

Understanding Vocabulary:
Making Sense of What We Measure,
Who We Measure, and How We Measure

Rebecca Knoph



Thesis submitted for the degree of Ph.D.

Department of Education

Faculty of Educational Sciences

University of Oslo

Norway

2023

© **Rebecca Knoph, 2023**

*Series of dissertations submitted to the
Faculty of Educational Sciences, University of Oslo
No. 367*

ISSN 1501-8962

All rights reserved. No part of this publication may be
reproduced or transmitted, in any form or by any means, without permission.

Cover: UiO.

Print production: Graphic center, University of Oslo.

Acknowledgements

The PhD process has certainly taught me that hard things often cannot be accomplished alone. I owe my deepest gratitude to the many people who played instrumental roles during my PhD journey.

To my supervisors Joshua Lawrence and David Francis, who, after climbing the academic ladder of success themselves, have extended their hands to help others up. You inspire me to be the best version of myself, and to support others along the way. To Josh, you have helped me grow as an academic, an expat in Norway, and a parent; and for that I will always be grateful. To David, thank you for welcoming me to your team with open arms, and investing your time (and occasional car cup-holder change) in me. To my AVEL collaborators, Paulina Kulesz and Autumn McIlraith, thank you for your unwavering support throughout this process; especially your patience as I learned the ropes.

To my colleagues, thank you for listening to my endless barrage of thoughts. I won the office mate lottery not once but *twice* with Rachel Sweetman, who supported me at the beginning of my journey; and again with Germán Grande, who supported me until the end. Junyi Yang helped me find my way many times, literally in Valencia, and otherwise. Emily Oswald supported me through my first Norwegian teaching experience and along with Kristi Barcus gave me a safe space to be my authentic American self. Olga Mukhina, I cannot fathom coming anywhere close to success without your assistance in nearly every aspect related to UiO and the PhD project. To the TextDIM and LinCon research teams, who gave me space to think and learn when my work was still in progress—at best. To my fellow PhD peers, Tonje Amland, Andres Araos Moya, Sofie Bastiansen, Ymkje Haverkamp, Lisa-Marie Karlsen, Svitlana Kucherenko, Natalia Latini, Dilman Nomat, Hedda Wahl, and so many others at Helga Engs: without opportunity to make friendships with you, discuss my work with you, and simply engage in fun conversations with you, this dissertation would have been abandoned long ago. To Marina Prilutskaya up in Bodø, thank you for allowing me to be a part of *your* PhD process, so that I could be inspired to embark on mine.

Life unfolded amidst my PhD journey, too—an international move, a global pandemic, and a child—all forced me into paradigm shifts and introduced me to the friends I didn't know I needed. To my American barselgruppe, thank you for supporting me during moments when being a mom away from home was overwhelming. To my parent friends at Tyrihans, thank you for encouraging me when academia and motherhood seemed insurmountable. Ane, you saved me on more than one occasion when times were brutal. Liz, you have supported me through the trials and tribulations of my twenties, which needs no more explanation. Ingrid, you helped me accept that I cannot be everything, all at once, all of the time.

To my mom, who was unmatched in turn of phrase; and my dad, who had an unparalleled affinity for mathematics: I don't know how I managed to find a job that combines your loves for words and numbers, but I guess I did it. I miss you, and I hope I make you proud.

To my husband, Martin, who has made immense sacrifices in our decade together. I could not have finished this work without your support—both emotionally when I didn't feel I was cut out for this work, and physically when you took on the challenge of being a “single dad” with minimal support. And to my case study of one, Tobias: you taught me that life is more than the pursuit of perfection. I recognize that you too made sacrifices during this time, but I hope this work is proof that your mom is up for incredible challenges, and with dedication and help from your friends, you can accomplish great things.

Rebecca Knoph

Summary

Through three empirical papers, this dissertation explores the relationships between lexical features and vocabulary knowledge in diverse student groups.

Paper I examines multiple exploratory factor analyses on 22 non-behavioral lexical characteristics across three word lists—the General Service List (West, 1953), Academic Word List (Coxhead, 2000), and Academic Vocabulary List (Domain-Specific subsample; Gardner & Davies, 2014)—as a proxy of the different words we encounter. Across the three lists, the models indicated five related but distinct and stable latent factors: Frequency, Complexity, Proximity, Polysemy, and Diversity. However, for domain-specific words, Polysemy and Diversity combined into a single factor.

Papers II and III demonstrate the methodological utility of the latent dimensions by modelling target-word characteristics as a predictor of item difficulty. In both studies, explanatory item response theory bridges the gap between educational research on individual students' reading and language proficiency and cognitive research on individual words' lexical features. Although methodologically similar, the studies diverge based on how vocabulary knowledge is assessed and who is sampled in the studies.

Paper II reports cross-classified interactions between the reading proficiency scores of monolingual English speakers and the five latent dimensions on a synonym identification task. This study found that lexical features influence item difficulty, such as identifying the synonyms of more polysemous and frequent target words, and that students with higher reading comprehension scores found items less difficult overall. More importantly, this study found that student and target-word characteristics interact: Students with low reading comprehension scores were less sensitive to lexical advantages, such as word frequency, and more susceptible to disadvantages, such as word complexity.

Paper III compares student responses on the synonym task to a similar definition task, then explores interactions between the five latent lexical dimensions and English Language Learner classifications across the two tasks. Overall, students found synonym identification easier than definition identification when assessed on identical words. However, the lower-proficiency ELL students did not exhibit a preference, instead finding both tasks equally difficult. This study also found that identifying synonyms when a target word has many meanings was easier (similar to Paper II). When comparing lexical characteristic effects across ELL classifications, the lowest-proficiency ELL group (Limited English Proficient students) were less sensitive to the advantages of word frequency, but also less sensitive to the burden of word complexity than their peers.

Table of Contents

Part I: Extended Abstract

1. Introduction	1
1.1 Vocabulary Knowledge and Reading Comprehension	1
1.2 Second Language Vocabulary and Reading	3
1.3 Vocabulary in the Academic Setting	4
1.4 Thesis Objective and Overview	5
2. Theoretical Frameworks and Prior Research	6
2.1 Reading Comprehension in the Reading Systems Framework	6
2.1.1 Knowledge Systems	7
2.1.2 Word Identification Systems	8
2.1.3 Representation Systems	9
2.1.4 Relationships across Reading Development	11
2.2 Development of Vocabulary Knowledge in Multiple Components	12
2.2.1 Vocabulary Breadth	13
2.2.2 Vocabulary Depth	17
2.2.3 Vocabulary Mastery	22
2.2.4 Vocabulary Integration	23
2.2.5 Factors Affecting Vocabulary Learning	23
2.3 Vocabulary Assessment	24
3. Methodological Considerations	26
3.1 Overarching Issues	26
3.1.1 Defining a Critical Unit of Language	27
3.1.2 Selecting Vocabulary Assessment(s)	29
3.1.3 Measuring Reading Comprehension	32
3.1.4 Explanatory Item Response Theory	33
3.2 Study-specific Issues	35
3.2.1 Exploratory versus Confirmatory Approaches in Latent Estimation	35
3.2.2 Allowance of Missing Data in Factor Analyses	36
3.2.3 Assessing the Monolingual Subsample	37
3.2.4 Learner Classification in the Multilingual Sample	38
3.3 Ethical Dilemmas	39
4. Main Features of the Papers	41
4.1 Paper 1: Latent Lexical Dimensions	41
4.2 Paper 2: Lexical Relationships for the Monolingual Subsample	46
4.3 Paper 3: Lexical Relationships across English Language Learners	47
5. Discussion	50
5.1 Methodological Contributions	51
5.1.1 Estimated Latent Lexical Dimensions are Valuable and Practical	51
5.1.2 Explanatory Item Response Theory is Complex but Advantageous	52
5.2 Theoretical Contributions	53
5.2.1 Lexical Characteristics can be Empirically Grouped into Dimensions	53
5.2.2 General Academic Words are Not Equally Difficult	53
5.2.3 Academic Vocabulary Learning is not Linear	54
5.2.4 Monolingual and Multilingual Learning is Qualitatively Different	55
5.3 Implications and Future Directions	55
6. References	57

Part II: Papers

Paper I.....86

Knoph, R.E., Lawrence, J. F., & Francis, D. (2023). The dimensionality of lexical features in general, academic, and disciplinary vocabulary. *Manuscript Submitted to Scientific Studies of Reading*.

Paper II.....138

Lawrence, J. F., Knoph, R. E., McIlraith, A., Kulesz, P. A., Francis, D. J. (2022). Reading comprehension and academic vocabulary: Exploring relations of item features and reading proficiency. *Reading Research Quarterly*, 57(2), 669–690. <https://doi.org/10.1002/rrq.434>

Paper III.....162

Knoph, R.E., Lawrence, J. F., & Francis, D. (2023). How we measure vocabulary matters: Exploring differential effects of task type, English proficiency, and lexical characteristics. *Manuscript submitted to Language Learning*.

List of Tables and Figures

Figure 1. The Reading Systems Framework.....	7
Figure 2. An Example of the Situational Model.....	10
Figure 3. A Model of the Development of Vocabulary Knowledge	13
Figure 4. Example of Different Units of Language	27
Figure 5. The Explanatory Item Response Theory Model for Paper II.....	34
Figure 6. The Explanatory Item Response Theory Model for Paper III.....	35
Figure 7. Results of the EFA for GSL words	42
Figure 8. Results of the EFA for AWL words	44
Figure 9. Results of the EFA for domain-specific words	45
Table 1. Estimated latent factor scores for the word "elevator".....	52

Part I

Extended Abstract

1

Introduction

Vocabulary knowledge is fundamental to reading comprehension and academic achievement (Grabe, 2012; Pearson et al., 2007), and understanding the variety of language experiences that students have is both practically important for educators, and theoretically interesting to researchers. The words students encounter, the way students learn novel words, and how vocabulary knowledge is assessed varies greatly across contexts. Thus, when it comes to understanding vocabulary knowledge, considering what words we measure, which students we are measuring, and how we are measuring their knowledge is imperative. The aim of this dissertation is to explore word characteristics, learner differences, and variations in assessment simultaneously to further our understanding of the interplay between these factors, especially within the academic setting.

1.1 Vocabulary Knowledge and Reading Comprehension

It is well-established that vocabulary knowledge correlates with reading comprehension (Nation & Snowling, 2004; Ouellette, 2006; Joshi, 2005; McKeown, Beck, Omanson, & Perfetti, 1983; Quinn et al., 2015; Cain & Oakhill, 2014; Cromley & Azevedo, 2007; Kieffer

& Box, 2013; see Jeon & Yamashita, 2014 for a meta-analysis on second language learners). However, the relationship between vocabulary knowledge and reading comprehension is neither direct nor in one direction. In their seminal work, Anderson and Freebody (1981) formulated three hypotheses that continue to permeate reading research: the verbal aptitude hypothesis, the knowledge hypothesis, and the instrumental hypothesis.

The verbal aptitude hypothesis states, "Vocabulary performance is a reflection of verbal ability, and it is verbal ability that mainly determines if text will be understood" (Anderson & Freebody, 1981, p. 6). Higher-order ability is related to reading comprehension and vocabulary knowledge separately, which in turn manifests as a correlation between reading comprehension and vocabulary. For example, students with larger working memory capacity can free up cognitive space for higher-level comprehension, such as inferencing (Cain, Oakhill, & Bryant, 2004), and simultaneously leverage their working memory to learn better and retain new vocabulary (Baddeley, 2003). Similarly, students with high levels of metalinguistic awareness tend to be strong readers and possess more vocabulary knowledge (Nagy, 2007; Bialystok, Peets, & Moreno, 2014).

The knowledge hypothesis states that "The person who scores high [on vocabulary measures] has a deeper and broad knowledge of the culture [which is] crucial for text understanding" (Anderson & Freebody, 1981, p. 7). In this case, knowledge about the world causes both strong reading comprehension skills and vocabulary knowledge. Following this concept, van Dijk & Kintsch's (1983) situation model posits that we interpret a new text by relating it to existing background knowledge (Kintsch, 1988; 2014). Similarly, O'Reilly, Wang, and Sabatini (2019) suggest that readers must meet some threshold of background knowledge to understand a text, even if the individual vocabulary words themselves are relatively understood. For example, reading a statistics textbook requires understanding new concepts using words many students already know, such as "slope" and "significant." Moreover, both strong and weak readers perform better when reading about familiar topics (Recht & Leslie, 1988).

The instrumental hypothesis states that "knowing the words enables text comprehension" (Anderson & Freebody, 1981, p. 6). Indeed, current research estimates that, in order to read an authentic text independently, a reader needs to know 95–98% of words in a text (Schmitt, Jiang, & Grabe, 2011; Laufer & Ravenhorst-Kalovski, 2010; Nation, 2006; Schmitt et al., 2017). While the raw number of words a reader must know varies between texts, researchers posit that a reader needs to know somewhere between three thousand (Laufer, 1992) and nine thousand (Hu & Nation, 2000) word families (i.e., "talk", "talked", and "talks" are one word family).

At the same time, there is evidence that reading comprehension impacts vocabulary growth, indicating a reciprocal relationship between the two (Biemiller & Slonim, 2001; Nagy & Scott, 2000; Verhoeven, van Leeuwe, & Vermeer, 2011; Stanovich, 1988, 2000). When we know more words in a text, we have more information to leverage in the process of inferring meanings of novel vocabulary (Cain, Oakhill, & Lemmon, 2004). Additionally, readers who know more words can also read more fluently and automatically, which in turn positively impacts reading comprehension (Klauda & Guthrie, 2008). It is evident that vocabulary knowledge and reading comprehension are not simply related by one direct path, but instead through many reciprocal, moderating, and mediating effects, as we see in many hypothesized reading models (Cromley & Azevedo, 2007; Ahmed et al., 2016; Joshi, 2005; amongst many others).

