One study found that, controlling for frequency, polysemous scientific words are more difficult for elementary-age students at pretest. However, polysemous target words were learned more effectively during the school year (when they were introduced as part of a language-rich science curriculum). Controlling for pretest scores, the number of target word meanings was a better predictor of posttest knowledge than word frequency measures were (Cervetti, Hiebert, Pearson, & McClung, 2015). In contrast, Hiebert, Scott, Castaneda, and Spichtig (2019) did not find a relation between target word knowledge and the number of word senses and meanings in an analysis of synonym task data from students across grades 2–12. These mixed results suggest that this may be a productive space for further study.

Hypothetical Relations Between Vocabulary and Reading

Explanations of the possible mechanisms underlying the correlations between measures of vocabulary knowledge and reading ability have focused on the importance of efficient lexical access, the importance of knowing a word encountered by readers in target passages, the relation between word knowledge and world knowledge, and the correlations across verbal skills (Anderson & Freebody, 1981; Quinn et al., 2015). Here, we provide a brief overview of these hypotheses, none of which is exclusive of the others.

Efficient Lexical Access

Accurate and efficient retrieval of word knowledge is essential for skilled reading (Mezynski, 1983; Perfetti, 1988), a point emphasized in text comprehension models that focus on efficient lexical access (Perfetti & Hart, 2002; Perfetti & Stafura, 2014). There are both individual differences in lexical access and differences in access speeds associated with lexical characteristics. Not surprisingly, efficient lexical retrieval (measured by speeded lexical decision tasks) at the individual level correlates with subject vocabulary scores (Yap et al., 2012). There are also word-level differences that influence speeded lexical retrieval tasks. For instance, less complex words and high-frequency words are retrieved more efficiently (see, e.g., Brysbaert & New, 2009). Interestingly, words with multiple senses are also retrieved more efficiently, possibly because the process of learning words with multiple senses provides the learner with the opportunity to compare and integrate usages across encounters. There is much less known about how word characteristics relate to student performance on educationally relevant tasks. However, if words that are more efficiently accessed are also better known, orthographically complex words will be more challenging, whereas frequent words with more meanings will be easier.

Instrumental Word Knowledge

The instrumentalist perspective is based on the finding that when a reader knows more words in a specific passage, the reader comprehends it better (Schmitt, Jiang, & Grabe, 2011). Vocabulary training produces improved comprehension when the target words are in the tested comprehension passages (Beck, Perfetti, & McKeown, 1982; McKeown et al., 1983; for a review, see Wright & Cervetti, 2017). Unfortunately, these results can be hard to translate into instructional practice across instructional contexts. Given the volume and diversity of texts that students are expected to read across classes in secondary schools, it can be challenging to provide tailored prereading support for unknown words. Instead, some researchers have resorted to examining textual corpora to identify frequent, widely dispersed words that students are most likely to encounter, and which may therefore be good candidates for instruction (Coxhead, 2000; Hiebert et al., 2018; Praninskas, 1972). However, vocabulary interventions usually analyze data aggregated at the individual, class, or school level: they do not shed light on the efficacy of target word selection strategies. Intervention research has demonstrated that academic vocabulary can be improved through targeted instruction (Lawrence, Francis, Paré-Blagoev, & Snow, 2017; Lesaux, Kieffer, Faller, & Kelley, 2010; Pany, Jenkins, & Schreck, 1982). However, meta-analyses of vocabulary interventions have suggested only moderate effects on passage comprehension as measured by researcher-developed instruments, and no impact on standardized reading measures (Elleman, Lindo, Morphy, & Compton, 2009; Stahl & Fairbanks, 1986).

Word Knowledge and World Knowledge

The knowledge hypothesis is predicated on the idea that knowing a word entails knowing something about the world and that the more learners know about the world, the better their reading comprehension. For instance, knowledge of domain- and topically relevant words predicted improvement in scenario-based reading measures (McCarthy et al., 2018). Among general academic vocabulary, there may also be words that help students understand the world or the way things can be related to each other. Knowledge of these concepts may relate to the skilled comprehension of a particular text passage, even if these words do not appear in the passage. For instance, a class of academic words known as connectives allows students to understand and make connections across ideas (Crosson, Lesaux, & Martiniello, 2008). Thus, knowledge of the word *notwithstanding* might be a marker of a student's understanding of how integrative arguments work. This understanding of integrative argumentation might help the student comprehend a text in which such a relation is implied, even if the word *notwithstanding* is not used in the text to signal the nature of the relation.

Words with multiple senses mark world knowledge as well. Words acquire these multiple senses through the countless ways their usage is broadened, extended, and refined (Aitchison, 2012). Students who know two or more senses of the same word have the opportunity to reflect on these relations and on the abstract conceptual relations that may link related meanings. Nagy and Townsend (2012) suggested that one class of these relations, grammatical metaphor, is one of the defining characteristics of academic language. Grammatical metaphor extends the range of a word's most frequent or etymologically primary meaning by metaphorical usage (e.g., *boils down to*), nominalization (employing derived inflections or zero derivation), or idiomatic phrasing. Grammatical metaphor is ubiquitous in academic writing and "is the largest diversion from social/conversational language and presents the most significant issue for students" (Nagy & Townsend, 2012, p. 94). Knowledge of polysemous words may support students' understanding of linguistic and conceptual relations that have broad utility.

Verbal Skill and Metalinguistic Ability

General factors can explain high correlations across discrete cognitive skills (Spearman, 1904; Tucker-Drob, 2009). Carroll (1941) argued that verbal ability is connected to how well one can infer and retain the meanings of newly encountered words (see also Sternberg & Powell, 1983). Tunmer and Herriman (1984) identified metalinguistic awareness as a similarly general verbal ability that learners use to "reflect on and manipulate the structural features of spoken language" (p. 136). Nagy (2007) pointed to metalinguistic awareness in explaining individual differences in vocabulary learning and retention rates. Whereas some researchers have pointed to a common underlying cause, such as metalinguistic awareness, or general verbal ability, to account for the correlation between reading comprehension and vocabulary knowledge, others have linked vocabulary knowledge and reading comprehension in a relation of reciprocal causality (Stanovich, 1986; Verhoeven, van Leeuwe, & Vermeer, 2011). The reciprocity argument views vocabulary as causally implicated in understanding language in written form, and exposure to word usage through written language as one way in which word meanings are acquired.

We argue that interaction between readers' abilities and word features in predicting word knowledge is not directly compatible with the spurious correlation view without modification, whereas these interactions are more easily explained through reciprocal causality models. Although these two views of the basis for the correlation between vocabulary and reading imply quite different causal models for the role of vocabulary in reading, that vocabulary knowledge and reading comprehension are strongly correlated is not in dispute. The magnitude of interindividual differences complicates any investigation of word-level features which might seek to average over individuals to get at relations at the word level and suggests the need for intensive data collection that is both wide (i.e., many words) and deep (i.e., many individuals), with many covariates at both the word and person levels. The present study was not intended to arbitrate these different views of the correlation between reading and vocabulary but to determine which characteristics of academic words are associated with item difficulty and to examine some characteristics of readers that might affect vocabulary knowledge and possibly alter the relation between word characteristics and item difficulty.

Research Questions

General academic word knowledge is strongly related to reading comprehension (Townsend et al., 2012; Lawrence, Hagen, Hwang, Lin, & Lervåg, 2019). However, little is known about which lexical features may make an academic word difficult for students to learn or if the word features that make these words challenging for students are consistent across students at different reading performance levels. Therefore, three research questions guided our study:

1. What are the characteristics of middle school readers (measured via reading ability, socioeconomic status [SES], gifted and talented education [GATE] status, and grade level) that account for individual differences in vocabulary knowledge?
2. What is the relation between features of academic vocabulary (measured via item frequency, complexity, proximity, polysemy, and diversity) and item difficulty on a test of academic word knowledge for middle school students?
3. How does student knowledge of words with different features relate to reading ability? Specifically, to what extent is the influence of word features on item difficulty different for good and poor readers?

Method

To answer our research questions, we needed to model item difficulty with word- and person-level data and explore interactions. We now present a technical description of the approach we used. This description is essential for scientific replication purposes, although readers with more substantive interests may wish to skip the next couple of paragraphs. We used explanatory item response theory (EIRT) models and examined middle school students' performance on a test of academic vocabulary. EIRT models are multivariate, cross-classified random-effects models that can be used to jointly explain differences in person ability and item difficulty by modeling item responses on a test in terms of (a) the effects of student characteristics on a latent ability (θ_p; in our case, vocabulary knowledge as measured by academic words), (b) the effects of word features on item difficulty (β_i; difficulty of an item designed to measure the latent ability; De Boeck & Wilson, 2004), and (c) cross-level interactions between person characteristics and word features. These models are particularly advantageous when one is interested in investigating moderating effects of test features (in our case, item/word features) on relations between students' characteristics and students' performance on an outcome measure (i.e., student–test interactions). Although interaction effects often account for a small proportion of variance explained in EIRT models (controlling for main effects of student characteristics and test features), interaction effects provide unique insights about how the same item feature affects students differently depending on their individual characteristics. Importantly, these insights cannot be easily examined when looking at interaction effects based on composite scores.

The specific EIRT models used in the current study are well suited for binary outcome data. A general mathematical formulation of the EIRT model proposed for the present study can be found in Kulesz, Francis, Barnes, and Fletcher (2016). We applied the binary form of the model because item responses to test items had a correct/incorrect format (missing values were coded as incorrect responses). We used a multivariate structure because item difficulty was simultaneously modeled for all items. We used a cross-classified random-effects structure to deal with dependencies among the responses to items, as these dependencies result from administering all items to all students and students responding to all items. Treating items as random effects further improves the estimation of the model and has the inferential advantage of treating items as being sampled from a universe of items. Thus, inferences about item features are not specific to the sample of items but to the universe of items from which the specific items have been sampled. The specific cross-classified structure employed in the present study comprised two levels: The first level was responses to items (dummy variables where 0 = *incorrect*, and 1 = *correct*), and the second level was item and student parameters, which are completely crossed in this design because all students completed all test items. Thus, we considered item responses cross-classified within a person and item. In all EIRT models, we standardized continuous student characteristics and word features to provide a correct and meaningful interpretation of parameter estimates.

We estimated the models in several steps. Step 1 fits an unconditional variance components model (model 1). We compared the unconditional variances from model 1 with residual variances of subsequent models that included student characteristics and word features, to estimate the variance explained by student characteristics and word features. Step 2 incorporated predictors of student ability, including grade, reading comprehension, SES status, and GATE status, that were sequentially entered in models 2–4. We used sequential entry of student characteristics to the models to estimate unique variance explained by different student characteristics. In step 3 (model 5), we added word features to model 1 (frequency, complexity, proximity, polysemy, and diversity) to explain item difficulty in the absence of student characteristics. In step 4 (model 6), we integrated student characteristics from model 4 and word features from model 5 to explain student ability and item difficulty, respectively, without inclusion of interactions between student characteristics and word features. In models 7–11, we extended model 6 by adding interaction terms individually. We added the interaction effects one at a time to examine their statistical significance in the absence of other interaction terms. In the final model, model 12, we included predictors from model 6 and interaction effects of reading comprehension with all word features (five interaction terms) to assess the importance of interaction terms relative to one another. Because the interaction terms are correlated with one another and the main effect terms, examining them individually and in conjunction with one another allowed us to evaluate their individual and joint contributions to the prediction of word difficulty and student ability. All EIRT models were estimated in R using the glmer function of the lme4 package (Bates et al., 2021) using nonlinear optimization of the Nelder–Mead and bound optimization by quadratic approximation methods.