Moreover, the strength and structure of the relationship change as language skills improve. When initially learning to read, readers spend nearly all of their working memory on decoding slowly and laboriously, with fragmented word knowledge and minimal integration (Nagy & Scott, 2000). As a reader becomes more skilled, semantic knowledge becomes more structured; learning becomes more efficient, and making inferences comes to the forefront (Cain, Oakhill, & Bryant, 2004). Conversely, consolidating the meaning of new words can remain difficult for students with poor reading comprehension; therefore, retaining meanings or engaging in higher-order thinking becomes challenging (Nation, Snowling, & Clarke, 2007).

1.2 Second Language Vocabulary and Reading

Growth in vocabulary knowledge and reading comprehension are reciprocally related for multilingual learners across languages as well (Nation, 2011, 2022; Schmitt, 2008; Laufer, 1992; Carlisle & Beeman, 2000; Qian, 2002); however, second language (L2) learning is accompanied by additional factors that complicate the language-acquisition process. A re-evaluation Anderson and Freebody's (1981) verbal aptitude, general knowledge, and instrumental hypotheses for multilingual learning exemplifies this fact.

In relation to the verbal aptitude and general knowledge hypotheses, multilingual learners can often utilize their linguistic knowledge and background knowledge initially coded in their first language (L1) when reading in their L2 (Lucas & Katz, 1994). When orthographies between languages are similar, decoding skills are transferrable between languages (Cummins, 2007; Kuo & Anderson, 2006). Moreover, multilingual learners often show an advantage in metalinguistic awareness (Adesope et al., 2010; Reder et al., 2013; Galambos & Goldin-Meadow, 1990; Nagy, 2007; Cummins, 1978).

Finally, while the instrumental hypothesis posits that knowing more words enables text comprehension, how vocabulary knowledge develops differs between L1 and L2 (Webb & Chang, 2012). Initially, multilingual learners often rely on direct vocabulary instruction and specific texts selected by their teachers (Laufer, 2003), and use direct translations between a novel L2 target word and their closest L1 translation (Ringbom, 1987,1992, 2007; Jarvis & Pavlenko, 2008). As language proficiency increases, learners increase and strengthen their vocabulary connections in the L2 (Joyce, 2018) and move towards more independent and automatic use of the L2. When shifting from direct L1 translations to more L2 lexical connections, learners begin to understand language-dependent nuances lost in literal translations—a problem that never exists for monolingual speakers.

Hence, it is clear that understanding how L1 speakers know words is not necessarily representative of how multilingual learners know words, even in the same language. Instead, research must consider diverse learners across diverse linguistic contexts, which requires more sophisticated modelling, which I will discuss in the Methods chapter.

1.3 Vocabulary in the Academic Setting

Reading is an important way to learn new vocabulary (Brett, Rothlein, & Hurley, 1996; Laufer, 2003) and is especially crucial in the school setting via textbook reading (Alexander, 2012; Weisberg, 2011). In addition, as students advance through the educational system, grade-appropriate textbooks become increasingly complex and therefore require stronger reading skills to comprehend (Goodwin, Gilbert, & Cho, 2013; Bar-Ilan & Berman, 2007; Landauer, Kireyev, & Panaccione, 2011).

Texts increase in complexity partly by introducing more morphologically complex and conceptually abstract vocabulary (Nagy & Townsend, 2012; Carlisle, 2000). For example, Marzano & Simms' (2011) *Vocabulary for the Common Core* list provides “circle” as a core mathematics vocabulary word for kindergarten and “circumference” for fifth grade. Students are expected to improve reading comprehension and increase vocabulary knowledge throughout school; for example, students in early grades learn new root words, while students in later grades shift their focus to more complex derivations (Kuo & Anderson, 2006; Anglin, Miller, & Wakefield, 1993). However, progression through the educational system does not *guarantee* linguistic preparedness for advanced texts, particularly in the case of English Language Learners (ELLs), who experience unique challenges in English vocabulary learning compared to their monolingual peers.

1.4 Thesis Objective and Overview

This dissertation explores relationships between lexical features and the words diverse students know in different ways, particularly in the academic setting. Across the three empirical papers, I address the following questions:

- Can we develop a factor structure for the lexical characteristics of words, and, is the structure consistent across English vocabulary contexts, such as conversational versus academic English?
- How do these lexical characteristics impact vocabulary-item difficulty, and, is the impact consistent across reading comprehension levels for monolingual students?
- Does the impact of lexical characteristics on item difficulty depend on how we assess knowledge, and, does the impact change as a function of English proficiency when assessing ELLs?

The dissertation consists of two parts: the extended abstract, which explains how the empirical papers add to the literature, followed by the three empirical papers, which address the above research questions in sequence.

The extended abstract contextualizes, interprets, and synthesizes the complex information across three empirical studies. The first chapter has introduced and described the main objectives of the dissertations as a whole. The next chapter will situate the dissertation within relevant reading comprehension and vocabulary development frameworks. Chapter 3 details the research methods across the papers, clarifies the reasoning behind methodological and analytical decisions, and explains ethical dilemmas that arose during the project. Chapter 4 reviews and comments on the main findings of the papers. Finally, the concluding chapter elaborates on the main findings and contributions and extends into perspectives on teaching and assessment.

The three empirical papers in Part II represent significant contributions to the research community, either as peer-reviewed and published papers or as papers still under review at the time of thesis submission. Paper I discusses the factor structures of lexical features for different types of words using exploratory factor analyses. Paper II explores the relationship between lexical characteristics and vocabulary-item difficulty, particularly by evaluating varying levels of reading comprehension for a monolingual subset. Finally, Paper III expands upon previous works with a similar analytical methodology but also assesses the relationship between lexical characteristics and vocabulary-item difficulty for ELLs, particularly the differences in relationships between different English-learner classifications and across two receptive vocabulary tasks.

2

Theoretical Frameworks and Prior Research

Although many competing models for reading comprehension and language acquisition frameworks proliferate research, this thesis is situated within the *Reading Systems Framework* (Perfetti & Stafura, 2014) for understanding the complex task of reading and a componential framework for understanding the development of vocabulary knowledge (Nation, 2022). I will summarize these frameworks in turn and explain how situating the current work within the frameworks aligns with current research.

2.1 Reading Comprehension in the Reading Systems Framework

The Reading Systems Framework is a general framework of reading comprehension that builds upon the top-down, higher-level processes of reading comprehension and the bottom-up, word-level processes, initially developed in Perfetti's (1999) *Blueprint of the Reader*. It integrates knowledge-driven research from the Situation and Construction-Integration Models (van Dijk & Kintsch, 1983; Kintsch, 1988; Kintsch & Rawson, 2005; see McNamara & Magliano, 2009 for a review) with text-driven processes, such as those in Gough and Tunmer's (1986) *Simple View of Reading* as well as with decoding, phonological awareness, and sight

recognition in Scarborough’s (2001) *Reading Rope* model and Duke & Cartwright’s (2021) *Active View of Reading* model. Unlike more individual word-reading models, this more general reading comprehension model highlights the importance of multiple “knowledge sources, basic cognitive and language processes, and the interactions among them” (Perfetti & Stafura, 2014, p. 24)

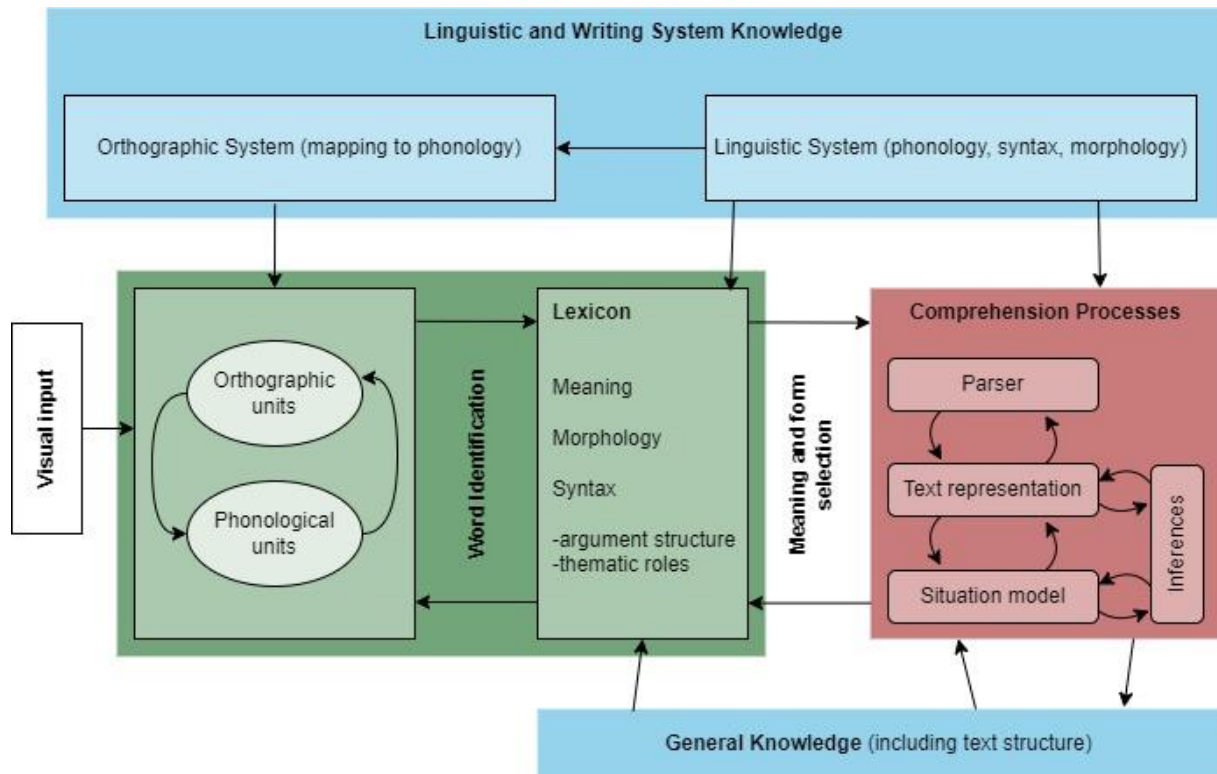


Figure 1. The Reading Systems Framework.

Note: This figure is reproduced from Perfetti & Stafura (2014) and colorized by me.

2.1.1 Knowledge Systems

Readers use a meaning-driven, top-down approach to comprehend text via their linguistic system knowledge, writing system knowledge, and general background knowledge (Perfetti & Stafura, 2014; Stanovich, 1980), which are represented in the blue boxes at the top and bottom of Figure 1. Knowledge about the linguistic system includes information about morphology, phonology, and syntax, which maps onto the orthographic (written) system. General knowledge includes information about the current text structure, the relevant writing genre, and the world in general.

In the *Reading Systems Framework*, linguistic system knowledge directly impacts our orthographic system knowledge, our vocabulary knowledge in the lexicon, and our comprehension processes (Perfetti & Stafura, 2014). Readers with high morphological

awareness are better able to infer the meaning of new words by splitting up words into known parts; for example, knowing that “jumped” contains the root word “jump” and the suffix “ed” to indicate “jump in the past” (Nation, Snowling, & Clarke, 2007; Fowler, Liberman, & Feldman, 1995; Kuo & Anderson, 2006; Bhattacharya & Ehri, 2004). Efficient readers who excel at decoding also free up cognitive resources to engage in higher-order processes, such as using the working memory to make inferences instead of laboriously parsing sentences (Hamilton, Freed, & Long, 2016; Bohn-Gettler & Kendeou, 2014; Ehri, 1995).

Linguistic knowledge alone is not sufficient to ensure text comprehension. In the *Reading Systems Framework*, general knowledge directly impacts vocabulary knowledge in the lexicon and is reciprocally related to the comprehension processes (Perfetti & Stafura, 2014). Background knowledge is a key component in many reading comprehension models (Cromley & Azevedo, 2007; Kintsch & Rawson, 2005; Myers & O’Brien, 1998). For example, in O’Reilly, Wang, and Sabatini’s (2019) study of 3,534 high-school students reading an ecology text, they found that reading comprehension was nearly impossible below a certain threshold of background knowledge. In essence, while students could know the individual words in the text, without the necessary background knowledge to connect to, students were unable to process and learn the new content.

2.1.2 Word Identification Systems

The next fundamental component in the *Reading Systems Framework* is individual word comprehension, represented as the green box in Figure 1. Knowing both the form of a word (orthographic and phonologic) and the meaning(s) of individual words is crucial to comprehending more complex text, such as sentences, paragraphs, etc., from a bottom-up approach (Perfetti & Stafura, 2014; Stanovich, 1980). The word identification system proposed by Perfetti and Stafura (2014) parallels earlier work, coined the *Lexical Quality Hypothesis* (Perfetti, 2007; Perfetti & Hart, 2002).

The *Lexical Quality Hypothesis* originated after observing that “when either children or adults were separated by their scores on a reading comprehension test, they sorted themselves also on their speed of written word and pseudo-word identification” (Perfetti & Hart, 2002; p. 189). The *Lexical Quality Hypothesis* suggests that word identification involves retrieving three constituents: orthography, phonology, and semantics. High-quality word representations require fully-specified and tightly-bound constituents, and a “skilled reader, in addition to having foundational resources (decoding, spelling, and grammatical skills), is one who has many high-quality word representations” (Perfetti & Hart, 2002; p. 192). Readers with stronger

lexical representations access words more efficiently during reading, while less skilled readers who struggle with individual word reading show deficits in overall text comprehension (Perfetti, 2007). Hence, understanding how readers map orthographic forms onto lexical representations is key (Perfetti & Adlof, 2012).