Student Sample

Students who contributed data to this study attended schools participating in the randomized efficacy trial of the Word Generation program (Strategic Education Research Partnership, 2021). The students were recruited from 12 middle schools from a large urban school district in California. The students participating in the initial study included a diverse range of language speakers. Linguistic diversity presented a challenge in this analysis because cognate advantages varied across language–word dyads. Therefore, we restricted this analysis to all monolingual English speakers from the initial study who contributed valid data. Our analytic sample of monolinguals is not typical of the district because only 34% of students in participating schools were monolingual English speakers. The monolingual students in our sample were similar to other monolingual students in the district in being less likely than their peers to be eligible for free or reduced-price lunch ($M_{\text{monolingual\_English}}$ = 37% vs. $M_{\text{nonmonolingual\_English}}$ = 64%). Forty-six percent of the students in our analytic sample were identified as being enrolled in the GATE program. This rate was similar to the district's identification rate (41%). Our analytic sample consisted of students in grades 6 (28%), 7 (38%), and 8 (34%; see Table 1). Performance levels on the Comprehension subtest (which has been nationally normed) of the Gates–MacGinitie Reading Tests (GMRT) indicate that our sample was typical to somewhat above average in reading performance relative to students in similar grades nationwide.

Student Measures

In addition to information about home language (which we used to determine the analytic sample), the district also provided information about students' grade level, eligibility for free or reduced-price lunch, and identification for the district's GATE program.[2]

Grade-Level Cohort

To control for differences across grade levels, we assigned values for dummy variables to each of the students according to their grade level.

SES Status

We used eligibility for free or reduced-price lunch as an indicator of students' SES status. We created a student-level dummy variable to indicate students who received free or reduced-price lunch (SES status = 1) and those who did not (SES status = 0).

GATE

The district used eight categories, such as "specific academic achievement" and "high potential," to identify students as gifted. The GATE variable indicates whether students were identified as being enrolled in the GATE program (GATE = 1) or not (GATE = 0).

Reading Comprehension

We used the Comprehension subtest of the GMRT to measure overall reading comprehension. Sixth-grade students completed level 6 of the assessment. Seventh- and eighth-grade students completed level 7/9, as suggested by the testing manual. The GMRT is a nationally normed test composed of 48 multiple-choice questions. Each item relates to a short reading passage. Kuder–Richardson formula 20 reliability coefficients were high (.92 for level 6 and .91 for level 7/9; Maria, Hughes, MacGinitie, MacGinitie, & Dreyer, 2007). We used the extended scale scores in this analysis because they place scores from different GMRT test levels onto a common scale, which allows progress in reading to be tracked over time and across grades on a single, continuous scale. For the present study, the extended scale scores allowed us to place students' performance on levels 6 and 7/9 of the GMRT on a common scale. The internal reliability of the test in our sample was high (Cronbach's α = .91). The extended scale scores ranged from 361 to 643 (*M* = 536.3, *SD* = 35.8) in our sample.

Academic Vocabulary Test

This researcher-developed test was group administered to measure students' academic vocabulary knowledge. Students were presented with target words placed within a neutral context suggesting a part of speech and were then asked to choose from four options, with the correct option indicating the target word's synonym. For instance, the key for the target word *suspended* was "The tests were suspended," and the choices were (a) *allowed*, (b) *hard for students*, (c) *suspicious*, and (d) *stopped for a while*. Target words were general academic words, and stems reference common senses of the target words. There were 50 items administered each year for two years. We included 22 anchor items both years, so this analysis uses information for 78 different words. These words were mostly taken

TABLE 1
Reading Score, Total Academic Vocabulary Score, GATE Identification Rate, and Percentage of Students Eligible for Free or Reduced-Price Lunch

Grade	Reading score *M* (*SD*)	Academic vocabulary score[a] *M* (*SD*)	GATE *M* (*SD*)	SES *M* (*SD*)
6 (*n* = 492)	514.3 (40.1)	32.7 (9.8)	.45 (.5)	.37 (.5)
7 (*n* = 661)	537.4 (40.5)	35.4 (10.4)	.48 (.5)	.39 (.5)
8 (*n* = 594)	550.1 (43.0)	37.7 (9.6)	.45 (.5)	.35 (.5)
Total (*N* = 1,747)	535.5 (43.9)	35.4 (10.1)	.46 (.5)	.37 (.5)

Note. GATE = enrollment in the Gifted and Talented Education program; Reading score = the extended scale score on the Comprehension subtest of the Gates–MacGinitie Reading Tests; SES = eligibility for free or reduced-price lunch.
[a]The maximum score is 50.

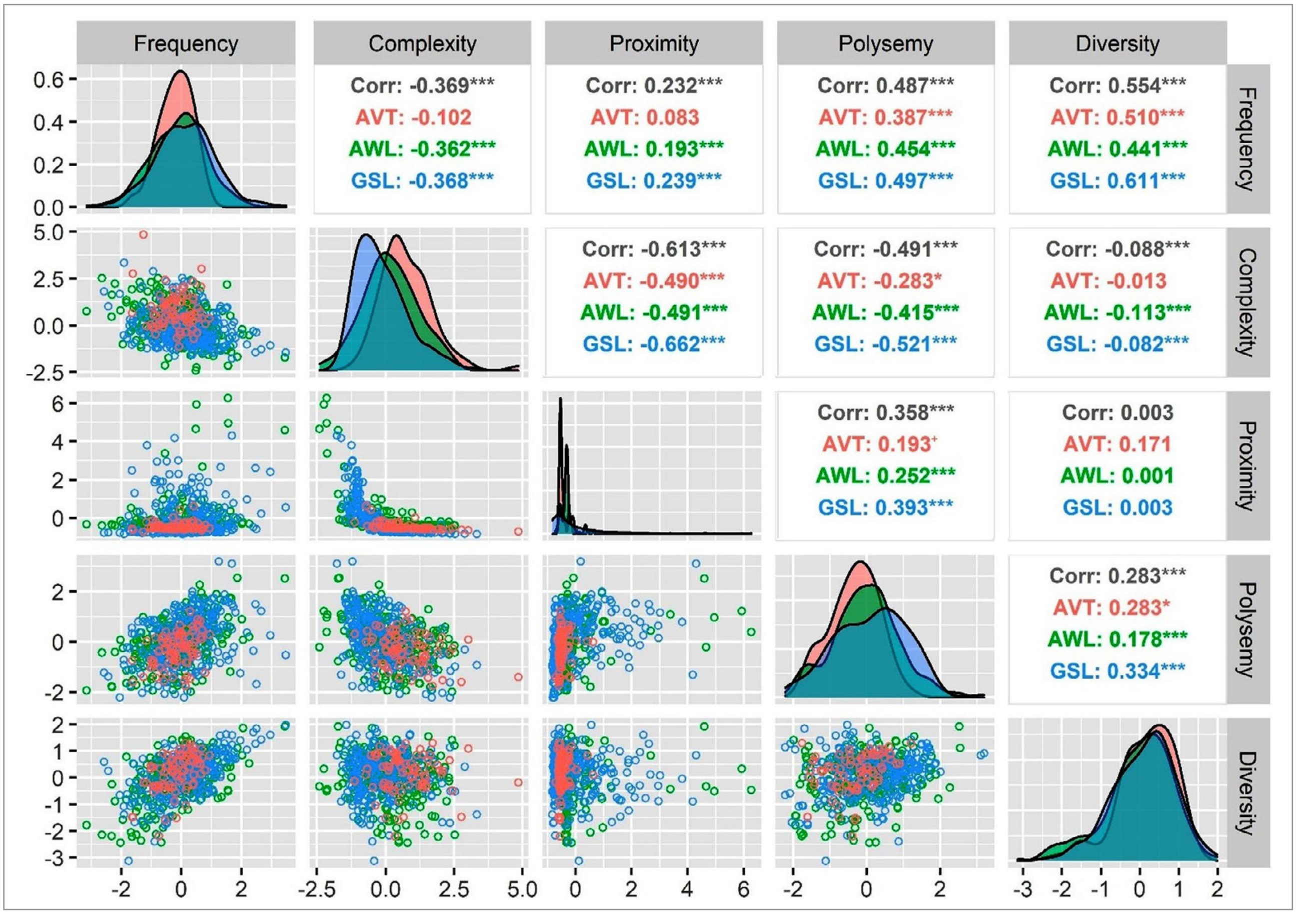
from the AWL (Coxhead, 2000) and seem to represent the class of words on the AWL with respect to word characteristics, as we subsequently discuss in detail. Within-sample internal consistency reliabilities for grades 6–8 ranged from .81 to .93. All Academic Vocabulary Test forms that were developed by the Word Generation research team can be found in the IRIS digital depository (https://www.iris-database.org/).

Factor Scores

Insofar as the words on the Academic Vocabulary Test are considered a sample of academic words, it is important to consider how the sample of 50 words included on the test relate to the universe of academic words. As such, we considered their characteristics in comparison with the characteristics of words from Coxhead's (2000) AWL and also West's (1957) GSL, a list of approximately 2,000 high-frequency words considered important for basic understanding of the English language.

We fitted exploratory factor models with a set of high-frequency words (*n* = 2,136; GSL), and general academic words (*n* = 1,082; AWL). Inspection of the factor scores provides some useful information about the generalizability of our findings to other academic and nonacademic words. We used the factor structure derived from the analysis of the AWL and GSL to create factor scores for the Academic Vocabulary Test words. These factor scores are used in the analyses reported here (see Tables A1 and A2 in the Appendix for a complete list of the variables used in determining the factors and estimating the beta weights used to estimate the factor scores). Figure 1 presents distributions of and correlations among the five factor scores,[3] color-coded according to the words' source. Notice that the distribution of each factor for our sample (Academic Vocabulary Test) largely overlaps with the distribution of a random sample of 500 words from the larger class of academic words (AWL). Similarly, the correlations across factors are similar in our set of words and the larger set of academic

FIGURE 1
Correlations and Density Plots for the Word Feature Factor Scores



Note. AVT = Academic Vocabulary Test; AWL = Academic Word List; Corr = correlation; GSL = General Service List. Correlation coefficients above the diagonal include all 78 words on the AVT, 1,082 words on the AWL, and all 2,136 words on the GSL. The diagonal includes density plots color-coded by list: red for the AVT words, green for the AWL words, and blue for the GSL words. Scatterplots below the diagonal contain a random sample of 500 words for the AWL and GSL each, plus the entire set of AVT words, using the same color scheme. The color figure can be viewed in the online version of this article at http://ila.onlinelibrary.wiley.com.
$^{+}$*p* < .10. **p* < .05. ****p* < .001.

words. These results gave us confidence that the findings presented here generalize to other academic words. We also present information about these factor scores for a random sample of 500 words from a set of high-frequency words (the GSL). Not surprisingly, these words appear to have higher frequencies and are less complex than academic words. Still, the relations between factors in the GSL are similar to those in the AWL sample, meaning that to some extent, our findings here may generalize to nonacademic words. For a full discussion, see Knoph et al. (2021).

Figure 1 can also help in understanding the relations between factors. Note the strong negative correlation between complexity and proximity ($r = -.513$, $p < .001$), which we expected given the large number of relatively simple words with related forms in English (e.g., *bat*, *cat*). Note also the relatively high correlation between polysemy and frequency ($r = .370$, $p < .001$) and between polysemy and diversity ($r = .283$, $p < .001$), which we expected because polysemous word forms have more semantic utility for writers. Clearly, the five factor scores that we used to summarize the characteristics of words and their meanings are correlated, or overlapping. As such, the individual factors will account for both unique and shared variance in predicting word difficulty in our EIRT models. It is important to recognize that the coefficient attached to a factor in any model that involves multiple factor scores will reflect both the relation of the factor to word difficulty and to the other factor scores. In the analyses that follow, we have not attempted to identify the best prediction model of a given size but rather to understand each feature's possible contribution in light of the contribution of other factors, as well as to examine possible interactions with characteristics of readers. Still, even with these 22 characteristics reduced to only five dimensions, there is still a rich diversity in the data trends across word forms, as seen in the example words presented in Table 2. Take the words *controversy* and *retain*, for example. *Controversy* is more frequent (frequency = 0.096) and complex (complexity = 2.085) than the word *retain* (frequency = −0.048; complexity = −0.526). Given that *retain* is less complex, it is not surprising that it has more orthographic and phonological neighbors (proximity = 0.429). Interestingly, *retain* has a higher polysemy rating (0.015) than *controversy* (−1.529) even though *controversy* is more frequent.

Results

EIRT Models

All models are based on the analysis of binary test items using a logit link function. Thus, model parameters estimate the effect of a particular feature on the log odds of answering an item correctly, either via an effect on person ability or an effect on item easiness. Tables 3 and 4 contain estimates of logistic regression parameters and their standard errors for models involving (a) only main effects of student characteristics (models 2–4), (b) only main effects of word features (model 5), (c) main effects of student characteristics and word features (model 6), and (d) interaction effects of student reading ability and word features (models 7–12).

Table 5 provides fit indices and random effects for all 12 models. Each regression parameter describes the difference in log odds for a unit change in the student characteristic or word feature associated with the regression parameter. Bearing in mind that we standardized all continuous predictors for inclusion in the models, a unit change in the associated variable implies a change of one standard deviation. For dichotomous student predictors (e.g., participation in the GATE program) in models 2–4, the regression parameter describes the difference in mean log odds of correctly answering an item of average item easiness for the group

TABLE 2
Example Academic Vocabulary Test Words and Factor Scores

Word	Frequency	Complexity	Proximity	Polysemy	Diversity
retain	−0.048	−0.526	0.429	0.015	0.834
controversy	0.096	2.085	−0.613	−1.529	0.259
circumstances	0.668	3.017	−0.649	0.079	1.095
concept	0.789	0.099	−0.366	−1.246	0.273
constrain	−1.638	0.658	−0.546	−0.744	−1.495
disproportionately	−1.269	4.850	−0.703	−1.414	−0.182
equity	−0.166	0.006	−0.537	−0.306	−1.525
maintained	0.286	0.736	−0.543	1.208	1.284
obtain	0.516	−0.399	−0.519	−0.466	0.891
subsequent	0.135	1.723	−0.598	−1.757	1.299

TABLE 3
Fixed Effects for the Main Effects Models

	Model 1		Model 2		Model 3		Model 4		Model 5		Model 6	
Fixed effect	***b***	***SE***	***b***	***SE***	***b***	***SE***	***b***	***SE***	***b***	***SE***	***b***	***SE***
Intercept	1.39	0.17	0.98	0.18	1.52	0.17	1.38	0.17	0.90	0.13	1.38	0.15
Grade 7			0.41***	0.08	−0.21***	0.05	−0.13**	0.05			−0.13***	0.05
Grade 8			0.75***	0.08	−0.20***	0.05	−0.07	0.05			−0.07	0.05
Reading					1.17***	0.02	1.00***	0.03			1.00***	0.03
GATE							0.37***	0.05			0.36***	0.05
SES							−0.25***	0.04			−0.25***	0.04
Frequency									0.12	0.14	0.10	0.16
Complexity									−0.21	0.15	−0.24	0.18
Proximity									−0.14	0.15	−0.15	0.17
Polysemy									0.35**	0.13	0.39*	0.16
Diversity									0.25	0.14	0.23	0.15

Note. N = 1,747 for models 1–6. *b* = log odds; GATE = enrollment in the Gifted and Talented Education program; Reading = the extended scale score on the Comprehension subtest of the Gates–MacGinitie Reading Tests; *SE* = standard error of log odds; SES = eligibility for free or reduced-price lunch.
*p < .05. **p < .01. ***p < .001.

TABLE 4
Fixed Effects for the Interaction Effects Models

	Model 7		Model 8		Model 9		Model 10		Model 11		Model 12	
Fixed effect	***b***	***SE***	***b***	***SE***	***b***	***SE***	***b***	***SE***	***b***	***SE***	***b***	***SE***
Intercept	1.38	0.15	1.38	0.15	1.38	0.15	1.38	0.15	1.37	0.15	1.39	0.15
Grade 7	−0.13**	0.05	−0.13**	0.05	−0.13**	0.05	−0.13**	0.05	−0.13**	0.05	−0.13**	0.05
Grade 8	−0.07	0.05	−0.07	0.05	−0.07	0.05	−0.07	0.05	−0.07	0.05	−0.07	0.05
Reading	1.01***	0.03	1.01***	0.03	1.01***	0.03	1.00***	0.03	1.01***	0.03	1.02***	0.03
GATE	0.37***	0.05	0.37***	0.05	0.37***	0.05	0.36***	0.05	0.37***	0.05	0.37***	0.05
SES	−0.25***	0.04	−0.25***	0.04	−0.25***	0.04	−0.25***	0.04	−0.25***	0.04	−0.25***	0.04
Frequency	0.12	0.16	0.11	0.16	0.11	0.16	0.10	0.16	0.10	0.16	0.13	0.16
Complexity	−0.24	0.18	−0.23	0.18	−0.26	0.18	−0.24	0.18	−0.23	0.18	−0.27	0.18
Proximity	−0.16	0.17	−0.15	0.17	−0.15	0.17	−0.15	0.17	−0.13	0.17	−0.15	0.17
Polysemy	0.39*	0.16	0.40*	0.16	0.38*	0.16	0.39*	0.16	0.39*	0.16	0.38*	0.16
Diversity	0.23	0.15	0.23	0.15	0.23	0.15	0.24	0.15	0.23	0.15	0.23	0.15
Frequency × Reading	0.07***	0.01									0.07***	0.01
Polysemy × Reading			0.03**	0.01							−0.01	0.01
Complexity × Reading					−0.08***	0.01					−0.07***	0.01
Diversity × Reading							0.02	0.01			−0.01	0.01
Proximity × Reading									0.06***	0.01	0.02	0.01

Note. N = 1,747 for models 7-12. *b* = log odds; GATE = enrollment in the Gifted and Talented Education program; Reading = the extended scale score on the Comprehension subtest of the Gates–MacGinitie Reading Tests; *SE* = standard error of log odds; SES = eligibility for free or reduced-price lunch.
*p < .05. **p < .01. ***p < .001.

TABLE 5
Computed Fit Indices and Random Effects

				Person side		Item side	
Model	AIC	BIC	Deviance	Variance (*SE*)	Variance reduction	Variance (*SE*)	Variance reduction
1	76,363.8	76,391.9	76,357.8	1.69 (1.30)		1.36 (1.17)	
2	76,283.9	76,330.7	76,273.9	1.61 (1.27)	0.05	1.36 (1.17)	0
3	74,428.8	74,484.9	74,416.8	0.45 (0.67)	0.73	1.36 (1.17)	0
4	74,319.2	74,394.1	74,303.2	0.41 (0.64)	0.76	1.36 (1.17)	0
5	360,150.2	360,236.6	360,134.2	1.48 (1.22)	0.12	0.91 (0.95)	0.33
6	74,315.3	74,437.1	74,289.3	0.41 (0.64)	0.76	1.03 (1.01)	0.24
7	74,272.3	74,403.4	74,244.3	0.41 (0.64)	0.76	1.04 (1.02)	0.24
8	74,308.1	74,439.2	74,280.1	0.41 (0.64)	0.76	1.03 (1.01)	0.24
9	74,265.5	74,396.6	74,237.5	0.41 (0.64)	0.76	1.02 (1.01)	0.25
10	74,315.4	74,446.5	74,287.4	0.41 (0.64)	0.76	1.03 (1.02)	0.24
11	74,287.7	74,418.8	74,259.7	0.41 (0.64)	0.76	1.02 (1.01)	0.25
12	74,233.1	74,401.7	74,197.1	0.41 (0.64)	0.76	1.03 (1.01)	0.24

Note. N = 1747 for models 1–12. AIC = Akaike information criterion; BIC = Bayesian information criterion; *SE* = standard error. Model 1 is the unconditional model, models 2–6 are the main effects models, and models 7–12 are the interaction effects models. We were interested in estimating variance reduction for models 2–12 using the unconditional model (model 1) as a reference point.

coded 1.0 on the dichotomous predictor for students in the group who are at the mean of any continuous predictors in the model. For dichotomous item predictors in model 5, the regression parameter describes the difference in mean log odds of correctly answering items of the type described by the dichotomous item feature as compared with items in the reference category for a person of average ability. When both item and person features and their interactions are in the model, the precise interpretation of individual regression parameters will depend on other effects in the model.

Research Question 1: Main Effects of Student Characteristics

Results indicated that reading comprehension was a statistically significant predictor of word knowledge, controlling for grade level, GATE status, and SES status. As expected, word knowledge was also positively related to student grade, with students in grades 7 (β = 0.41, standard error [*SE*] = 0.08, $p < .001$) and 8 (β = 0.75, *SE* = 0.08, $p < .001$) having better odds of answering an average item correctly than students in grade 6. Not surprisingly, reading comprehension was positively strongly related to vocabulary knowledge (β = 1.17, *SE* = .02, $p < 0.001$). When reading comprehension is in the model, the regression coefficients for grades 7 and 8 remain statistically significant but change in sign because these effects now compare students in grades 7 and 8 who are at the mean of reading comprehension with grade 6 students who are at the sample mean on the GMRT extended scale scores. Not surprisingly, a student in grade 6 who is reading at the mean for the full sample has a somewhat higher probability of answering an average item correctly, as this student is an above-average student for grade 6. Students who were eligible for free or reduced-price lunch and those who were not enrolled in the district's GATE program had a lower chance of answering an item correctly on average as compared with their peers. Effects of grade were not statistically significant for grade 8 when SES status and participation in GATE programs were included in the model. Although the negative effect of grade 7 remained statistically significant, it was substantially smaller (–0.13 vs. –0.21).