At the same time, a perfect one-to-one mapping between orthography, phonology, and meaning is not common in English. For example, a single orthographic form can map onto two different phonological forms (e.g., “crayon” can be pronounced CRAY-on or crown), a single phonological form can map onto two orthographic forms (e.g., “doughnut” and “donut” can both be pronounced DOW-nuht), and a single orthographic and phonological form can map onto multiple meanings (e.g., a vehicle “train” and the action “to train”). Priming research indicates that, even for the skilled reader, words with different phonologies and meanings are simultaneously and subconsciously activated (at least partially) when they share orthographies, which can lead to confusion in comprehension (Gernsbacher & Faust, 1991).

Recognizing the written and spoken forms of a word depends on both the decoding skills of the reader and the decodability of the word (Wang et al., 2013; Bhattacharya & Ehri, 2004). Longer words take more time to decode (Just & Carpenter, 1980; Hyönä & Olson, 1995), as do opaque words (i.e., not orthographically transparent; Spencer & Hanley, 2003), and readers are more likely to fixate longer on less-frequent words (Joseph, Nation, & Liversedge, 2013). However, word identification also requires retrieving the conceptual meaning. It is possible to decode a word but not activate the meaning—in fact, this is precisely what happens during lexical decision tasks with nonwords. Previous eye-tracking and semantic priming research has found that some poor readers do not lack decoding skills but instead spend extra time and cognitive resources in processing the meaning of the word, which leaves fewer resources for higher-order processing in the representation systems and results in poorer reading comprehension (Nation & Snowling, 1999; Veldre & Andrews, 2014; Walczyk et al., 2004).

2.1.3 Representation Systems

The last component of the *Reading System Framework* is connecting text-independent knowledge and word identification to van Dijk & Kintsch’s (1983) *Situation Model* by creating and updating mental representations of the text as a reader progresses through a reading (Perfetti, 1999), represented in Figure 1 as the red box. As a reader parses information, they form a text base (memory of what the text said) and a situational representation (a model of what situation is occurring). The text base and situational representation continuously inform one another and update as the reader progresses through the text, connects it to their general

and linguistic knowledge, and makes inferences (Perrig & Kintsch, 1985; Perfetti, 1999). Figure 2 exemplifies how the text base and situational representation might inform each other via inferencing as a reader processes a particularly tricky “garden path” sentence (Ferreira & Henderson, 1991): “The old man the boat.”



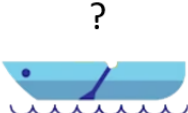

Processed Text	The old man	the boat	(backtracking)	
Potential Inferencing	The “man” is the subject of the sentence	There are two subjects and no verbs?	“Man” is not referring to the noun, but the verb “to operate”	“The old” is referring to a group of elderly people, not describing a single person
Situational Representation				
Textual Base	The old man	Old man and boat	Someone mans the boat	Old people run the boat

Figure 2. An Example of the Situational Model.

Note: Stock images from freepik.com for free commercial use.

For many readers, comprehending this sentence initially involves incorrect assumptions about the meanings of the words “old” and “man”, and subsequent reinterpretations based on linguistic knowledge about sentence structure (Ferreira, Christianson, & Hollingworth, 2001). In Figure 2, it first seems as though the sentence does not contain a verb, so a reader must recognize both that English sentences usually require a verb, and that “man” can also refer to an action, not just a person. Skilled readers are more likely to notice and re-read when ambiguous words break down comprehension (Joseph & Liversedge, 2013; Hacker, 1997).

Perfetti and Stafura recognized the interrelatedness between the components of reading comprehension, stating: “even with the best of efforts it is difficult to persuasively assess processes in isolation of other processes” (2014; p. 25). Hence, when examining the Reading Systems Framework, only one component features a single, unidirectional arrow: the visual input itself. Even in the caption of Perfetti’s (1999) original figure, he states: “In particular, whether bidirectional arrows are needed everywhere is an empirical question” (p. 169). While this thesis does not directly address multiple bidirectional arrows, this comment implies that reading comprehension is indeed highly complex and varies as a function of multiple components, both from the individual person and the individual text.

2.1.4 Relationships across Reading Development

The Reading Systems Framework is more process-oriented than development-oriented (Wang et al., 2019); it is centered more on the *current* text than the development of reading comprehension skills over time. However, relationships between the knowledge, word identification, and representation systems (and their subcomponents) are not time-invariant. This section reviews some empirical findings of how these relationships change over time.

Previous research indicates that decoding skills are strongly related to reading ability in early school grades when students are learning to read (Ehri, 1995, 2014), but as students approach decoding mastery and automaticity, it begins to take a backseat to higher-order processes, such as inferencing (García & Cain, 2014; Aarnoutse et al., 2001; Foorman, Petscher, & Herrera, 2018; Schatschneider et al., 2004; Gough & Tunmer, 1986). In their meta-analysis of 110 studies, García and Cain (2014) found that reading comprehension and decoding skills correlated strongly ($r = 0.80$) for readers under ten years old but weakly ($r = 0.47$) for older readers (where the distinction between strong and weak readers is more evident; Chall, Jacobs, & Baldwin, 1990). Likewise, Wang and colleagues (2019) proposed the Decoding Threshold Hypothesis, wherein a reader must meet some “minimum level of decoding skill before higher-level processing is operational” (p. 389). In both a cross-sectional study of over 10,000 students and their longitudinal study of over 30,000 students in grades 5–10, students who met the decoding threshold showed a positive linear relationship between decoding skills and reading comprehension (Wang et al., 2019). In contrast, students below the threshold exhibited no relationship and minimal growth in reading comprehension.

As reading skills develop and texts become more complex, vocabulary and background knowledge dominate in reading comprehension (Cromley & Azevedo, 2007; Ahmed et al., 2016; Stanovich, 1988; 2000; Leach, Scarborough, & Rescorla, 2003; Buly & Valencia, 2002; Sabatini et al., 2010). In their Direct and Inferential Mediation model, Cromley and Azevedo (2007) found that vocabulary and background knowledge were stronger predictors of reading comprehension than word-reading skills (direct paths = 0.366, 0.234, and 0.151, respectively) for 175 students in 9th grade and that struggling readers at this age were more likely to struggle with vocabulary knowledge than with word reading skills—replicating previous findings (Leach, Scarborough, & Rescorla, 2003; Buly & Valencia, 2002). At the same time, younger readers are less likely to know low-frequency words (Jenkins et al., 2003; McNamara, Graesser, & Louwse, 2012), the development of cognitive skills, such as inferencing, also improves as the brain develops (Carretti et al., 2014), and world knowledge accumulates as learners gain more life experience. The fact that relationships between multiple components and skills with

reading comprehension evolve over time highlights the complexity and intertwining of person and text characteristics in reading comprehension.

Developing reading comprehension skills for the multilingual learner also leverages qualitatively different mechanisms compared to monolingual reading development. Multilingual learners possess linguistic/writing-system knowledge and mental lexicons, to some degree, in at least two languages (Borodkin et al., 2016), and general knowledge can be encoded or related to multiple languages. As a result, multilingual learners have the opportunity to connect words across languages. Initially, learners directly map new words in their L2 onto their corresponding L1 translations (Jiang, 2002; Kroll et al., 2010; Kroll et al., 2002). As their proficiency increases, multilingual learners require less scaffolding via translation and instead become more automatic and fluent users in their L2. It is therefore crucial that research involving multilingual readers consider not only how multiple components of reading comprehension relate to one another when comprehending a single text or how they relate to a single language throughout reading development but also the added complexity of developing reading comprehension skills in multiple languages.

2.2 Development of Vocabulary Knowledge in Multiple Components

Vocabulary knowledge plays a key role in every system of the Reading Systems Framework (Perfetti & Stafura, 2014), yet understanding how we develop vocabulary knowledge is not clear from the framework alone. Vocabulary knowledge itself is a complex phenomenon; therefore, there are many depictions in terms of the dimensions, aspects, components, or constructs involved in knowing vocabulary. For example, Meara (2005) presented vocabulary knowledge as the three components: size, organization, and accessibility, while Daller et al. (2007) proposed breadth, depth, and fluency, and Schmitt (2014) suggested an interrelation between components. Conversely, Laufer and Goldstein (2004) argued that vocabulary knowledge is not based on the word but on the skill, suggesting four elements: passive recognition, active recognition, passive recall, and active recall; later reformulated by Schmitt (2010) as form recognition, form recall, meaning recognition, and meaning recall.

Nation's (2001) three-component framework is one of the most prominent frameworks for vocabulary knowledge components: form, meaning, and use. This component framework posits that individual word knowledge is related to the word form (e.g., "What does the word look like?"), the word meaning (e.g., "What meaning does this word refer to?"), and usage (e.g., "Where, when, and how often can we use this word?"; Nation, 2022). However, these components only encompass vocabulary knowledge about a single word. In the latest edition

of his *Learning Vocabulary in Another Language*, Nation (2022) suggests that the development of vocabulary knowledge as a whole requires four parts: depth (the components summarized above), breadth (the number of words known), strength (how well words are known), and integration (relations to other known words; p. 89). Figure 3 illustrates Nation’s (2022) model of the development of vocabulary knowledge, and the following sections discuss each of the parts in turn.

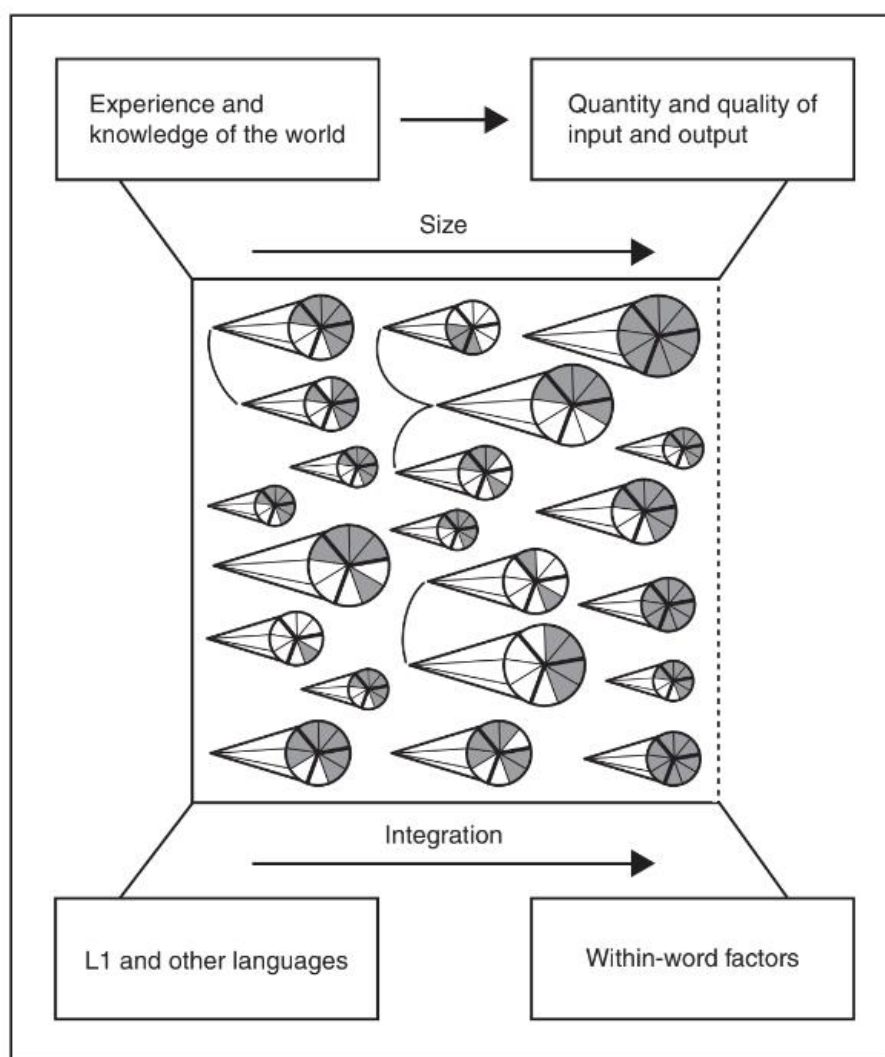


Figure 3. A Model of the Development of Vocabulary Knowledge.

Note: This figure is reproduced from Nation (2022, p. 89)

2.2.1 Vocabulary Breadth

Numerous studies have shown that vocabulary size (breadth) is one of the best predictors of reading comprehension for children in elementary school (Ouellette & Beers, 2010; Verhoeven & van Leeuwe, 2008) and above (van Gelderen et al., 2007). In addition, meta-analyses have also indicated that improving vocabulary size via educational intervention

improves reading comprehension (Stahl & Fairbanks, 1986; Wright & Cervetti, 2017; Elleman et al., 2009).

However, the English language contains an immense number of words; hence, learning all words is daunting—if not impossible. For example, *Webster's Third New International Dictionary* (1961), contains over 450 thousand words across 267 thousand distinct entries (Dupuy, 1974; Goulden, Nation, & Read, 1990). Nation and Coxhead (2021) report that vocabulary size estimates for adult native speakers range from 12,000 to over 200,000 words, depending on how individual words are distinguished (e.g. grouping word families or lemmas). Vocabulary breadth is represented in Figure 3 as the number of individual cones.

Even so, vocabulary breadth is defined not just by the sheer number of words you know but by *which* words. In their seminal work, Beck, McKeown, and Kucan (2002) identified three tiers of words: basic, general academic, and disciplinary. The three tiers vary in terms of when readers are most likely to encounter vocabulary and the lexical features of that vocabulary, on average (Hiebert & Fisher, 2005). This in turn impacts how, when, and if vocabulary instruction is necessary in the classroom (Coxhead, 2000; Hiebert et al., 2018).

Everyday English. Conversational, everyday English words (basic Tier I words; Beck, McKeown, & Kucan, 2002) are the high-frequency words we use in everyday conversation and “rarely require instructional attention to their meanings in school” (p. 8). However, a lack of instructional attention to meanings does not imply a lack of instructional attention to other word aspects, such as decodability or phonological awareness (Melby-Lervåg, Lyster, & Hulme, 2012).

Paper I samples these words via the General Service List (GSL; West, 1953), which is a compiled list of high-utility and high-frequency words containing “1,907 main entries and 3,751 orthographically different words” (Gilner, 2011, p. 71). The GSL was compiled both based on objective frequency estimates and subjective criteria about utility and accessibility. While it is often criticized for being outdated and subjective (Richards, 1974; Gilner, 2011; Gardner & Davies, 2014), it is undeniable that the GSL has permeated educational research and pedagogical materials for decades, as Gilner states: “Indeed, the GSL is part of the collective consciousness and one would be hard pressed to find someone in the field who has not heard of this word-list” (2011, p. 80).

Most words we encounter, regardless of age or context, are these words. Researchers estimate that the GSL accounts for nearly 90% of spoken corpora (Nation, 2004), 80% of written corpora (DeRocher, 1973), and almost 70% of academic corpora (Coxhead & Hirsh,

2007). Even with these large estimates, 10–30% of English tokens are left unaccounted for, thus necessitating the distinction of the remaining two tiers.

Academic Vocabulary. Academic vocabulary is “used in academic settings that facilitates communication and thinking about disciplinary content” (Nagy & Townsend, 2012, p. 92). Knowing academic vocabulary is imperative for success in school (Schleppegrell, 2004), starting in primary grades (Biemiller, 2012), middle school (Townsend & Collins, 2009), and beyond. Academic vocabulary is encountered less frequently than conversational English but is also more morphologically complex (Hiebert, Goodwin, & Cervetti, 2018; Nagy & Anderson, 1984) and abstract (Nagy & Townsend, 2012; Lawrence, Maher, & Snow, 2013). Abstract words can be more difficult to process, given that learners are less likely to have associated memory contexts from prior exposures and experiences (Schwanenflugel, Harnishfeger, & Stowe, 1988; Crosson et al., 2019). Additionally, academic vocabulary is often divided into two categories: general academic and domain-specific vocabulary (Nagy & Townsend, 2012; Baumann & Graves, 2010; Snow & Uccelli, 2009; Gardner & Davies, 2014; Hiebert & Lubliner, 2008), as discussed next.

General Academic Vocabulary. General academic vocabulary includes words used more frequently in academic contexts but across multiple disciplines (Nagy & Townsend, 2012). These words are Tier II words, and are “high frequency words for mature language users” (Beck, McKeown, & Kucan, 2002; p. 16). While general academic words are of high utility for reading academic texts (Townsend et al., 2012), these words pose unique learning challenges. For example, general academic vocabulary provides language for academic skills, such as integrating ideas across texts to form arguments by way of connectives (e.g., “however”, “thus”, and “moreover”; Crosson, Lesaux, & Martiniello, 2008). Additionally, many general academic words have multiple meanings (i.e., are polysemous or homonymous), where a concrete meaning is more frequent (e.g., furniture *table*) and an abstract meaning is less frequent (e.g., data *table*); meanings can even vary by discipline (e.g., medicinal *culture* versus humanities *culture*; Lawrence, Maher, & Snow, 2013; Hyland & Tse, 2007).

Explicit instruction of general academic vocabulary is not as common as technical vocabulary (Hiebert & Lubliner, 2008), though it is possible and effective, particularly in middle-school grades, as we see with three widely-known vocabulary interventions: Word Generation, ALIAS, and RAVE (Snow, Lawrence & White, 2009; McKeown, et al., 2018; Lesaux et al., 2010, respectively). The Word Generation intervention focuses on learning general academic vocabulary in middle school by incorporating five novel vocabulary items weekly into activities across multiple classrooms (English Language Arts, Science, Math,

Social Studies) and through essay writing (Snow, Lawrence, & White, 2009), with vocabulary gains still evident over a year later (Lawrence et al., 2012). The Academic Language Instruction for All Students (ALIAS) intervention follows an incremental sequence to develop vocabulary knowledge: seeing the word in text, activating background knowledge, exploring meaning in context, introducing additional meanings, morphological analysis, and ultimately production in independent writing (Lesaux et al., 2010). The Robust Academic Vocabulary Encounters (RAVE) intervention teaches academic vocabulary by focusing on multiple meanings, enriching semantic representations, building fluency, integrating meaning and context, and analyzing morphology (McKeown et al., 2018). Students who participated in RAVE demonstrated more efficient lexical access to instructed target words, greater knowledge of instructed target words, increased morphological awareness, and improved reading comprehension in general (McKeown et al., 2018).

Similar to the Word Generation, ALIAS, and RAVE interventions, Paper I samples general academic words via Coxhead's (2000) Academic Word List (AWL), and Papers II and III used the AWL to identify target words. The AWL is based on 3.5 million words of academic texts across four disciplines (Arts, Commerce, Law, and Science) and includes 570 headwords and their word families, such as "react" (the headword) with "reacted," "reaction," and "reactive" as some of the family members (Coxhead, 2000). Words were included when they frequently appeared in academic texts (at least 100 times in Coxhead's Academic Corpus) *and* across diverse disciplines (at least ten times in each discipline). However, given that the purpose of the AWL was not to describe all high-frequency words, but *academic* ones, words on the General Service List (West, 1953) were automatically excluded. As we have seen, the Academic Word List is widely-used in vocabulary interventions and is highly-regarded in educational research (e.g., Hiebert & Lubliner, 2008; Baumann & Graves, 2010; Nagy & Townsend, 2012; see Coxhead, 2011 for a decade review). However, to date, minimal research has examined lexical features of academic words and their impact on how well students know these words.

Discipline-Specific Vocabulary. Discipline-specific words, or technical words (Gardner & Davies, 2014) encompass the last tier of words (Beck & McKeown, 2002). These words are academic vocabulary with limited disciplinary range, in that they appear in only a few content areas (e.g., “hypotenuse,” “microorganism,” or “migratory”). Although these words are low-frequency *overall*, they can refer to crucial concepts in their fields; and they are often derived from Latin or Greek roots (Goodwin, Gilbert, & Cho, 2013), which can make guessing meaning difficult. Thus, curricula often require direct and explicit instruction of technical vocabulary, exemplified by Pearson, Hiebert, & Kamil (2007): “the bulk of text-centered science instruction is learning the meanings of hundreds of new scientific terms rather than experiencing the intellectual rush of hands-on inquiry”.

Paper I samples domain-specific words via a subset of the Academic Vocabulary List (AVL; Gardner & Davies, 2014). The AVL is based on a 120-million-word subset of academic texts from the Corpus of Contemporary American English (COCA). Words were excluded from the AVL if the frequency was not at least 50% higher in the academic portion of the COCA versus the non-academic and occurred with “at least 20% of the expected frequency in at least seven of the nine academic disciplines” (p. 315). Words that occurred more than three times the expected frequency in any discipline were then considered part of the domain-specific subset of the AVL (which we abbreviated as AVL-DS in Paper I).

2.2.2 Vocabulary Depth

Knowing many words is a necessary but not sufficient condition for success in reading comprehension (Schmitt, Jiang, & Grabe, 2011; Laufer, 1989; Hu & Nation, 2000). Previous research suggests that how well (or deeply) we know words independently predicts reading comprehension even after controlling for vocabulary breadth, decoding, and listening comprehension skills (Braze et al., 2007; Swart et al., 2017; Sénéchal, 2006). While vocabulary breadth and depth are highly correlated (Binder et al., 2017; McKeown et al., 2017) and reciprocal (Li & Kirby, 2015; Walley, Metsala, & Garlock, 2003), depth appears to be a stronger predictor, suggesting that how well we know words is more important than the number of unique words we know (Ouellette, 2006; Tannenbaum et al., 2006; Perfetti, 2007; Perfetti & Hart, 2002).

Nation’s (2022) model reflects the various aspects of word knowledge and depth as individual slices within each cone in Figure 3. Knowing a word encompasses understanding the form, meaning, and use of a word; each component comprises subcomponents that can be assessed either receptively or productively (Nation, 2001, Schmitt, 2019). This section gives

an overview of previous research related to lexical characteristics within this framework, though it by no means summarizes all characteristics in existence.

Form. The form of a word relates to its internal structure, such as the spelling, pronunciation, and word parts (Nation, 2022). Word length is one of the most common lexical characteristics related to word form—counting the number of letters, syllables, or morphemes in a word. Longer words take more time to decode, process, and remember (New et al., 2006; Ehri, 2005; Ellis, 2002) and read (Carlisle, 2000). Previous research also suggests that the learning burden for longer words is greater (Goodwin & Cho, 2016), in part because there is more information to remember. Additionally, longer words are more likely to contain prefixes and suffixes or be derivations (Ellis, 2016; Nagy & Anderson, 1984; Nippold & Sun, 2008).

Conversely, information theory posits that longer words are more likely to contain more information about meaning (Piantadosi, Tily, & Gibson, 2011; Mahowald et al., 2013), which could reduce the learning burden. By middle school, most readers no longer sound out words letter-by-letter or sound-by-sound; instead, they use larger units like syllables and morphemes to support their reading and understanding of morphologically complex words present within academic texts (Ehri, 2005). We find evidence of the reduced learning burden in Carlisle and Stone’s (2005) study of 2nd-6th graders who read phonologically-transparent derivations (e.g. “hilly”) more accurately than root words (e.g. “silly”) of the same length, frequency, and spelling.

Word length is not the only way to measure the form component in Nation’s (2022) model; the form can also be measured in relation to other words. Coltheart’s *N* is the orthographic neighborhood size, or the number of words that a target word can form by substituting a single letter (e.g., “stove” and “shove”; Coltheart, Davelaar, Jonasson, & Bessner, 1977), and words with many orthographic neighbors are processed faster in lexical decision tasks (Andrews, 1996; Dujardin & Mathey, 2022). Peereman and Content (1997) suggested a parallel measure for phonologic neighborhood size (substituting a single phoneme, e.g., “place” and “face”) and phonographic neighborhood size (substituting a single letter and phoneme, e.g., “stove” and “stone”). When comparing the multiple neighborhood size measures, Adelman and Brown (2007) found that phonographic neighborhood size produced unique facilitatory effects in word naming, while orthographic neighborhood size did not.

Given that phonological awareness is taught at young ages (see meta-analysis by Melby-Lervåg, Lyster, & Hulme, 2012), words in dense neighborhoods are more likely to be learned early (Storkel, 2004). For example, teaching phonological awareness includes rhyming games, which could facilitate not only learning the word “cat” but all of its phonological neighbors as

well. Additionally, Storkel (2001) and Edwards, Beckman, and Munson (2004) provide strong evidence that familiar phonotactic sequences in novel words facilitate learning in young children. Moreover, German and Newman (2004) found that older children (ages 8–12) were more likely to struggle in accessing and producing words in smaller neighborhoods, potentially because “those segments and segment combinations are not accessed frequently, and thus, may have relatively underdeveloped paths” (p. 633). And since short words tend to occur in more dense neighborhoods, some research suggests that the facilitative effect of word length may not be directly caused by length but by neighborhood size (Jalbert, Neath, & Suprenant, 2011).

At the same time, lexical priming studies indicate that words can prime one another in ways other than single substitutions, such as deletions/insertions (e.g., planet and plane) and transpositions (e.g., trail and trial; Davis, 2006), and that priming effects are not all-or-nothing (although neighborhood membership is binary). As a result, Yarkoni, Balota, and Yap (2008) suggested a new measure, the orthographic Levenshtein distance 20 (old20; and later the phonologic Levenshtein distance 20, pld20; Yap et al., 2012). These measures take the average number of insertions, deletions, and substitutions needed to move from one target word to its closest twenty neighbors. Across their entire dataset, Yarkoni, Balota, and Yap (2008) found the correlation between neighborhood size and Levenshtein distances was statistically significant ($p < .001$) but not necessarily strong ($r = -0.561$) and that the relationship between the two differed as a function of word length (e.g., monosyllabic words saw a near perfect correlation at $r = -0.925$). Hence, neighborhood sizes and Levenshtein distances, although conceptually related, are informative in different ways.

While ELLs are no more likely to struggle with decoding than their monolingual peers (Strange & Shafter, 2008), they are more likely to select distractors based on form as opposed to meaning or semantics, especially at lower levels of proficiency (Ellis & Shintani, 2013; Henning, 1973). Still, poor reading comprehension can result from more than just issues with decoding (Spencer & Wagner, 2017); there is more to vocabulary than the orthographic and phonologic form.

Meaning. A word's meaning component comprises the form-meaning link, concept and referents, and associations (Nation, 2022). Words are often quantitatively measured in terms of the number of meanings, as it is common for words to have multiple meanings (Youn et al., 2016). However, current research shows that multiple meanings can both enhance and hinder vocabulary knowledge (see Eddington & Tokowicz, 2015 for a review). Learning polysemous words may be an example of the trade-off between short-term performance and long-term learning (Soderstrom & Bjork, 2015), as polysemous words can be difficult to learn, but the process of learning results in a more robust lexical representation (Cervetti et al., 2015).

Previous research shows that words with many meanings are retrieved more efficiently (Azuma & Van Orden, 1997; Borowsky & Masson, 1996; Hino & Lupker, 1996; Eddington & Tokowicz, 2015). This may be because the highest-frequency words also have the most meanings (Ravin & Leacock, 2000), or because words with many meanings provide learners with more opportunity to compare and integrate usages across encounters (González-Fernández & Schmitt, 2019). The number of word meanings may even be more important than word frequency, as evidenced by the fact that it is a stronger predictor of vocabulary test scores than target word frequency in a variety of studies, including Papers II and III (Cervetti et al., 2015; but see Hiebert et al., 2019, which found no significant relation for multiple meanings).