As expected, adding reading comprehension to the model (model 3) substantially decreased the unexplained variance in student ability but had no effect on the variance in item difficulties (relative to the unconditional model, model 1). Model 3 accounted for 73.4% of the variance associated with student ability relative to the unconditional model, that is, (1.69 – 0.45)/1.69. At the same time, adding GATE status and SES status to the model (model 4) reduced the unexplained variance in student ability relative to model 3 by an additional 8.8%, that is, (0.45 – 0.41)/0.45. Compared with model 1, model 4 reduced the unexplained variance in student ability by 75.8%, that is, (1.69 – 0.41)/1.69.

Research Question 2: Main Effects of Word Features

The second research question asked about the relations between features of academic vocabulary (measured via item frequency, complexity, proximity, polysemy, and diversity) and item difficulty. We answered this question with reference to model 5. The model indicated that polysemy was the only statistically significant predictor of correct responses to the word knowledge items, over and above word frequency, complexity, proximity, and diversity. Words with more meanings were easier relative to words with fewer meanings (β = 0.35, *SE* = 0.13, *p* < .01). Adding word features to the model decreased the residual item variance and residual student variance relative to the unconditional model (model 1). Model 5 accounted for 33% of the variance in item difficulty and 12% of the variance in student ability as compared with model 1.

Combined Main Effects of Student Characteristics and Word Features

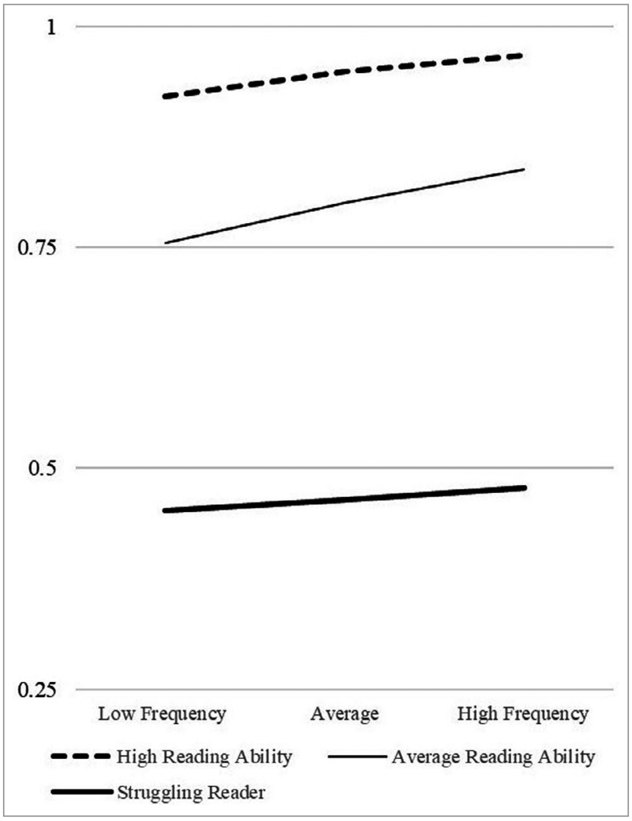
As expected, the combined model findings in model 6 for person characteristics and word features were identical to the results reported for these features separately in models 4 and 5, respectively, because person and word characteristics are not correlated in the design. That is, effects of student characteristics in model 6 parallel those observed in model 4, and effects of word features in model 6 parallel those observed in model 5. As such, student characteristics predominantly explain variance in student ability, and word features predominantly explain variance in item difficulty. At the same time, we expected that in the interaction effects model, the two sets of characteristics would jointly impact student ability and item difficulty.

Research Question 3: Interaction of Student Characteristics and Word Features

Although results suggested statistically significant main effects of reading comprehension, SES status, participation in GATE programs, and polysemy, these main effects discussed in regard to research question 1 may not tell the whole story with respect to vocabulary learning insofar as student characteristics and word features may interact in determining students' responses to vocabulary items. Models 7–11 examined the interaction of reading comprehension and word features individually and found statistically significant interactions between reading comprehension and (a) word frequency (β = 0.07, *SE* = 0.01, *p* < .001), (b) polysemy (β = 0.03, *SE* = 0.01, *p* = .002), (c) complexity (β = −0.08, *SE* = 0.01, *p* < .001), and (d) proximity (β = 0.06, *SE* = 0.01, *p* < .001), over and above the main effect of word and person features in the models. Although the magnitude of individual main effects in models 7–11 were comparable to those reported above for the same effect, the main effect of any term involved in an interaction should not be interpreted, as the interaction indicates that the effect is moderated by another variable, either another student characteristic or word feature.

Insofar as models 7–11 examine the interactions individually, these effects are correlated and must be considered in combination with one another to identify those that exert a unique influence on student responses to the vocabulary items. When all interactions of reading comprehension and word features were simultaneously entered in model 12, only interactions of reading comprehension with word frequency (β = 0.07, *SE* = 0.01, *p* < .001) and complexity (β = −0.07, *SE* = 0.01, *p* < .001) remained statistically significant. These interaction effects were small compared with the main effects. The interpretation of the main effects in light of the interactions is best appreciated by examining graphs depicting the interaction effects. As can be seen in Figure 2, there were large differences in the probability of answering an item correctly associated with overall reading ability. Although the interactions with reading ability are continuous by continuous interactions and generalize across reading skill abilities, we present prototypical plots of stronger (1.5 *SD*) and weaker (−1.5 *SD*) readers to demonstrate how these interactions work. Strong readers (dashed line) were more likely to answer items correctly than struggling readers (bold solid line). Figure 2 also demonstrates that high-frequency words were easier for both strong and struggling readers (based on the statistically nonsignificant main effect of frequency).

FIGURE 2
Probability of Correctly Answering an Item About a Low-, Average-, or High-Frequency Word, by Student Reading Proficiency Level



What is harder to see in the figure is that in addition to these two main effects, there is an interaction such that the effect of frequency is slightly stronger for high-ability readers than for struggling readers. Figure 3 is similar in many ways. However, in this case, more complex words are harder for all students, but it is the struggling readers (bold solid line) who are more sensitive to the effects of complexity.

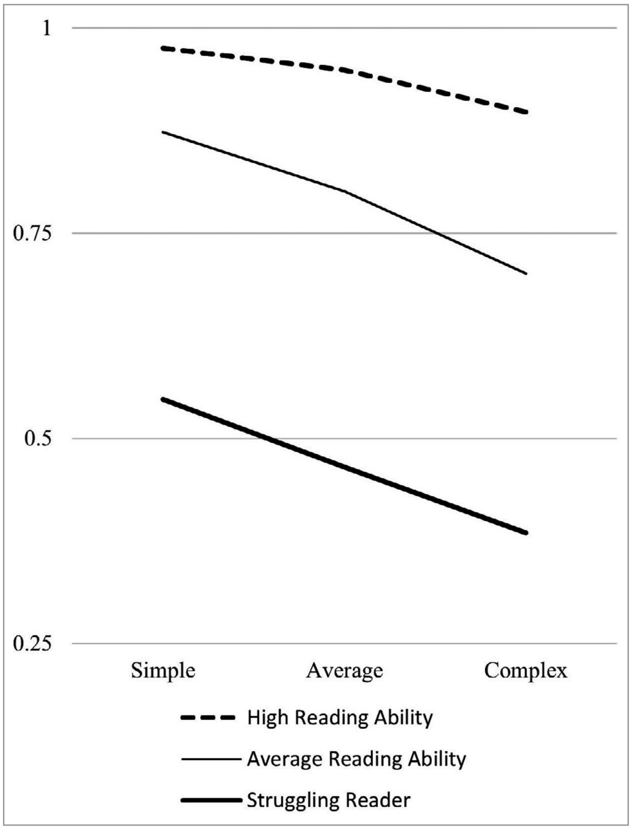
Interaction effects were generally small in their magnitudes. We can conceptualize this difference between variance accounted for in the two sides of the model as indicating that readers who are higher in ability tend to know more words regardless of the features of the words being tested. Although polysemy affects the probability of knowing a word, it exerts a similar effect on knowledge for good and poor readers. In contrast, although complexity and frequency interacted with reader ability, these interaction effects were relatively small in comparison with the main effect of reader ability.

Discussion

Summary of Findings

In this study, we explored the relations between five lexical dimensions and academic vocabulary knowledge by simultaneously modeling the effects of student and word characteristics. Our results affirm the strong relation between vocabulary and reading at the individual level. Strong readers were more likely than struggling readers to know the meanings of words, regardless of the features of the academic words tested. Our results also show that words with more meanings were easier for students, which aligns with an extensive literature showing that polysemous words are accessed more efficiently in adults (Eddington & Tokowicz, 2015). We tested reading ability by item characteristic interactions. These analyses showed that the relation between frequency and item difficulty is stronger for better readers and that the relation between complexity and difficulty is stronger for weaker readers.

FIGURE 3
Probability of Correctly Answering an Item About a Low-, Average-, or High-Complexity Word, by Student Reading Proficiency Level



The strong relation between reading and vocabulary achievement at the individual level is not surprising. Word knowledge has long been considered one of the best measures of general verbal skill, and vocabulary knowledge is strongly correlated with reading ability. Including individual-level covariates related to student SES status and academic achievement reduced the partial correlations between reading ability and academic vocabulary. In other words, the relation between reading ability and vocabulary is due in part to differences among students in characteristics such as SES status, participation in GATE programs, and grade level. These results align with one of the hypotheses presented in our introduction, namely, that the correlation between reading and vocabulary is at least partly spurious and due to differences in general skill, such as verbal ability, or metalinguistic awareness. Although our models lack a direct measure of general verbal ability, or metalinguistic awareness, the reduction in the correlation due to the inclusion of such student characteristics is consistent with this idea. Thus, although we expected these findings, this study's novel contribution is the exploration of these relations within the class of words known as academic words and using random effects models that allow generalization of the demonstrated relations back to the universe of academic words.

Given the large, multivariate space of word characteristics and the small set of words on which students can reasonably be tested, in the interest of parsimony, we relied on prior work by our group (Knoph et al., 2021) to reduce the dimensionality of the word characteristics for inclusion in the models. This prior work suggested five underlying factors related to frequency, complexity, proximity, polysemy, and diversity. We used factor scores on these five dimensions to examine the relation between word characteristics and item difficulty for words from the Academic Vocabulary Test, while treating the words as a source of random variation in the data. This treatment of words as random effects allows our findings to generalize back to the broad class of academic words from which we chose words on the Academic Vocabulary Test. We found that words with more senses were easier for students than words with fewer senses. Researchers using data from speeded lexical decision tasks with adults have also found an advantage for words with related senses. In contrast, Cervetti et al. (2015) found that polysemous words were more challenging for second, third, and fourth graders at the start of the intervention. Interestingly, students in that study learned polysemous words faster during instruction. These seemingly incongruent findings may make sense from a word-learning perspective, which we subsequently discuss.