However, research also indicates that words with many meanings are processed slower in lexical decision tasks (Beretta, Fiorentino, & Poeppel, 2005; Rodd, Gaskell, & Marslen-Wilson, 2002;2004) and semantic categorization tasks (Hino, Lupker, & Pexman, 2002). This divergence in research is due in part because multiple meanings can be related (i.e., polysemous, where related meaning senses share a core meaning; Srinivasan & Snedeker, 2011) or distinct (i.e., homonymous). Recent research has demonstrated that facilitative effects are often found when meanings are related (polysemous), while the opposite holds when meanings are distinct (ambiguous or homonymous; Eddington & Tokowicz, 2015; Floyd & Goldberg, 2020).

At the same time, whether a word has multiple meanings (related or not) is unimportant if a learner is unaware of them. For example, when studying students' production of polysemous words in an intensive English-language program, Crossley, Salsbury, and McNamara (2009) found that students used polysemous words in their first months but did not begin to extend the core meanings until later. Additionally, it can be challenging to learn new senses of conceptually rich words that are already partially known (González-Fernández & Schmitt, 2019; 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 their knowledge about

the word form and register how this meaning or sense is novel. Though the process may be challenging, it likely supports a rich representation of the word, especially when the word encounters are staggered (Soderstrom & Bjork, 2015; Crossley, Salsbury, McNamara, 2009).

Use. A word's use component comprises grammatical functions, collocations, and constraints on use. One common usage measure is word frequency, or how often a word occurs in a given text. Higher-frequency words are processed more quickly and accurately than lower-frequency words (coined the *word frequency effect*; Monsell, Doyle, & Haggard, 1989; see Brysbaert, Mander, & Keuleers, 2018 for a review). Part of this is because of statistical or incidental word learning—learning through encounters instead of explicit and direct instruction. Incidental word learning requires many encounters (Stahl & Fairbanks, 1986), and the likelihood of learning a word correlates with text exposure. Swanborn and de Groot's (1999) meta-analysis shows that 5th–11th graders have a 15% chance of learning a novel word from an incidental encounter.

Nevertheless, it is difficult to measure the exact number of times a specific individual has encountered a specific word without using nonwords or other controlled experiments; hence word frequency across a large collection of language data (a corpus) is commonly used as a proxy for average word experience. However, in their 2018 review, Brysbaert, Mander, and Keuleers emphasize that the word frequency effect is contingent on the corpus from which the frequency measure is drawn and that individual differences, particularly concerning a person's specific vocabulary size (breadth) and exposure (Preston, 1935; Cop et al., 2015) still exist.

Corpora can collect language data across nearly any context, such as subtitles (Subtlex; Brysbaert & New, 2009), blogs (Worldlex; Gimenes & New, 2016), child language (CHILDES; MacWhinney, 2000), scientific texts (SciTex; Degaetano-Ortlieb et al., 2013), and L2 learner texts (ICLE; Granger, 2003), to name a few. Furthermore, with more advanced technology available, corpora have exploded in size, such as the British National Corpus (BNC; Leech, 1992) with over 100 million words, and the Corpus of Contemporary American English (COCA; Davies, 2009) with over 450 million words, both of which contain spoken and written English (although more written than spoken).

Still, multiple encounters is not enough to develop a rich lexical representation; encounters must be diverse and high quality (Adelman, Brown, & Quesada, 2006; Webb & Nation, 2017; Nakata, 2011; Frances, Martin, & Duñabeitia, 2020). Previous research has found that the word frequency effect may not be due to the number of raw encounters but because high-frequency words occur in more contexts (Adelman, Brown, & Quesada, 2006; Frances, Martin, & Duñabeitia, 2020; Brysbaert & New, 2009). In their 2006 study, Adelman,

Brown, and Quesada found that the number of documents in which a word appeared (coined *contextual diversity*) eliminated the unique effect of word frequency when both were included in regression models to predict lexical reaction times. Moreover, Hoffman, Lambon Ralph, and Rogers (2013) estimated *semantic diversity* using Latent Semantic Analysis as a way to measure sentence-level diversity and also found that their semantic diversity measure was a significant predictor of reaction times, while frequency was not (recently replicated by Cevoli, Watkins, & Rastle, 2021). Recently researchers have found that words which appear in redundant contexts are recognized slower and less accurately than words used in diverse contexts (Jones, Johns, & Recchia, 2012; Johns, Dye, & Jones, 2016).

However, this does not mean that contextual diversity effects completely subsume all effects of word frequency. In their *Nature* report, Frances, Martin, and Duñabeitia (2020) explain that, in certain situations, word frequency and contextual diversity exhibit opposite effects. For example, word recall can be facilitated by word frequency (frequent words are recalled better) and hindered by contextual diversity (diverse words are recalled worse), showing the push and pull between the multiple-exposure and salience effect.

2.2.3 *Vocabulary Mastery*

Vocabulary strength (or mastery), suggests that individual word knowledge exists on a continuum, and is represented in Figure 3 as the horizontal stretching of each cone. Dale and O'Rourke (1986) suggested four stages of word knowledge, which emphasize partial knowledge more holistically: (1) completely unknown, (2) implicitly known, (3) partially known but mastered in some contexts, and (4) completely known in all contexts. Similarly, Stahl (2003) proposed three levels: (1) association processing, i.e., associating with familiar concepts; (2) comprehension processing, i.e., comprehending in particular contexts; and (3) generation processing, i.e., using the word in a new context (Brown, Frishkoff, & Eskenazi, 2005). More recently, Deane et al. (2014) hypothesized that word knowledge moves from familiarity to semantic representations, then conceptual representations, consolidation with world knowledge, and finally, encyclopedic understanding.

Conversely, mastery can be considered in terms of knowledge of the three components (and various subcomponents) suggested by Nation (2001). Schmitt (2014;2019) proposed that vocabulary knowledge not only exists across multiple components, but also that mastery of each component exists on its own continuum. Different components of word knowledge develop simultaneously (Gonzalez-Fernandez & Schmitt, 2020; Nation, 2022), such as learning

the written form and grammatical function in parallel, and are represented in Figure 3 by the different shaded areas of each component for individual words.

Additionally, all words are not learned in the same way. For example, learning first words aurally before learning to read, learning words incidentally through spoken and written exposure, and deliberate vocabulary learning vary drastically, as evidenced by the diverse vocabulary interventions for different learners (see Marulis & Newman, 2010 for a meta-analysis of young children’s vocabulary interventions). It is, therefore, critical to consider individual differences in learning and textual differences.

2.2.4 Vocabulary Integration

Vocabulary knowledge “is not a feature of individual words: rather it is a characteristic of the test taker’s entire vocabulary” (Meara & Wolter, 2004; p. 87). Perfetti and Hart’s (2002) *Lexical Quality Hypothesis* states that words with identical orthography or phonology are directly linked in the mental lexicon. Moreover, words with similar orthography and phonology (“neighbors”) are simultaneously activated in lexical decision tasks (e.g., Andrews & Hersch, 2010; Meade et al., 2018; Forster et al., 1987; Andrews, 1996; Luce & Pisoni, 1998), indicating that lexical representations are interconnected.

Neuroimaging studies using priming tasks also indicate that words with semantic relationships and similar meanings are also simultaneously and implicitly activated (e.g., Copland et al., 2003; Giesbrecht, Camblin, & Swaab, 2004; Kotz et al., 2002; Matsumoto et al., 2005; Rissman, Eliassen, & Blumstein, 2003; Rossell, Price, & Nobre, 2003; Wible et al., 2006). Words may also be integrated with a hierarchical structure, wherein more specific concepts like “poodle” inherit semantic information from more general super-ordinates like “dog.” Collins and Quillan (1969) proposed this hypothesis when they found that participants spent longer confirming that “canaries have feathers” as opposed to “birds have feathers” (Fellbaum, 2010). Overall, it is likely that words are integrated in many ways simultaneously, which may explain why no uniform model exists for how the mental lexicon is organized (Dóczy, 2019).

2.2.5 Factors Affecting Vocabulary Learning

Lastly, Nation’s (2022) framework includes factors that influence vocabulary learning, represented as the four boxes on the outside of Figure 3: content knowledge, quality of input and output, language knowledge, and within-word factors. Content knowledge refers to the within-person experience and knowledge of the world; similar to the general background

knowledge component in the Reading Systems Framework (Perfetti & Stafura, 2014). One point of departure is that, in Nation's (2022) framework, general knowledge affects opportunities to learn vocabulary (Nation & Coxhead, 2021).

Similar to the Reading Systems Framework (Perfetti & Stafura, 2014), the within-person linguistic knowledge impacts vocabulary learning (Nation, 2022). The point of departure here is that Nation (2022) somewhat clarifies that linguistic knowledge from multiple languages impacts vocabulary learning. Current theories about how lexical representations in multiple languages are stored suggest that words exist in a shared system and are activated when sharing orthographic, phonological, or semantic similarities, regardless of language (Dijkstra et al., 2019; Meade et al., 2018).

However, lexical connections between languages also vary over time. In the case of sequential bilinguals (who learn L2 after initial development in L1), it is common to learn L2 words initially by associating them with L1 translations (Barcroft, 2009; O'Malley & Chamot, 1990). This is why priming studies have found that the link between direct translations—particularly novel L2 target word to L1 translation—is strong (Ferré et al., 2017; Wen & van Heuven, 2017). Yet, as proficiency in L2 increases, linking between L1 words and their L2 translations strengthen (Duñabeitia et al., 2010; Nakayama et al., 2018).

Lastly, while features of individual words can impact vocabulary learning by increasing or decreasing the learning burden, multilingual learners can leverage different advantages based on their cross-linguistic knowledge. For example, learning the meaning of a cognate is easier than learning a unique word in L2 (Jacobs, Fricke, & Kroll, 2016; De Groot & van Hell, 2005). However, which words are considered cognates depends on the learners' linguistic knowledge; the fact that “animal” is a Spanish–English cognate is not facilitative to a Norwegian–English bilingual (where “dyr,” also meaning animal, is not a cognate), nor to a Spanish–English bilingual who doesn't know “animal” in either language. Within-word and within-person factors are interconnected, and thus considering aspects of both simultaneously when modelling vocabulary knowledge and reading comprehension is critical, even though complex.

2.3 Vocabulary Assessment

Although not explicit in Nation's (2022) model of vocabulary learning, diverse aspects of word knowledge imply the need for diverse vocabulary assessment, as poignantly stated by Crosson, McKeown, and Ward (2019): “If students are simply asked to practice definitions, their breadth of vocabulary increases, *as measured by the ability to recognize definitions...* To

understand the success of instructional interventions we need to be able to characterize the kinds of knowledge that students have acquired.” (p. 197; my emphasis).

Vocabulary assessment can be approached in a variety of ways, such as contrasting receptive versus productive skills (Schmitt, 2014; Webb, 2008), assessing the form-meaning link versus lexical integration (Read & Dang, 2022), or assessing form, meaning, and use (Nation, 2001).

In educational research, it is common to assess receptive knowledge of the form-meaning link via diverse multiple-choice assessments. For example, Deane et al. (2014) hypothesized that, under “normal word learning conditions,” word knowledge would move from familiarity to semantic representations, then conceptual representations, consolidations with world knowledge, and finally, encyclopedic understanding. To test their hypotheses, word knowledge was assessed via four novel assessments: a collocation task, topical associate task, hypernym task, and definition task. Lawrence et al. (2019) found that scores on the collocation, topical associate, and definition tasks each explained unique variance above a synonym identification task, indicating that the tasks tapped into different aspects of knowledge; however, the research design limited analytical ability to compare the novel tasks to one another statistically.

Another way to assess different dimensions is to construct specific foils that differentiate various aspects of word knowledge. In their EL-RAVE intervention study, Crosson et al. (2019) assessed ELLs on a cloze task for academic vocabulary. Response options included: the key (correct answer), a syntactic foil, a topic foil, and a semantic foil. The syntactic foil would be in the wrong part of speech (e.g., *confine* in “We saw a __ on the busy highway”). The topic foil would be the correct part of speech but not related to the topic (e.g., *confine* in “The dog did not __ the food”). The semantic foil would be the correct part of speech and topically related but not make sense (e.g., *confine* in “Prisoners often __ letters to their families from jail”). While they do not report the results of a distractor analysis (i.e., information about how often specific foils were selected), students who selected different foils received different scores, implying that different responses served as a proxy to varying levels of knowledge.

This is, by no means, an exhaustive list of all vocabulary assessments; however it highlights that various aspects of and approaches to vocabulary knowledge necessitate diverse vocabulary assessments.

3

Methodological Considerations

The three empirical papers in this dissertation aim to elucidate what makes vocabulary challenging for diverse learners across diverse tasks at the word level. Unpacking complex relationships requires explicit consideration of how to approach research, define terms, and interpret results. This chapter opens with a section on overarching issues that extend across papers, then delves into a selection of issues pertaining to specific papers. I conclude with a brief section on data handling and scientific transparency.

3.1 Overarching Issues

A few methodological considerations span papers and are worthy of overall consideration. In particular, I focus on how a critical unit of language was defined, how vocabulary and reading comprehension were assessed, and why explanatory Item Response Theory was selected as the analytical approach.

3.1.1 Defining a Critical Unit of Language

When discussing “vocabulary,” we often think of individual words, yet, defining what counts as an individual “word” is not universal. Different definitions, even if all logically-consistent, imply specific research questions and change what conclusions can be drawn from any particular study (Schmitt, 2014). From a methodological view, I considered three common approaches to determining the critical unit of language: the lexeme (a single meaning unit, regardless of size), the lemma (a base word and its inflections), and the letter string (unique combinations of letters). The figure below illustrates the differentiations between each approach.