We also modeled interactions between reading ability and item characteristics in predicting item difficulty. In this way, we could test whether the relation between word characteristics and item difficulty would vary as a function of student reading ability, controlling for the main effects of both. We found that the relation between word frequency and target word knowledge was stronger for better readers. This finding aligns with research showing that better readers are skilled at inferring the meanings of words they independently encounter in print (Swanborn & de Glopper, 1999). Poor readers are less efficient at inferring the meanings of newly encountered words, so the relation between the number of encounters they experience with a new word and their knowledge of it may be relatively weak. Less skilled readers are probably similarly inefficient at learning new words from other contexts. In either case, if we accept that item-level frequency measures are an appropriate proxy for student encounters with a word across print and other contexts, our findings align with those of research on incidental word learning. In contrast, because differences in students' independent reading diets are related to their reading ability, it is likely that item-level frequency measures are not an equally good proxy for encounters with texts for all students. If so, the interaction may also be related to differences in reading amounts.

We also found a stronger relation between word complexity and item difficulty for weaker readers. These results suggest that poor readers were more likely to struggle with the orthographic representation of a word and may need extra assistance to learn the meanings of orthographically complex words. For stronger readers, this dimension of word knowledge was not as related to word difficulty. These results align with theories suggesting that novice readers need to attend to decoding more than skilled readers do and that skilled readers can allocate more attention to higher order comprehension (LaBerge & Samuels, 1974).

Possible Implications for Instruction

Our findings align with those of research suggesting the importance of considering orthographic and morphological aspects of academic word instruction in middle grades and suggest that these dimensions may be particularly important for struggling readers. Intervention research with middle school students has shown the importance of morphological training, especially for students with weaker baseline scores (Lesaux et al., 2014). Instructional texts emphasize morphological and orthographic considerations in terms of how teachers select words for middle school learners (Beck et al., 2013) and support them (Dobbs, 2013; Templeton et al. 2015).

Research in incidental word learning has demonstrated individual differences in determining the meanings of newly encountered words (Swanborn & de Glopper, 1999). Our results align with those from incidental learning studies but also demonstrate why item selection can be challenging for vocabulary instruction. In our models, the best readers are the most likely to know words, and high-frequency words are more likely to be known than less frequent words are. On top of these effects, stronger readers are even more likely than poor readers to know high-frequency words. These differences present teachers with an instructional challenge. High-frequency words are essential for reading, so struggling students need to master them. However, stronger readers likely know these words well. This skill disparity may make it difficult for struggling students to feel comfortable acknowledging their difficulty with words that their classmates may consider easy. Instructional leaders acknowledge these challenges. They can be addressed in part by supporting an open and exploratory classroom culture (Scott, Skobel, & Wells, 2008) and providing students with explicit strategies for learning about new words when encountering them. In particular, support should be provided to help students master the spellings and morphological structures of complex words.

Polysemy is ubiquitous in English, which provides challenges and opportunities for vocabulary instruction. Researchers have shown how difficult it can be to learn new senses of conceptually rich words that are already partially known (González-Fernández & Schmitt, 2020; Nagy, Anderson, & Herman, 1987). To do so, learners must first notice that a newly encountered usage is novel by referencing both what they currently know about the word and the semantic constraints of the new context. Next, they must update what they know about the word form and register how this new meaning or sense is novel. This entire process is likely to support a rich representation of the word, especially when the word encounters are staggered. From this perspective, the learning of polysemous words may be another example of the trade-off between short-term performance and long-term learning (Soderstrom & Bjork, 2015). Younger students may find polysemous words harder to learn (Cervetti et al., 2015), but the process of learning them results in a more robust lexical representation, which explains the posttest results reported by Cervetti et al. (2015) and the results reported here.

The explicit teaching of word forms with distinct meanings (homophones and homonyms) is a staple in elementary classrooms. There is more variability in how strongly instructional texts and approaches emphasize the instruction of words with multiple related senses. Beck et al. (2013) noted that words with distinct meanings can be confusing, and suggested emphasizing multiple meanings "when introducing a word that has a meaning that students already know" (p. 79). Beck et al. noted that examining words with multiple senses provides an opportunity for teachers to talk about how language grows and how the same word can be used in several different ways. In *Teaching Words and How They Work: Small Changes for Big Vocabulary Results*, Hiebert (2019) extended this approach. She presented an instructional schema for talking about how words develop multiple meanings. Remixing is when a word takes a new meaning, and recycling is when words are combined in novel ways. Hiebert devoted a chapter to the history of English and a second chapter to these two processes, thereby emphasizing these aspects of vocabulary instruction to a significant degree.

Our research findings suggest that this emphasis is warranted. Although the effect of polysemy is modest relative to person-level variables in our models, our analysis is of words that had not been systematically taught at the time they were assessed. Students had no structured encounters with the multiple meanings of words or instructional support for learning them prior to testing. Thus, the advantage that students may have enjoyed while learning polysemous words was probably not fully realized in these results. The approaches advocated by Hiebert (2019) could help students extend and consolidate their learning in productive ways. If the relation between vocabulary and reading comprehension is driven in part by the fact that knowledge of words is also knowledge of the world and conceptual relations, this approach may also be particularly valuable in supporting reading comprehension.

Methods and Limitations

The contrasting results from the separate models involving individual interaction terms and the joint model involving all terms highlight the need for additional research. It is important to understand that the dynamic between specific terms across models changes as a function of effects being added or removed from these models due to the correlations betweem terms. In the current study, the observed effects were small, although the study was not designed to use words that would be explicitly sampled for specific features. As a result, effects of word features are correlated in this sample. A different study design could sample words so word feature effects were less correlated and so words were targeted to differ more on dimensions of interest. Such design would aid in disentangling relations among specific features of words, reading comprehension, and word knowledge. Furthermore, it is important to keep in mind that the power of the design for detecting the effects of word features is more a function of the numbers of words with specific features and has little to do with the sample size in terms of students. Standard errors for the regression parameters for word features could be reduced by increasing the sample size with respect to the number of items on the test, whereas effect sizes could be increased by sampling words to differ more on the dimensions of interest.

Clearly, as evidenced in Figure 1 and the table of factor correlations in the figure, the five factor scores that we used to summarize the characteristics of words and their meanings are correlated, leading them to account for both unique and shared variance in predicting word difficulty. The same can be said for the interaction terms in our models. When effects are correlated, the dynamic that plays out across different models for a given factor reflects variations in the unique contribution of the specific term after accounting for other terms in one model relative to another. Our analytic approach did not attempt to identify the best prediction model of a given size, but rather was designed to understand each feature's possible contribution in light of the contribution of other factors. Due to the intercorrelation across interaction terms, we examined these both individually and collectively. The fact that only two of the five interactions were statistically significant when all five terms were included in a single model, whereas four of five interaction terms were statistically significant when examined individually, reflects the fact that the different interaction terms account for overlapping information. As such, the specific interaction terms that are retained in the final model should be regarded with a certain degree of caution, as one might expect that the specific retained terms may fluctuate across replicate studies using different sets of words, and/or samples of readers, and could be expected to fluctuate if the sample size were varied, leading to greater or lesser power for detecting unique effects of specific terms (e.g., as sample size is increased or decreased, all other things being equal).

Given that we used factors scores as predictors in these models, it might be objected that the two-step approach we employed ignores the errors in estimating factor scores, which can lead to bias in the stage 2 regression parameters. This bias stems from the attenuation of correlations due to treating the factor scores as if they have been measured without error. In general, bivariate relations are biased toward 0, suggesting a reduction in power. However, in regression with multiple predictors measured with error, the relations among the predictors are also attenuated, which can result in some regression parameters being biased upward (i.e., inflated), whereas others are biased toward 0. This problem is most acute when scores differ in their precisions, with some scores having low standard errors and others having substantially larger ones. At the same time, from the standpoint of prediction through multiple regression, this bias is most concerning when our interest is tied to causal inferences based on the regression parameters. In prediction, this bias due to error is viewed less problematically because it contributes to the overall lack of precision in prediction.

There are at least two potential remedies. One is to conduct all analyses in a single step. Such an approach is unwieldy here because of the cross-classified random-effects structure and the relative sparseness of the design matrix for variable on factor regressions if untested words are included in the estimation of the factors in a single-stage model. More than likely, one would be forced to drop words from the AWL and GSL that were not tested on the vocabulary test. Restricting the single-stage analysis to the tested words would lead to poorer estimation of the factor scores for the tested words, which would then lead to bias in estimating the regression coefficients associated with the factors even though a single-stage analysis was used. An alternative is to conduct the second-stage regression by carrying forward the standard errors of the factor scores and using these standard errors to weight the second-stage regression analyses. If we were inclined toward causal inference for the second-stage regression coefficients rather than simple prediction of item difficulty, this added complexity in the second-stage regressions would be essential.

Whether the effects of word features on item difficulty can be leveraged to improve vocabulary instruction has only begun to be researched (Cervetti et al., 2015; Goodwin & Cho, 2016). We did not investigate the possibility of higher level interactions (e.g., Frequency × Complexity × Reading Ability × SES Status), primarily because our sample of words is limited to 50 items, which limits the number of interaction terms that should reasonably be included in the models. At the same time, based on the present findings, it is not unreasonable to speculate that a study designed specifically to investigate such heterogeneity in the effects of word and reader characteristics could have important implications for the design of instruction aimed at improving vocabulary knowledge for struggling readers. Similarly, we excluded English learners from the sample, but it is plausible that word features may interact with other student characteristics other than reading ability. Our focus in the present study was on the moderating effects of reading ability on word knowledge, but other characteristics of students are at least as important to consider in future research.

Finally, it seems worthwhile to point out the value of cross-classified random-effects models in reading research. In the present study, we made use of EIRT models, one type of cross-classified random-effects model for simultaneously modeling effects of the stimulus and the respondent on the response. Goodwin, Gilbert, Cho, and Kearns (2014) were among the first to apply these models in reading research and showed the value of these models for exploring complex theoretical questions, such as the lexical quality hypothesis (Perfetti & Hart, 2002) in reading comprehension. Kulesz et al. (2016) used a similar item response model to integrate component skills and text and discourse frameworks to investigate reading comprehension on a standardized reading assessment. Francis, Kulesz, and Benoit (2018) expanded on this general idea to show how cross-classified random-effects models could integrate these frameworks in developmental contexts, while also incorporating reading purpose and other contextual moderators, simultaneously allowing the functional form of the model to vary across respondents. This extension allows the separation of person-specific and person-general effects of stimulus attributes on response probabilities. Allowing stimulus characteristics to exert both person-specific and person-general effects has important implications for teaching but poses significant challenges for research because of the need for intensive data collection (i.e., many stimulus items) on large numbers of subjects, if the person-specific functions are to be estimated with sufficient precision to support instructional decisions. However, as automated measurement becomes more ubiquitous through interaction with personal electronic devices in educational contexts, such data collection becomes feasible and minimally burdensome to a student while simultaneously creating the possibility for presenting learning opportunities tailored to the precise needs of the student. Our understanding of the student and stimulus characteristics that affect learning and the extent to which these features interact will determine the success of any such endeavors to craft effective student-specific instruction.

The current study has important limitations. First, our focus was on monolingual students. We are currently advancing these models with a more diverse range of students to understand how language proficiency may be related to word learning, item characteristics, and reading comprehension. Second, there are many ways to know a word, and the ways that one may know a word can vary according to the word (Nagy & Scott, 2000). In our analysis, we examined results from a synonym task. We are currently working to extend our analysis of word features to understand how they may support or disrupt word learning across a broader range of vocabulary assessment types. Third, we only examined general academic words in this analysis. Although the factor structures that we described among these words look similar in discipline-specific academic words, it is difficult to establish valid and reliable reading performance estimates across domains. Consequently, we have yet to replicate these analyses with discipline-specific vocabulary and discipline-specific measures of reading comprehension. Still, this study extends how we think about what makes academic vocabulary challenging for middle school students, the extent to which these challenges vary across students, and how student learning might be supported across a broad range of student vocabulary and reading proficiency ranges.

NOTES

This work was supported by the Institute of Education Sciences of the U.S. Department of Education through a grant (R305A120045, "Improving the Accuracy of Academic Vocabulary Assessment for English Language Learners") awarded to David Francis (principal investigator) and a grant (R305A090555, "Word Generation: An Efficacy Trial") awarded to Catherine Snow (principal investigator). The content and opinions are solely the responsibility of the authors and do not represent the official views of the Institute of Education Sciences or the U.S. Department of Education.

[1] These tasks require participants to indicate whether a particular isolated string of letters is an English word. Because participants can perform judgment tasks on hundreds of words per session, researchers have been able to establish estimates of processing efficiency for tens of thousands of words and related these to word characteristics.

[2] The district also provided other student-level data related to home language use, school entry date, language fluency, score on a language proficiency test, and the language guardians requested the report card be printed in. These data were used in previous analyses but are not relevant to the current analysis of data from English monolinguals. Thus, we selected individual-level variables used in this analysis for convenience. No other individual-level data were modeled in our analysis for this article.

[3] We provided descriptive labels for each factor to facilitate reference to them throughout the article. These labels are purely descriptive and were derived based on the best available evidence at this time with respect to the nature of each factor. It is important to note that the interpretation of factors is not strictly a matter of examining factor loadings when factors are correlated. However, it is even more important to realize that the precise nature of latent constructs is rarely, if ever, settled by a single study but is certainly never clear from a single exploratory factor analysis. Although we believe that the proposed working labels are reasonably accurate descriptions and reflect our current understanding, additional research is warranted and may lead to different understandings regarding the nature of the factors, as well as the number of required dimensions.

REFERENCES

Adelman, J.S., Brown, G.D.A., & Quesada, J.F. (2006). Contextual diversity, not word frequency, determines word-naming and lexical decision times. *Psychological Science*, *17*(9), 814–823. https://doi.org/10.1111/j.1467-9280.2006.01787.x

Aitchison, J. (2012). *Words in the mind: An introduction to the mental lexicon* (4th ed.). Malden, MA: John Wiley & Sons.

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

Azuma, T., & Van Orden, G.C. (1997). Why SAFE is better than FAST: The relatedness of a word's meanings affects lexical decision times. *Journal of Memory and Language*, *36*(4), 484–504. https://doi.org/10.1006/jmla.1997.2502

Balota, D.A., Yap, M.J., Hutchison, K.A., Cortese, M.J., Kessler, B., Loftis, B., … Treiman, R. (2007). The English Lexicon Project. *Behavior Research Methods*, *39*(3), 445–459. https://doi.org/10.3758/BF03193014

Bates, D., Maechler, M., & Dai, B. (2021). *lme4: Linear mixed-effects models using "Eigen" and S4* (Version 1.1-27) [Computer software]. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from http://cran.r-project.org/web/packages/lme4/index.html

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

Beck, I.L., Perfetti, C.A., & McKeown, M.G. (1982). Effects of long-term vocabulary instruction on lexical access and reading comprehension. *Journal of Educational Psychology*, *74*(4), 506–521. https://doi.org/10.1037/0022-0663.74.4.506

Beretta, A., Fiorentino, R., & Poeppel, D. (2005). The effects of homonymy and polysemy on lexical access: An MEG study. *Cognitive Brain Research*, *24*(1), 57–65. https://doi.org/10.1016/j.cogbrainres.2004.12.006

Biemiller, A., & Slonim, N. (2001). Estimating root word vocabulary growth in normative and advantaged populations: Evidence for a common sequence of vocabulary acquisition. *Journal of Educational Psychology*, *93*(3), 498–520. https://doi.org/10.1037/0022-0663.93.3.498

Borowsky, R., & Masson, M.E.J. (1996). Semantic ambiguity effects in word identification. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *22*(1), 63–85. https://doi.org/10.1037/0278-7393.22.1.63

Braze, D., Katz, L., Magnuson, J.S., Mencl, W.E., Tabor, W., Van Dyke, J.A., … Shankweiler, D.P. (2016). Vocabulary does not complicate the simple view of reading. *Reading and Writing*, *29*, 435–451. https://doi.org/10.1007/s11145-015-9608-6

Breland, H.M. (1996). Word frequency and word difficulty: A comparison of counts in four corpora. *Psychological Science*, *7*(2), 96–99. https://doi.org/10.1111/j.1467-9280.1996.tb00336.x

Brysbaert, M., Mandera, P., McCormick, S.F., & Keuleers, E. (2019). Word prevalence norms for 62,000 English lemmas. *Behavior Research Methods*, *51*, 467–479. https://doi.org/10.3758/s13428-018-1077-9

Brysbaert, M., & New, B. (2009). Moving beyond Kučera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behavior Research Methods*, *41*(4), 977–990. https://doi.org/10.3758/BRM.41.4.977

Carlisle, J.F., & Stone, C.A. (2005). Exploring the role of morphemes in word reading. *Reading Research Quarterly*, *40*(4), 428–449. https://doi.org/10.1598/RRQ.40.4.3

Carroll, J.B. (1941). A factor analysis of verbal abilities. *Psychometrika*, *6*(5), 279–307. https://doi.org/10.1007/BF02288585

Cervetti, G.N., Hiebert, E.H., Pearson, P.D., & McClung, N.A. (2015). Factors that influence the difficulty of science words. *Journal of Literacy Research*, *47*(2), 153–185. https://doi.org/10.1177/1086296X15615363

Coltheart, M., Davelaar, E., Jonasson, J.T., & Besner, D. (1977). Access to the internal lexicon. In S. Dornič (Ed.), *Attention and performance* (Vol. 6, pp. 535–556). Hillsdale, NJ: Erlbaum.

Coxhead, A. (2000). A new academic word list. *TESOL Quarterly*, *34*(2), 213–238. https://doi.org/10.2307/3587951

Cromley, J.G., & Azevedo, R. (2007). Testing and refining the direct and inferential mediation model of reading comprehension. *Journal of Educational Psychology*, *99*(2), 311–325. https://doi.org/10.1037/0022-0663.99.2.311

Crosson, A.C., Lesaux, N.K., & Martiniello, M. (2008). Factors that influence comprehension of connectives among language minority children from Spanish-speaking backgrounds. *Applied Psycholinguistics*, *29*(4), 603–625. https://doi.org/10.1017/S0142716408080260

Dale, E., & O'Rourke, J. (1981). *The living word vocabulary: A national vocabulary inventory*. Chicago, IL: World Book-Childcraft International.

Davies, M. (2009). The 385+ million word Corpus of Contemporary American English (1990–2008+): Design, architecture, and linguistic insights. *International Journal of Corpus Linguistics*, *14*(2), 159–190. https://doi.org/10.1075/ijcl.14.2.02dav

Deacon, S.H., Whalen, R., & Kirby, J.R. (2011). Do children see the *danger* in *dangerous*? Grade 4, 6, and 8 children's reading of morphologically complex words. *Applied Psycholinguistics*, *32*(3), 467–481. https://doi.org/10.1017/S0142716411000166

De Boeck, P., & Wilson, M. (2004). A framework for item response models. In P. De Boeck & M. Wilson (Eds.), *Explanatory item response models: A generalized linear and nonlinear approach* (pp. 3–41). New York, NY: Springer.

Dobbs, C.L. (2013). Vocabulary in practice: Creating word-curious classrooms. In J. Ippolito, J.F. Lawrence, & C. Zaller (Eds.), *Adolescent literacy in the era of the Common Core: From research into practice* (pp. 73–83). Cambridge, MA: Harvard Education Press.

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

Elleman, A.M., Lindo, E.J., Morphy, P., & Compton, D.L. (2009). The impact of vocabulary instruction on passage-level comprehension of school-age children: A meta-analysis. *Journal of Research on Educational Effectiveness*, *2*(1), 1–44. https://doi.org/10.1080/19345740802539200

Francis, D.J., Kulesz, P.A., & Benoit, J.S. (2018). Extending the simple view of reading to account for variation within readers and across texts: The complete view of reading (CVR*i*). *Remedial and Special Education*, *39*(5), 274–288. https://doi.org/10.1177/0741932518772904

Gilbert, J.K., Compton, D.L., & Kearns, D.M. (2011). Word and person effects on decoding accuracy: A new look at an old question. *Journal of Educational Psychology*, *103*(2), 489–507. https://doi.org/10.1037/a0023001

González-Fernández, B., & Schmitt, N. (2020). Word knowledge: Exploring the relationships and order of acquisition of vocabulary knowledge components. *Applied Linguistics*, *41*(4), 481–505. https://doi.org/10.1093/applin/amy057

Goodwin, A.P., & Cho, S.-J. (2016). Unraveling vocabulary learning: Reader and item-level predictors of vocabulary learning within comprehension instruction for fifth and sixth graders. *Scientific Studies of Reading*, *20*(6), 490–514. https://doi.org/10.1080/10888438.2016.1245734

Goodwin, A.P., Gilbert, J.K., Cho, S.-J., & Kearns, D.M. (2014). Probing lexical representations: Simultaneous modeling of word and reader contributions to multidimensional lexical representations. *Journal of Educational Psychology*, *106*(2), 448–468. https://doi.org/10.1037/a0034754

Hiebert, E.H. (2019). *Teaching words and how they work: Small changes for big vocabulary results*. New York, NY: Teachers College Press.

Hiebert, E.H., Goodwin, A.P., & Cervetti, G.N. (2018). Core vocabulary: Its morphological content and presence in exemplar texts. *Reading Research Quarterly*, *53*(1), 29–49. https://doi.org/10.1002/rrq.183

Hiebert, E.H., & Lubliner, S. (2008). The nature, learning, and instruction of general academic vocabulary. In A.E. Farstrup & S.J. Samuels (Eds.), *What research has to say about vocabulary instruction* (pp. 106–129). Newark, DE: International Reading Association.