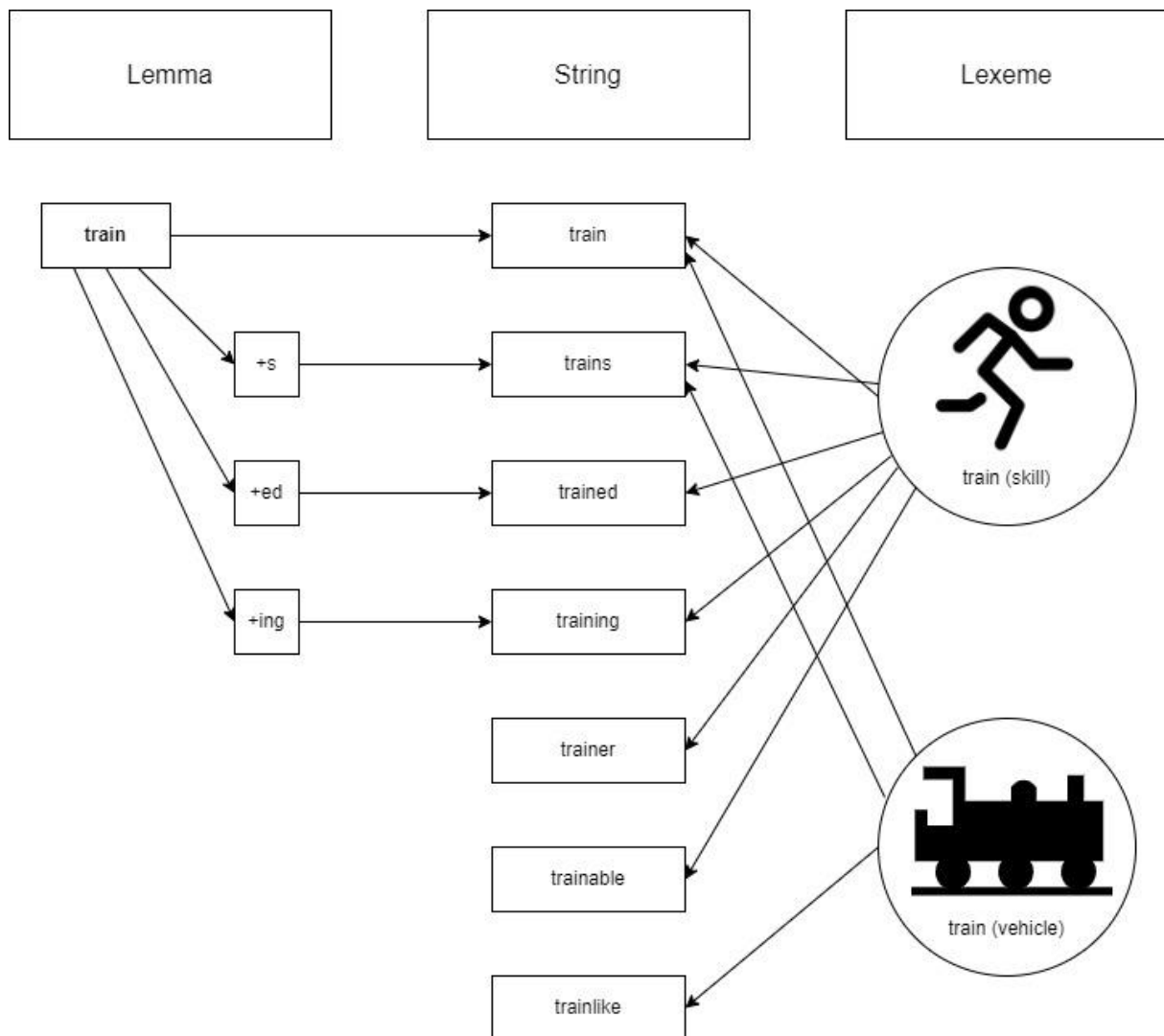


Figure 4. Example of Different Units of Language.

Note: Stock images from freepik.com

A lexeme is a unit of language with a single meaning, regardless of orthographic form. Thus, using lexemes as the critical unit distinguishes between homonyms (e.g., a locomotive “train” and to “train” for a marathon are distinct meanings, as seen in Figure 4). At the same time, combining senses and drawing boundaries between distinct meanings is not necessarily a binary process. For example, “paper” can refer to both sheets of writing and the content written on the paper (such as the papers in this dissertation, if you will). While these are often regarded as related senses, they share few related semantic features (wood pulp versus intellectual content), which hinders facilitative processing when wrapping paper and empirical paper are paired together (Klein & Murphy, 2001; Klepousniotou, Titone, & Romero, 2008).

Additionally, multiword expressions such as “give up” and “take off” are also distinguished as individual lexemes. Multiword expressions are common in English, and comprehending them goes beyond understanding the individual words (Hinkel, 2022; Webb & Nation, 2017). From a methodological standpoint, and similar to homophones, distinguishing between multiword expressions that share similar meanings can be challenging. For example, “to each his own” and “to each their own” both convey the idea that people have different preferences, but they are not identical—so, how close is too close for distinct entries? Drawing boundaries between lexemes quickly becomes cumbersome when decisions must be made across thousands of words. Moreover, separating homonyms and parts of speech ignores lexical connections between similar orthographies, and combining multiword units ignores individual word processing.

Grouping words by lemmas includes a common stem (or root word) and inflections from the same part of speech (Gardner & Davies, 2014). For example, the root word “train” can be split into two lemmas: the verb “to train” with the inflections “trains”, “trained”, and “training” (illustrated in Figure 4); and the noun “train”, with the inflection “trains”. Research in natural learning has indicated that known vocabulary can be leveraged when learning novel words, such as inflections (Biemiller & Slonim, 2001). For example, when a learner already knows “train” and “-ed”, the learning burden for “trained” is lower.

At the same time, restricting lemmas to root words and their inflections, which must be identical parts of speech, ignores research that learning derivations across parts of speech is easier as well (e.g., “talker” and “talkative”). Priming and partial activation studies provide evidence that words belonging to the same word family (i.e., base words plus their inflections *and* transparent derivations; Coxhead, 2000; Schmitt & Zimmerman, 2002) are likely grouped in the mental lexicon (Nagy et al., 1989). Consequently, many vocabulary word lists—particularly those with a pedagogical purpose—are also grouped by lemmas or word families

(West, 1953; Coxhead, 2000). However, doing so ignores that some words in the same family do not share the same core meaning (e.g., “react” versus “reactivate”; Gardner & Davies, 2014; Nagy & Townsend, 2012; Hyland & Tse, 2007) and that different words in the same family vary in other ways, such as collocational use and decodability.

A letter string is a group of letters that form a word. In this case, “train” and “trains” are distinct units, but “train” and “train” are not, as illustrated in Figure 4. There is evidence that words that share orthography (i.e. homonyms) and words with many senses (i.e. polysemes) are all initially activated when encountered (Klepousniotou & Baum, 2007); so when a reader sees “train”, both the verb and noun are activated even if the context suggests only one is correct. However, distinguishing units by letter strings ignore multiword units where new meanings come to fruition when specific words are joined together (“the whole is greater than the sum of its parts”, so to speak). Still, distinguishing by letter string is relatively clear-cut and requires significantly less data processing compared to distinguishing words in relation to distinct meanings.

Different conceptualizations of the critical unit of language are useful depending on the particular research questions and approaches in any given study. From a methodological standpoint, distinguishing words by lexeme or lemma in Paper I would be challenging because many lexical measures, such as word frequency, do not distinguish between lexemes or lemmas. In addition, when unique letter strings define the critical unit of language, then the number of meanings a word possesses becomes a lexical measure itself—instead of the distinguishing feature between words. Consequently, we chose to distinguish by unique letter string in Paper I; which naturally leads to a similar conclusion in Papers II and III, given that the latter use lexical dimension estimates derived from the former.

3.1.2 Selecting Vocabulary Assessment(s)

Because there are many components to vocabulary knowledge, there are also diverse ways to assess vocabulary knowledge; yet we often take for granted that assessments are not equal and ignore the implications of selecting specific assessments. Even when evaluating the knowledge of a single component, such as the form-meaning link, assessment can still vary in terms of receptive versus productive knowledge (Nation, 2001), partial versus complete knowledge (Dale & O’Rourke, 1986; Deane et al., 2014), and different ways of defining “meaning” itself (Eddington & Tokowicz, 2015). We need research across all varieties of assessment to understand the complexity of vocabulary knowledge and learning, but students

are rarely assessed in such an in-depth way (Crosson, McKeown, & Ward, 2019; Elleman et al., 2009).

In Papers II and III, students were assessed on their receptive knowledge via multiple-choice assessment(s) partly because administering and scoring multiple-choice tests is relatively easy and requires less resources than other methods, such as open responses. Moreover, the data analyzed in Papers II and III are pre-test data from previous intervention work, where target words were selected specifically *because* the researchers expected participants would not know the words well (Snow, Lawrence, & White, 2009). Hence, assessing receptive knowledge via identification tasks instead of productive knowledge via recall tasks would alleviate potential floor effects.

Additionally, the multiple-choice items were scored on a binary scale—either correct or incorrect (versus partial credit, for example in Crosson et al.’s 2019 study). Binary scoring can be advantageous because it is easy to score and has high internal reliability (Haladyna, Downing, & Rodriguez, 2002; Haladyna, 1994; Ben-Simon, Budescu, & Nevo, 1997), but it also assumes that knowledge is all-or-nothing because there is no possibility to receive partial credit when demonstrating partial knowledge (Lau et al., 2011; Kurz, 1999). Given that vocabulary knowledge exists on a continuum, and selecting the correct answer can depend on which aspect of word knowledge is assessed, a binary scoring system can be problematic. Take for example a case where a student knows the locomotive meaning of the word “train” but not the verb “to train”. If an item with the target word “train” only assesses the verb meaning, then it is not unlikely that the student will incorrectly answer the item. With a binary scoring system, the underlying assumption of an incorrect response is that the student does not know “train” *at all*, despite evidence to the contrary.

Furthermore, it is possible to answer a multiple-choice item correctly by sheer coincidence (Jaradat & Tollefson, 1988; Lau et al., 2011), yet multiple-choice tests often assume that correct answers are due to genuine knowledge, not luck. Statistical models can correct for the probability of a “lucky guess”; for example, estimating the lower asymptote in a three-parameter Item Response Theory Model (i.e., the *guessing parameter*; Lord, 1980, de Ayala, 2009; Embretson & Reise, 2013). However, doing so invokes different issues. Aside from the fact that adding more parameters to estimate decreases statistical power (Wainer, 1983), including a guessing parameter makes assumptions on when correct answers are due to luck and not knowledge (e.g., based on the pattern of student responses and their “risk-taking” tendencies; de Ayala, 2009, p. 126); this penalizes students who genuinely know the answer

when it is statistically *unlikely*, which is especially true for lower performers (Stemler & Naples, 2021).

Additionally, Paper II explores student responses on the synonym task as opposed to definitions, hypernyms, or antonyms. While there is a hyperfocus on synonym tasks in the educational field, this does not negate the usefulness of the task itself. The fact that much research involving vocabulary assessment via synonym identification is a potential strength of this study, as it allows for results of sophisticated statistical modelling to be contextualized within a well-established research field. Instead of presenting the research field with a newer analytical design and a less-common assessment type, we opted to begin with the former, assuming that future research will build upon these results and expand to diverse assessments.

Paper III compares performance on the synonym tasks to a similar definition task. In the initial data collection process, all students were given the same 50 synonym items plus one of 16 experimental forms (the four assessment types from Deane et al., 2014 x four different sets of 12 target words). We chose to include one of the experimental assessments—the definition task—as a comparison to the synonym task and to assess the range of vocabulary knowledge. Comparing two tasks adds another dimension of complexity to the analytical model, yet this is a step in understanding vocabulary knowledge when words are multi-componential, as described previously. Moreover, because the initial research design limited students to one of 16 experimental forms, each experimental item has approximately $1/16^{\text{th}}$ of the sample size of synonym identification tasks; modelling one experimental assessment alone with less data will result in increased standard errors during estimation and thus increase the likelihood of a Type II error (Field, 2013). Including the synonym task as a comparison lends these estimations some precision because of the inclusion of more data points per word (at least 4x as much) and more data points per person (usually 62 data points instead of 12). Inclusion also allows for interesting comparisons of student responses to identical words across tasks.

Still, we could choose from four experimental assessments: a multiword expression, topical associate, hypernym, or definition task (Deane et al., 2014). While we initially considered modelling across all assessments, it was immediately apparent that doing so made interpretation significantly more complex. More importantly, the baseline probability of a random correct guess was not identical between item types because the number of response options differed. While including statistical control by estimating a guessing parameter could be problematic as discussed previously, completely ignoring the differences in guessing probability is also problematic. Hence, since the synonym and definition tasks are often compared as different aspects of vocabulary depth (Ouellette, 2006; Deane et al., 2014), and

because they both included the same number of response options, the focus of Paper III was limited to these two item types.

3.1.3 Measuring Reading Comprehension

Reading comprehension is a complex process, and thus, different assessments can be valid and reliable, yet also tap into different aspects of reading comprehension (Keenea, Betjemann, & Olson, 2008; Francis et al., 2005; Nation & Snowling, 1997; Cutting & Scarborough, 2006). For example, Nation and Snowling's (1997) covariance analysis, and later Francis et al.'s (2005) latent trait models, indicated that decoding was more strongly related to sentence completion assessments (i.e., cloze tasks) than to multiple-choice assessments; and that decoding skill was more strongly related to cloze task performance than listening comprehension skill. Moreover, when comparing students across four different assessments, Keenan, Betjemann, and Olson (2008) found that the oral reading assessments with moderate text lengths saw stronger relationships with listening comprehension, while silent reading assessments with short passages or single sentences were more strongly related with decoding skills. Later, Keenan and Meenan (2014) found that identifying students with comprehension deficits was contingent upon which task was used, as comprehension difficulty diagnoses overlapped, on average, by only 43%.

Papers II and III use the Gates-MacGinitie reading comprehension subtest (GMRT; MacGinitie et al., 2000), a widely-used, nationally normed, and internally reliable test (Maria et al., 2007). This subtest requires test takers to read short passages (three to five sentences) and answer three to six multiple-choice questions before moving to the next set. Cutting and Scarborough (2006) found that decoding skill and oral language proficiency explained 46% of the variance in GMRT scores together, plus an additional 6% uniquely explained by decoding skills and 15% uniquely explained by oral language proficiency.

Notably, GMRT scores in Papers II and III were also standardized (z-scored) in relation to the relevant samples prior to statistical modelling as opposed to standardizing using published national norms. Transforming scores in any way changes the interpretation of estimated slopes in regression analyses (see Pek, Wong, & Wong, 2017 for a mathematical overview). When models have multiple predictors, any specific main effect is interpreted with the caveat “when all other predictors are zero” (Field, 2013; Willett, Singer, & Martin, 1998; see also Preacher, Curran, & Bauer, 2006). In the case of z-scores, zero indicates “average”, and thus is distinctly meaningful. This is worthy of explicit mention because reading comprehension scores were z-scored against *monolinguals* in Paper II and *everyone* in Paper

III. Thus, interpreting the main effect of target-word frequency in both papers would read as “the effect of frequency for the average monolingual reader” or “the average middle-school student,” respectively.

3.1.4 Explanatory Item Response Theory

Papers II and III employ one-parameter, doubly-explanatory Item Response Theory models (eIRT; De Boeck & Wilson, 2016) to explain student responses to vocabulary items both as a function of student characteristics, item features, building off of the general mathematical formulation (Wilson & De Boeck, 2004; p. 66; equation 2.1):

$$\eta_{pi} = \sum_{j=1}^J \theta_j Z_{pj} + \varepsilon_p + \sum_{k=1}^K \beta_k X_{ik} + \varepsilon_i$$

Where η_{pi} is a linear component for each person p and item i pair.

In the first large operator, j represents the person predictors, such as grade level or reading comprehension score. θ_j represents the regression weight of that predictor on student ability θ , and Z_{pj} represents student p 's score on predictor j (e.g. “6” if a student is in grade 6).

In the second large operator, k represents the item predictors, specifically target word lexical dimension scores and item type in Paper III. Parallel to θ_j , β_k represents the regression weight of a specific predictor on item difficulty β , and X_{ik} represents item i 's score on predictor k (e.g. “0” for a target word with average Frequency).

Both large operators also include an estimate for residual variance, ε_p for person and ε_i for item, which are assumed to be normally distributed with a mean of 0 (see Kulesz et al., 2016 for more mathematical information and statistical code). This is worth explicitly noting because models in Papers II and III aim to reduce *both* the person and item variance by using different person and item parameters to explain variance in student responses. As Kulesz and colleagues (2016) state, “This approach provides tremendous flexibility in investigating complex reader-test interactions by simultaneously examining the influence of reader, passage, and item characteristics on readers’ comprehension as measured by their performance on test questions” (p. 1085).

Paper II uses measured student characteristics to estimate student ability θ , latent factor estimates from Paper I to estimate item difficulty β , and cross-classified interactions between student reading comprehension scores and estimated latent factor estimates to simultaneously

predict student ability and item difficulty across 50 synonym identification items for the monolingual student subsample. Figure 5 illustrates the final eIRT model used in Paper II.

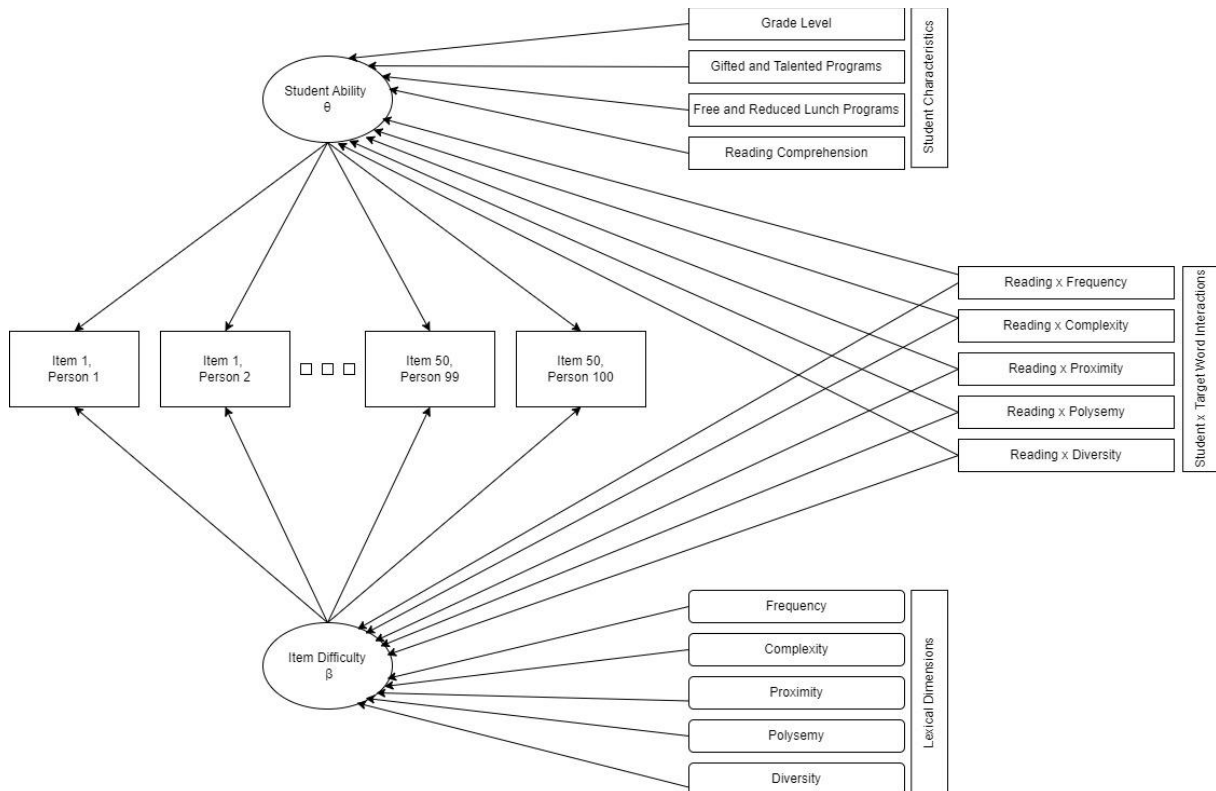


Figure 5. The Explanatory Item Response Theory Model for Paper II.

Paper III uses measured student characteristics to estimate student ability θ , latent factor estimates from Paper I, and item type, and interactions between them to estimate item difficulty β . Additionally, Paper III uses cross-classified interactions between ELL classification and estimated latent factor scores and item type to simultaneously predict student ability and item difficulty across 50 synonym identification items and 48 definition identification items for the entire sample of students. Figure 6 illustrates the final eIRT model used in Paper III.

Alternatively, Papers II and III could have analyzed data using a multiple regression approach, where the same student characteristics, lexical features, and interactions predict item responses. However, multiple regression with “correctness” as a single dependent variable cannot disentangle differences between items from differences between people (Khorramdel et al., 2020; de Boeck & Wilson, 2016); only relationships with the resulting behavior (i.e., correctly or incorrectly answering an item). Item Response Theory, instead, takes a multivariate regression approach by modelling multiple dependent variables (item difficulty and person ability in the case of a one-parameter model; Cai & Thissen, 2014).

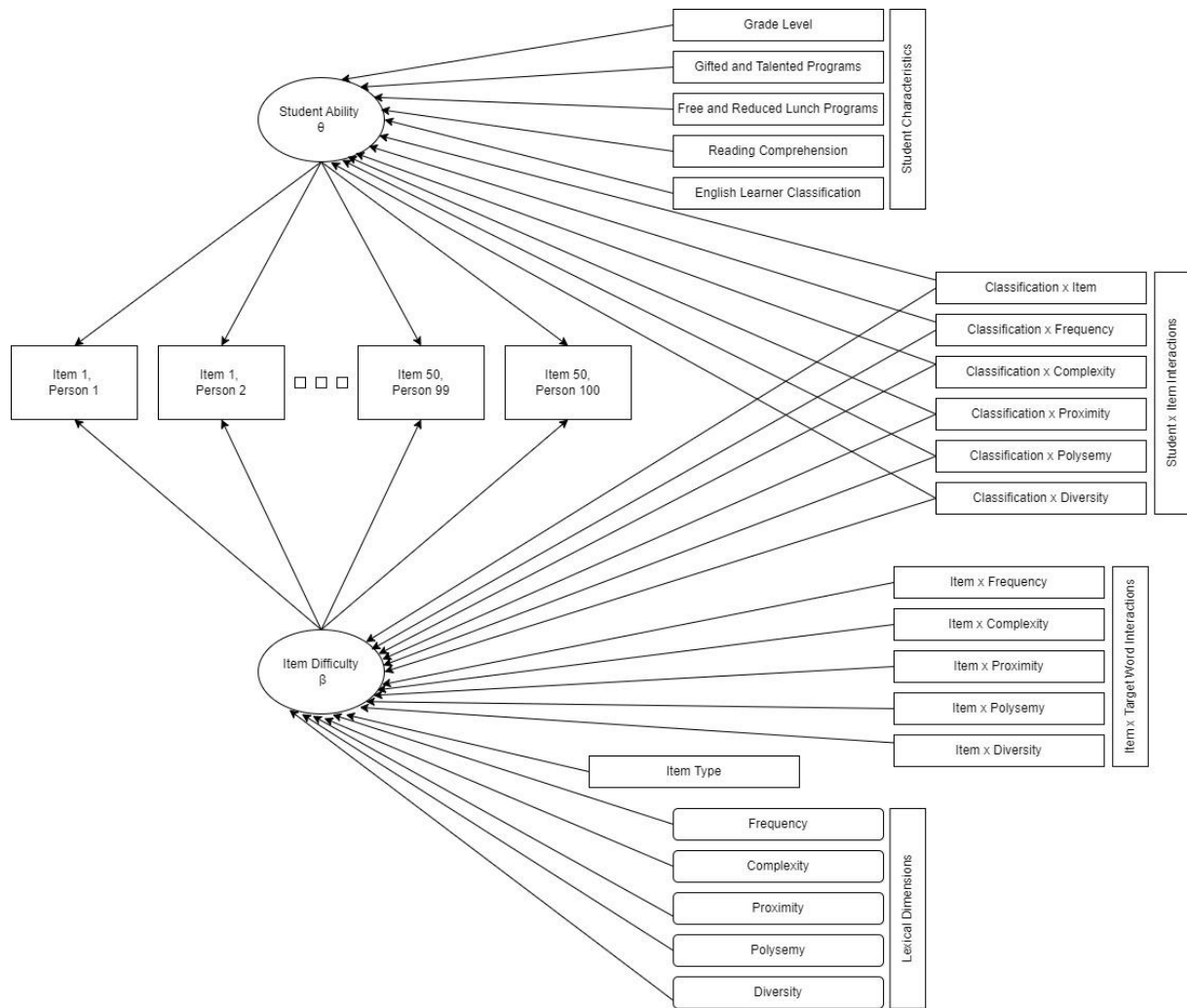


Figure 6. The Explanatory Item Response Theory Model for Paper III.

3.2 Study-specific Issues

3.2.1 Exploratory versus Confirmatory Approaches in Latent Estimation

Paper I aimed to explore the relationships between lexical characteristics, which can be answered via factor analysis. Factor models use the correlation matrix between many individual measures to identify likely latent factors (Field, 2013). Because we did not want to use theory to determine how individual measures would group into factors, it was clear that at least one model would need to be based on an exploratory factor analysis (EFA). However, determining what to do with subsequent samples was less clear: continue with EFA, or take a confirmatory approach and use the initial model as theory for preceding confirmatory factor analysis (CFA) models.

The EFA and CFA approaches both have important implications for the overall conclusions. EFA is a model-building, inductive approach that works from word-level data to

theory development, while CFA is a model-testing, deductive approach that works from a pre-existing theory to word-level data (Tukey, 1980; Schmitt, 2011; Field, 2013). Using a CFA approach allows for direct statistical comparisons of model fit between a pre-specified model and new observed data; but given the exploratory nature of our hypotheses, it would likely be inappropriate (Schmitt, 2011; Marsh et al., 2011). Moreover, the CFA approach would have required a model to be specified, and it was unclear *which* sample of words would be the appropriate original reference. Instead, multiple EFAs allow different factor structures to emerge if said structures are a better fit for the different samples (Schmitt, 2011). For example, the words from the General Service List could have been estimated against a four-factor model like the domain-specific words, but doing so would miss that the *better* model for these words was five factors, not four.

3.2.2 Allowance of Missing Data in Factor Analyses

An important assumption for factor analyses is complete data (McNeish, 2017): no observations can be missing from the dataset used to model. Given the exploratory procedure for collecting all possible data regardless of quality, it was clear that many observations would be missing. In some cases, missing data indicated nonwords with data on a single measure (e.g., “-Feb”). However, many letter strings with missing data do have dictionary entries (e.g., “boardroom,” missing 14%; and “boardrooms,” missing 75%).

There are two options for handling missing data in factor analyses: omit any letter string with any missing data or replace missing data with estimates. Replacing missing data has been available to researchers for decades (see Dempster et al., 1977; Heckman, 1979; Rubin, 1976) and can involve, for example, multiple imputation with chained equations (also called fully conditional specification imputation or sequential regression; van Buuren, 2007; White, Royston, & Wood, 2011), joint imputation (van Buuren et al., 2006), or even simply assigning missing values to the mean (though this is no longer recommended given the more advanced techniques and technology available; Graham, 2002; Allison, 2002).