Hiebert, E.H., Scott, J.A., Castaneda, R., & Spichtig, A. (2019). An analysis of the features of words that influence vocabulary difficulty. *Education in Science*, *9*(1), Article 8. https://doi.org/10.3390/educsci9010008

Hino, Y., & Lupker, S.J. (1996). Effects of polysemy in lexical decision and naming: An alternative to lexical access accounts. *Journal of Experimental Psychology: Human Perception and Performance*, *22*(6), 1331–1356. https://doi.org/10.1037/0096-1523.22.6.1331

Hino, Y., Lupker, S.J., & Pexman, P.M. (2002). Ambiguity and synonymy effects in lexical decision, naming, and semantic categorization tasks: Interactions between orthography, phonology, and semantics. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *28*(4), 686–713. https://doi.org/10.1037/0278-7393.28.4.686

Hoffman, P., Lambon Ralph, M.A., & Rogers, T.T. (2013). Semantic diversity: A measure of semantic ambiguity based on variability in the contextual usage of words. *Behavior Research Methods*, *45*(3), 718–730. https://doi.org/10.3758/s13428-012-0278-x

Hoover, W.A., & Gough, P.B. (1990). The simple view of reading. *Reading and Writing*, *2*(2), 127–160. https://doi.org/10.1007/BF00401799

Joshi, R.M. (2005). Vocabulary: A critical component of comprehension. *Reading & Writing Quarterly*, *21*(3), 209–219. https://doi.org/10.1080/10573560590949278

Joshi, R.M., & Aaron, P.G. (2000). The component model of reading: Simple view of reading made a little more complex. *Reading Psychology*, *21*(2), 85–97. https://doi.org/10.1080/02702710050084428

Kieffer, M.J., & Box, C.D. (2013). Derivational morphological awareness, academic vocabulary, and reading comprehension in linguistically diverse sixth graders. *Learning and Individual Differences*, *24*, 168–175. https://doi.org/10.1016/j.lindif.2012.12.017

Knoph, R.E., Lawrence, J.F., & Francis, D.J. (2021). *The dimensionality of empirical lexical features in general, academic, and disciplinary vocabulary*. Manuscript in preparation.

Kučera, H., & Francis, W.N. (1967). *Computational analysis of present-day American English*. Providence, RI: Brown University Press.

Kulesz, P.A., Francis, D.J., Barnes, M.A., & Fletcher, J.M. (2016). The influence of properties of the test and their interactions with reader characteristics on reading comprehension: An explanatory item response study. *Journal of Educational Psychology*, *108*(8), 1078–1097. https://doi.org/10.1037/edu0000126

Kuperman, V., Stadthagen-Gonzalez, H., & Brysbaert, M. (2012). Age-of-acquisition ratings for 30,000 English words. *Behavior Research Methods*, *44*(4), 978–990. https://doi.org/10.3758/s13428-012-0210-4

LaBerge, D., & Samuels, S.J. (1974). Toward a theory of automatic information processing in reading. *Cognitive Psychology*, *6*, 293–323. https://doi.org/10.1016/0010-0285(74)90015-2

Lawrence, J.F., Crosson, A.C., Paré-Blagoev, E.J., & Snow, C.E. (2015). Word Generation randomized trial: Discussion mediates the impact of program treatment on academic word learning. *American Educational Research Journal*, *52*(4), 750–786. https://doi.org/10.3102/0002831215579485

Lawrence, J.F., Francis, D., Paré-Blagoev, J., & Snow, C.E. (2017). The poor get richer: Heterogeneity in the efficacy of a school-level intervention for academic language. *Journal of Research on Educational Effectiveness*, *10*(4), 767–793. https://doi.org/10.1080/19345747.2016.1237596

Lawrence, J.F., Hagen, A.M., Hwang, J.K., Lin, G., & Lervåg, A. (2019). Academic vocabulary and reading comprehension: Exploring the relationships across measures of vocabulary knowledge. *Reading and Writing*, *32*(2), 285–306. https://doi.org/10.1007/s11145-018-9865-2

Lawrence, J.F., Lin, G., Jaeggi, S., Kreger, N., Hwang, J.K., & Hagen, Å. (2021). *Measures of lexical ambiguity for 62,954 words from WordNet*. Manuscript in preparation.

Lesaux, N.K., Kieffer, M.J., Faller, E., & Kelley, J. (2010). The effectiveness and ease of implementation of an academic vocabulary intervention for linguistically diverse students in urban middle schools. *Reading Research Quarterly*, *45*(2), 196–228. https://doi.org/10.1598/RRQ.45.2.3

Lesaux, N.K., Kieffer, M.J., Kelley, J.G., & Harris, J.R. (2014). Effects of academic vocabulary instruction for linguistically diverse adolescents: Evidence from a randomized field trial. *American Educational Research Journal*, *51*(6), 1159–1194. https://doi.org/10.3102/0002831214532165

Lin, Y., Michel, J.-B., Lieberman Aiden, E., Orwant, J., Brockman, W., & Petrov, S. (2012). Syntactic annotations for the Google Books Ngram Corpus. In *Proceedings of the 50th annual meeting of the Association for Computational Linguistics: Vol. 2. Demo papers* (pp. 169–174). Stroudsburg, PA: Association for Computational Linguistics.

Lund, K., & Burgess, C. (1996). Producing high-dimensional semantic spaces from lexical co-occurrence. *Behavior Research Methods, Instruments, & Computers*, *28*(2), 203–208. https://doi.org/10.3758/BF03204766

Maria, K., Hughes, K.E., MacGinitie, W.H., MacGinitie, R.K., & Dreyer, L.G. (2007). *Lexile conversions for the Gates–MacGinitie Reading Tests* (4th ed.). Rolling Meadows, IL: Riverside.

McCarthy, K.S., Guerrero, T.A., Kent, K.M., Allen, L.K., McNamara, D.S., Chao, S.-F., … Sabatini, J. (2018). Comprehension in a scenario-based assessment: Domain and topic-specific background knowledge. *Discourse Processes*, *55*(5/6), 510–524. https://doi.org/10.1080/0163853X.2018.1460159

McKeown, M.G., Beck, I.L., Omanson, R.C., & Perfetti, C.A. (1983). The effects of long-term vocabulary instruction on reading comprehension: A replication. *Journal of Reading Behavior*, *15*(1), 3–18. https://doi.org/10.1080/10862968309547474

Mezynski, K. (1983). Issues concerning the acquisition of knowledge: Effects of vocabulary training on reading comprehension. *Review of Educational Research*, *53*(2), 253–279. https://doi.org/10.3102/00346543053002253

Miller, G.A. (Ed.). (1990). WordNet: An on-line lexical database. *International Journal of Lexicography*, *3*(4), 235–312.

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*(4), 438. https://doi.org/10.1037/1040-3590.5.4.438

Nagy, W. (2007). Metalinguistic awareness and the vocabulary–comprehension connection. In R.K. Wagner, A.E. Muse, & K.R. Tannenbaum (Eds.), *Vocabulary acquisition: Implications for reading comprehension* (pp. 52–77). New York, NY: Guilford.

Nagy, W.E., Anderson, R.C., & Herman, P.A. (1987). Learning word meanings from context during normal reading. *American Educational Research Journal*, *24*(2), 237–270. https://doi.org/10.3102/00028312024002237

Nagy, W.E., & Scott, J.A. (2000). Vocabulary processes. In M.L. Kamil, P.B. Mosenthal, P.D. Pearson, & R. Barr (Eds.), *Handbook of reading research* (Vol. 3, pp. 269–284). Mahwah, NJ: Erlbaum.

Nagy, W., & Townsend, D. (2012). Words as tools: Learning academic vocabulary as language acquisition. *Reading Research Quarterly*, *47*(1), 91–108. https://doi.org/10.1002/RRQ.011

New, B., Ferrand, L., Pallier, C., & Brysbaert, M. (2006). Reexamining the word length effect in visual word recognition: New evidence from the English Lexicon Project. *Psychonomic Bulletin & Review*, *13*(1), 45–52. https://doi.org/10.3758/BF03193811

Olson, R., Forsberg, H., Wise, B., & Rack, J. (1994). Measurement of word recognition, orthographic, and phonological skills. In G.R. Lyon (Ed.), *Frames of reference for the assessment of learning disabilities: New views on measurement issues* (pp. 243–277). Baltimore, MD: Paul H. Brookes.

Oxford online dictionary. (2015). Retrieved from https://en.oxforddictionaries.com/

Pany, D., Jenkins, J.R., & Schreck, J. (1982). Vocabulary instruction: Effects on word knowledge and reading comprehension. *Learning Disability Quarterly*, *5*(3), 202–215. https://doi.org/10.2307/1510288

Parks, R., Ray, J., & Bland, S. (1998). *Wordsmyth English dictionary-thesaurus*. Chicago, IL: University of Chicago Retrieved from https://www.wordsmyth.net/

Perfetti, C.A. (1988). Verbal efficiency in reading ability. In M. Daneman, T. MacKinnon, & T.G. Waller (Eds.), *Reading research: Advances in theory and practice* (Vol. 6, pp. 109–143). New York, NY: Academic.

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

Perfetti, C., & Stafura, J. (2014). Word knowledge in a theory of reading comprehension. *Scientific Studies of Reading*, *18*(1), 22–37. https://doi.org/10.1080/10888438.2013.827687

Praninskas, J. (1972). *American university word list*. London, UK: Longman.

Qian, D.D. (2002). Investigating the relationship between vocabulary knowledge and academic reading performance: An assessment perspective. *Language Learning*, *52*(3), 513–536. https://doi.org/10.1111/1467-9922.00193

Quinn, J.M., Wagner, R.K., Petscher, Y., & Lopez, D. (2015). Developmental relations between vocabulary knowledge and reading comprehension: A latent change score modeling study. *Child Development*, *86*(1), 159–175. https://doi.org/10.1111/cdev.12292

RAND Reading Study Group. (2002). *Reading for understanding: Toward an R&D program in reading comprehension*. Santa Monica, CA: RAND.

Rodd, J., Gaskell, G., & Marslen-Wilson, W. (2002). Making sense of semantic ambiguity: Semantic competition in lexical access. *Journal of Memory and Language*, *46*(2), 245–266. https://doi.org/10.1006/jmla.2001.2810

Rydland, V., Aukrust, V.G., & Fulland, H. (2013). Living in neighborhoods with high or low co-ethnic concentration: Turkish–Norwegian-speaking students' vocabulary skills and reading comprehension. *International Journal of Bilingual Education and Bilingualism*, *16*(6), 657–674. https://doi.org/10.1080/13670050.2012.709224

Schmitt, N., Jiang, X., & Grabe, W. (2011). The percentage of words known in a text and reading comprehension. *Modern Language Journal*, *95*(1), 26–43. https://doi.org/10.1111/j.1540-4781.2011.01146.x

Scott, J.A., Skobel, B.J., & Wells, J. (2008). *The word-conscious classroom: Building the vocabulary readers and writers need*. New York, NY: Scholastic.

Snow, C.E., Porche, M.V., Tabors, P., & Harris, S.R. (2007). *Is literacy enough? Pathways to academic success for adolescents*. Baltimore, MD: Paul H. Brookes.

Soderstrom, N.C., & Bjork, R.A. (2015). Learning versus performance: an integrative review. *Perspectives on Psychological Science*, *10*(2), 176–199. https://doi.org/10.1177/1745691615569000

Spearman, C. (1904). "General intelligence", objectively determined and measured. *The American Journal of Psychology*, *15*(2), 201–292. https://doi.org/10.2307/1412107

Stahl, S.A., & Fairbanks, M.M. (1986). The effects of vocabulary instruction: A model-based meta-analysis. *Review of Educational Research*, *56*(1), 72–110. https://doi.org/10.3102/00346543056001072

Stanovich, K.E. (1986). Matthew effects in reading: Some consequences of individual differences in the acquisition of literacy. *Reading Research Quarterly*, *21*(4), 360–407. https://doi.org/10.1598/RRQ.21.4.1

Sternberg, R.J., & Powell, J.S. (1983). Comprehending verbal comprehension. *The American Psychologist*, *38*(8), 878–893. https://doi.org/10.1037/0003-066X.38.8.878

Strategic Education Research Project. (2021). *WordGen Weekly: Academic Language strategies for today's youth*. Retrieved from https://www.serpinstitute.org/wordgen-weekly

Swanborn, M.S.L., & de Glopper, K. (1999). Incidental word learning while reading: A meta-analysis. *Review of Educational Research*, *69*(3), 261–285. https://doi.org/10.3102/00346543069003261

Tannenbaum, K.R., Torgesen, J.K., & Wagner, R.K. (2006). Relationships between word knowledge and reading comprehension in third-grade children. *Scientific Studies of Reading*, *10*(4), 381–398. https://doi.org/10.1207/s1532799xssr1004_3

Templeton, S., Bear, D.R., Invernizzi, M., Johnston, F., Flanigan, K., Townsend, D.R., … Hayes, L. (2015). *Words their way: Vocabulary for middle and secondary students*. Upper Saddle River, NJ: Pearson.

Townsend, D., Filippini, A., Collins, P., & Biancarosa, G. (2012). Evidence for the importance of academic word knowledge for the academic achievement of diverse middle school students. *The Elementary School Journal*, *112*(3), 497–518. https://doi.org/10.1086/663301

Tucker-Drob, E.M. (2009). Differentiation of cognitive abilities across the life span. *Developmental Psychology*, *45*(4), 1097–1118. https://doi.org/10.1037/a0015864

Tunmer, W.E., & Herriman, M.L. (1984). The development of metalinguistic awareness: A conceptual overview. In W.E. Tunmer, C. Pratt, & M.L. Herriman (Eds.), *Metalinguistic awareness in children: Theory, research, and implications* (pp. 27–36). Berlin, Germany: Springer-Verlag.

Verhoeven, L., van Leeuwe, J., & Vermeer, A. (2011). Vocabulary growth and reading development across the elementary school years. *Scientific Studies of Reading*, *15*(1), 8–25. https://doi.org/10.1080/10888438.2011.536125

Wagner, R.K., Torgesen, J.K., Rashotte, C.A., Hecht, S.A., Barker, T.A., Burgess, S.R., … Garon, T. (1997). Changing relations between phonological processing abilities and word-level reading as children develop from beginning to skilled readers: A 5-year longitudinal study. *Developmental Psychology*, *33*(3), 468–479. https://doi.org/10.1037/0012-1649.33.3.468

West, M. (1957). *A general service list of English words: With semantic frequencies and a supplementary word-list for the writing of popular science and technology*. London, UK: Longmans, Green.

Wright, T.S., & Cervetti, G.N. (2017). A systematic review of the research on vocabulary instruction that impacts text comprehension. *Reading Research Quarterly*, *52*(2), 203–226. https://doi.org/10.1002/rrq.163

Yap, M.J., Balota, D.A., Sibley, D.E., & Ratcliff, R. (2012). Individual differences in visual word recognition: Insights from the English Lexicon Project. *Journal of Experimental Psychology: Human Perception and Performance*, *38*(1), 53–79. https://doi.org/10.1037/a0024177

Yarkoni, T., Balota, D., & Yap, M. (2008). Moving beyond Coltheart's *N*: A new measure of orthographic similarity. *Psychonomic Bulletin & Review*, *15*(5), 971–979. https://doi.org/10.3758/PBR.15.5.971

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

Ziegler, J.C., & Goswami, U. (2005). Reading acquisition, developmental dyslexia, and skilled reading across languages: A psycholinguistic grain size theory. *Psychological Bulletin*, *131*(1), 3–29. https://doi.org/10.1037/0033-2909.131.1.3

Submitted July 15, 2020
Final revision received May 11, 2021
Accepted May 14, 2021

JOSHUA F. LAWRENCE (corresponding author) is a professor in the Department of Education at the University of Oslo, Norway; email joshua.lawrence@iped.uio.no. His research interests relate to understanding adolescent literacy development, second-language acquisition, hybrid learning, and improving instruction through coaching and leadership.

REBECCA KNOPH is a doctoral student at the University of Oslo, Norway; email rebecca.knoph@iped.uio.no. Her research interests include second-language acquisition and testing fairness and equivalency for native and non-native students.

AUTUMN MCILRAITH was a postdoctoral fellow at the Texas Institute for Measurement, Evaluation, and Statistics at the University of Houston, Texas, USA, at the time this work was completed and is now a data analyst and independent researcher; email autumnlorayne@gmail.com. Her research interests include word reading, reading comprehension, reading disorders, and statistical methods.

PAULINA A. KULESZ is a research associate at the Texas Institute for Measurement, Evaluation, and Statistics at the University of Houston, Texas, USA; email paulina.kulesz@times.uh.edu. Her primary research focus is cognitive processes underlying reading comprehension.

DAVID J. FRANCIS is a Hugh Roy and Lillie Cranz Cullen Distinguished University Chair and the director of the Texas Institute for Measurement, Evaluation, and Statistics at the University of Houston, Texas, USA; email dfrancis@uh.edu. His research interests focus on the application of advanced statistical models to problems in education and child development, especially as related to the study of reading and language, the identification and treatment of reading and related developmental disabilities, and the education of at-risk populations, especially English learners.

APPENDIX

TABLE A1
Variable Names, Descriptions, and Citations

Variable name	Description	Citation
bg_mean	Mean frequency of bigrams in a word (e.g., *the* = *th* + *he*)	Balota et al. (2007)
cd	Number of documents in which a word appears (contextual diversity) in the Touchstone Applied Sciences Associates corpus	Adelman, Brown, and Quesada (2006)
cocazipf	Zipfian-transformed frequency in the Corpus of Contemporary American English	Davies (2009)
d	Number of subject areas in which a word appears (dispersion) in *The Educator's Word Frequency Guide*	Zeno, Ivens, Millard, and Duvvuri (1995)

(*continued*)

TABLE A1
Variable Names, Descriptions, and Citations (*continued*)

Variable name	Description	Citation
freqband	Frequency groupings from the *Oxford Online Dictionary* based on raw frequencies from Ngram	Oxford Online Dictionary (2015)
length	Number of letters	Balota et al. (2007)
lg10cd	Log-transformed percentage of film and television series transcripts in which a word occurs in the SubtlexUS corpus	Brysbaert and New (2009)
lg10wf	Log-transformed frequency per million words in the SubtlexUS corpus	Brysbaert and New (2009)
log_freq_hal	Log-transformed frequency in the Hyperspace Analogue to Language corpus	Lund and Burgess (1996)
log_freq_kf	Log-transformed frequency in the Brown University Standard Corpus of Present-Day American English	Kučera and Francis (1967)
nmorph	Number of morphemes	Balota et al. (2007)
nphon	Number of phonemes	Balota et al. (2007)
nsyll	Number of syllables	Balota et al. (2007)
og_n	Raw number of phonographic neighbors (e.g., *stove/stone*)	Balota et al. (2007)
old	Mean Levenshtein distance of 20 closest orthographic neighbors	Yarkoni, Balota, and Yap (2008)
ortho_n	Raw number of orthographic neighbors (e.g., *love/dove*)	Balota et al. (2007)
phono_n	Raw number of phonologic neighbors (e.g., *hear/hare*)	Balota et al. (2007)
pld	Mean Levenshtein distance of 20 closest phonologic neighbors	Balota et al. (2007)
semd	Mean cosine of latent semantic analysis vectors of all pairwise combinations of contexts containing a word (semantic diversity)	Hoffman, Lambon Ralph, and Rogers (2013)
sfi	Weighted frequency per million tokens divided by dispersion (standardized frequency index)	Zeno et al. (1995)
subzipf	Zipfian-transformed frequency in the SubtlexUS corpus	Brysbaert and New (2009)
word_age	Age of a word as of 2000, from the *Oxford Online Dictionary*	Oxford Online Dictionary (2015)
wordage	Age of a word as of 2000, from Google Ngram	Lin et al. (2012)
wordnet_lnapossam	Log-transformed senses and meanings across parts of speech from WordNet	G.A. Miller (1990)
wordsmyth_lnapossam	Log-transformed senses and meanings across parts of speech from Wordsmyth	Parks, Ray, and Bland (1998)
z_sem_prec	z-transformed depth scores averaged by part of speech from WordNet	G.A. Miller (1990)
zenozipf	Zipfian-transformed frequency from *The Educator's Word Frequency Guide*	Zeno et al. (1995)

TABLE A2
Factor Score Estimation Beta Weights for Word Features

	Beta weights for factor score estimates				
Variable name	Frequency	Complexity	Proximity	Polysemy	Diversity
bg_mean	Omitted because of low measure of sampling adequacy (MSA < .60)				
cd	0.036	0.002	0.006	0.000	0.023
cocazipf	0.365	0.008	0.000	−0.045	−0.061
d	0.026	−0.001	0.001	0.003	0.325
freqband	0.052	−0.005	−0.005	−0.003	−0.036
length	−0.007	0.324	−0.006	0.086	0.003
lg10cd	Omitted because of high correlation with lg10wf and subzipf				
lg10wf	Omitted because of high correlation with lg10cd and subzipf				
log_freq_hal	0.171	−0.012	−0.001	0.021	−0.094
log_freq_kf	0.138	0.010	0.000	0.019	0.061
nmorph	−0.006	0.047	0.002	−0.001	−0.002
nphon	−0.003	0.297	−0.004	0.045	−0.011
nsyll	0.009	0.114	−0.006	−0.035	−0.105
og_n	−0.003	0.018	0.565	0.007	−0.019
old	0.000	0.113	−0.013	−0.079	0.084
ortho_n	−0.005	−0.018	0.347	−0.011	−0.010
phono_n	0.001	−0.014	0.085	−0.002	0.009
pld	0.002	0.148	−0.008	−0.071	0.044
semd	−0.002	0.000	0.000	0.006	0.497
sfi	Omitted because of high correlation with zenozipf				
subzipf	0.053	−0.008	0.000	0.035	0.036
word_age	Omitted because of low MSA				
wordage	0.015	0.003	−0.001	0.024	0.105
wordnet_lnapossam	0.011	−0.010	−0.002	0.200	−0.048
wordsmyth_lnapossam	0.000	−0.009	0.002	0.705	0.021
z_sem_prec	0.011	0.003	0.000	0.045	−0.154
zenozipf	0.233	−0.001	0.002	0.032	0.101