Eliminating all missing data results in a smaller dataset (and, consequentially, less statistical power), which can be less representative of the whole population (Heckman, 1979). Initially, we noticed that lower frequency words were more likely to have some missing data, often being omitted from one or two specific measures. Excluding observations with any missing data would consequentially exclude many low-frequency but still genuine words, biasing the results towards higher-frequency words. Conversely, replacing any missing data means estimating possible values, which introduces error into the model. Put bluntly,

“imputation is making up the data” (Graham, 2002; p. 559), though the assumption in using sophisticated statistical estimation techniques is that the made-up data is reasonably accurate. Still, replacing missing data usually requires using the values already present in the dataset to make estimations, such as calculating the mean score based on available data and replacing the missing observation with the mean. Data replacement processes bias the relationships between observations and other variables; for example, replacing missing data with the mean artificially reduces the standard deviation, and multiple imputation relies on (and therefore inflates) correlations (Graham, 2009). Given that the explicit goal of Paper I was to explore the correlations between variables, avoiding imputation methods that artificially bias correlations and instead limiting the word sample in models was more appropriate.

3.2.3 Assessing the Monolingual Subsample

After estimating latent factor scores, Paper II used estimates from the general academic vocabulary model (AWL-reference) in Paper I to predict item difficulty for monolingual English speakers on a synonym task. While limiting the sample to monolingual students might appear linguistically ethnocentric, the exceptionally diverse sample of multilingual learners is analytically challenging. Modelling vocabulary knowledge is more complex for multilingual learners in part because there are more ways in which they can vary. All students can vary across demographic characteristics, such as age and socioeconomic status, and in proficiency in a specific target language. However, multilingual students carry additional key characteristics, particularly the age at which they began learning the target language, their exposure to the target language, similarity between the target language and their native language, and proficiency in their native language (Paradis, 2011). Mathematically, such a diverse sample can lead to unstable model estimates and exceptionally complex models that are more likely to be statistically underpowered (Dattalo, 2008; Cohen, 1988). This is why it is not uncommon to, for example, only sample ELLs with the same L1 or other similar background characteristics.

Similarly, we balanced statistical power, precision, and theoretical practicality by artificially simplifying the research questions and models, focusing the research questions to a more homogenous sample, thus limiting variability (Dattalo, 2008). Doing so does not negate the importance of the multilingual perspective, but sheds light on complex research questions as is and provides insights that allow for the framing of future research questions based on these initial results. Consequently, Paper II does not attempt to generalize beyond monolinguals, much like it does not generalize beyond academic vocabulary or synonym knowledge. Instead,

it takes a deeper dive into a partition that is—in its own right—interesting, even if not the complete story of vocabulary learning.

3.2.4 *Learner Classification in the Multilingual Sample*

Building upon the conclusions from Paper II, Paper III investigates what makes academic words challenging across monolingual students and multilingual students at different proficiency levels on different vocabulary assessments. Taking a multilingual perspective entailed many new variables for consideration in modelling. In particular, we needed to decide whether to include measures related to language backgrounds (e.g., L1) and/or target language (English) proficiency.

Ultimately, analyzing L1-related data was not a methodologically sound choice for Paper III. Students reported over 30 different L1s, with many students being the sole speaker of that language within our sample (e.g., Pashto and Cebuano), which would result in unstable estimates (if any at all) due to cell sizes of $n = 1$ (McNeish & Stapleton, 2016). We could address this issue by combining languages into larger groups or language families, but this artificial categorization ignores meaningful variation (Cohen, 1983; Peters & Van Voorhis, 1940; Humphreys & Fleishman, 1974; DeCoster, Gallucci, & Iselin, 2011). Conversely, we could compare a few dominating L1s, such as English, Spanish, and Cantonese; although we lacked data in L1 proficiency.

One particularly interesting contrast highlighted in the limitations section of Paper II is the cross-classified interactions between target-word lexical characteristics and the *exposure* to English that students receive (as opposed to their reading comprehension proficiency alone). In California (where the current sample was located), all students receive an ELL classification, which determines entitlement to language support in educational settings and therefore indicates the qualitatively different English environments where diverse students find themselves.

The ELL classification system automatically classifies all monolingual students as English Only (EO). An important caveat here is that monolingual students vary widely in their English proficiency (as seen in Paper II) and their exposure to English. For example, previous research shows that children with a low socioeconomic status hear fewer words overall and less diverse words (Hart & Risley, 1995; Hoff, 2006; Rowe, Pan, & Ayoub, 2005). Consequently, Paper III includes both a main effect for reading comprehension scores and an indicator of low socioeconomic status via participation in the free and reduced lunch program, but shifts the focus for the interaction terms.

Students who enter the California education system speaking another language at home can be given one of two classifications: Initially Fluent English Proficient (IFEP) or Limited English Proficient (LEP). As their English skills improve, students initially classified as LEP can be reclassified (Reclassified Fluent English Proficient; RFEP) if they later meet the threshold for English proficiency. California Education Code EDC § 313 (f) outlines four criteria students must meet to receive the IFEP or RFEP classifications:

1. Pass a language proficiency assessment
2. Receive teacher evaluation of curriculum mastery and classroom readiness
3. Receive a similar evaluation from parents
4. Demonstrate comparability to their English-proficient peers on testing

ELL classification requires students to not only reach some level of English proficiency but also be ready for the standard classroom *and* have mastered the curriculum comparable to their peers. Students must be identified as ready and proficient by three separate entities: teachers, parents, and standardized tests. This is a high bar compared to the EO students, who can be at any level of English proficiency as long as English is their only language. As a result, IFEP and RFEP students often outperform EO students (Hill, Weston, & Hayes, 2014; Saunders & Marcelletti, 2013; Hill, 2012; Gándara & Remberger, 2006).

3.3 Ethical Dilemmas

Because the data analyzed across the three empirical papers was initially collected by other research entities (i.e. lexical data from diverse corpora in Paper I and anonymized data from the Word Generation intervention study in Papers II and III; Lawrence, Capotosto, Branum-Martin, White, & Snow, 2012), the major ethical dilemma in throughout this dissertation related to the tension between principles of scientific transparency and principles of data confidentiality.

Scientific transparency is a hallmark of research and is even expressed in the Norwegian constitution as a right of the people in Article 100: “The authorities of the state shall create conditions that facilitate open and enlightened public discourse” (Forskningsetikk, 2019). At the same time, researchers operating within the EU and participating Schengen countries, such as Norway, must adhere to explicit rules regarding data protection set by the General Data Protection Regulation (GDPR; European Parliament and Council of European Union, 2016). The result is two conflicting principles: at one end of the spectrum, researchers can restrict all access to any data as a way to ensure privacy, compromising transparency; at the other end, researchers can make all data open access, violating confidentiality. It is uncommon for

researchers to fall at either extreme; thus, finding the balance between openness and privacy is an ethical dilemma many of us face.

The current project analyzes different types of data, including data from multiple corpora, participant data, and open access data. As a result, there is no public access to full datasets for any specific study. However, the syntax and output are readily available in the supplementary materials for inspection, and relevant materials were made available during the peer review process when possible. Similarly, estimates on the five latent lexical dimensions from Paper I are freely and openly available for all noncommercial use at <https://academicvocab.times.uh.edu/>.

4

Main Features of the Papers

In light of the previous research and methodological choices made this section summarizes findings from each paper and presents a short commentary regarding the findings.

4.1 Paper 1: Latent Lexical Dimensions

Paper I explored the latent relationships between 22 nonbehavioral measures using multiple exploratory factor analyses across three word lists: the General Service List (GSL; West, 1953), the Academic Word List (AWL; Coxhead, 2000), and the domain-specific subset of the Academic Vocabulary List (AVL-DS; Gardner & Davies, 2014). We found five relatively consistent and distinct latent factors: Frequency, Complexity, Proximity, Polysemy, and Diversity. Figure 7 illustrates the results of the factor analysis for words on the GSL.

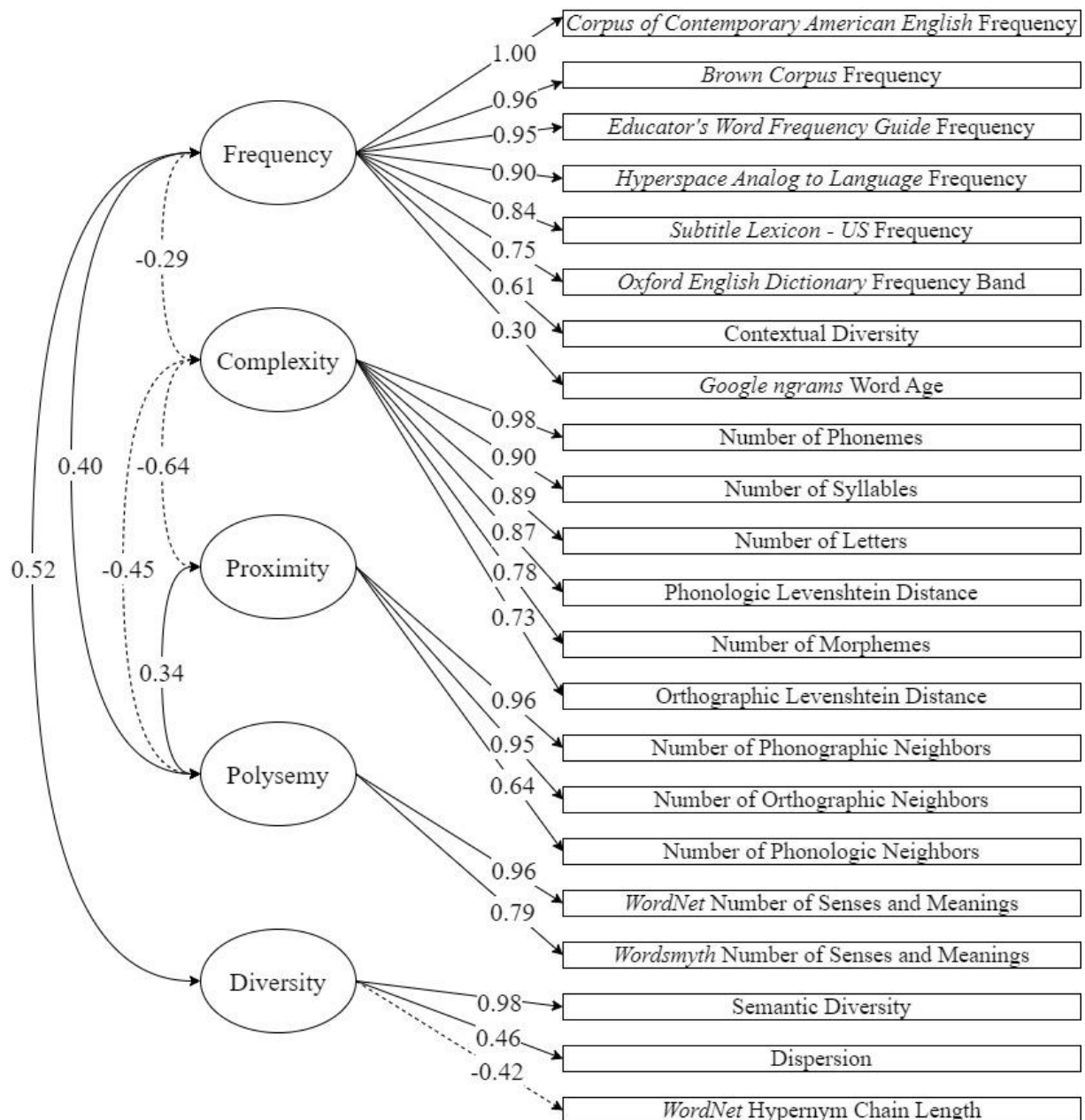


Figure 7. Results of the EFA for GSL words

The Frequency factor contained not only all of the measures of word frequency across a variety of corpora, such as the Corpus of Contemporary American English (COCA; Davies, 2009) and the Subtitle Lexicon (Subtlex-US; Brysbaert & New, 2009), but also measures such as Adelman, Brown, and Quesada's (2006) contextual diversity and Zeno's (1995) dispersion. However, both measures are a form of frequency at a larger grain size (number of documents versus the number of individual observations, for example). The Frequency factor explained about 24-26% of the variance in lexical features.

The Complexity factor contained measures related to word length, such as the number of letters and syllables. Additionally, the Complexity factor contained the orthographic, phonologic, and phonographic Levenstein distances. The Complexity factor explained 21-22% of the variance in lexical features.

The Proximity factor contained three measures of neighborhood density: orthographic, phonologic, and phonographic neighbors. The Proximity factor explained 11-12% of the variance in lexical features; a moderate decrease from Frequency and Complexity, though not negligible.

The Polysemy factor included measures for the number of senses and meanings across scraped data from WordNet (Fellbaum, 2010) and Wordsmyth (Parks, Ray, & Bland, 1998). Notably, we were unable to disentangle distinct meanings from related senses, particularly when unique letter strings—not part of speech—were used in the factor analysis, as discussed in the preceding Methods section. The Polysemy factor explained 7-8% of the variance in lexical features.

Finally, the Diversity factor included measures of semantic diversity and lexical precision. Semantic diversity measures diversity across sentences (Hoffman, Lambon Ralph, & Rogers, 2013), while lexical precision measures the length of the hypernym chain, for example a “poodle” is a “dog” which is an “animal” creates a hypernym chain of three for “poodle” (Fellbaum, 2010). The Diversity factor explained about 6-7% of the variance in lexical features.

We found considerable stability in this structure across conversational English (estimated by the General Service List; West, 1953), general academic English (estimated by the Academic Word List; Coxhead, 2000), and domain-specific English (estimated by the Academic Vocabulary List; Gardner & Davies, 2014). The notable difference was that domain-specific vocabulary better fit a model with four factors, combining Polysemy and Diversity into Poly-Div, which still explained 9% of the variance in lexical features. We theorize that this is because domain-specific words are, by definition, limited in their diversity, so variation on the Diversity factor would be a consequence of multiple meanings (e.g., “slope” has a mathematics-specific meaning but also an everyday meaning). Figures 8 and 9 illustrate the results of the factor analyses for words on the AWL and AVL-DS, respectively.

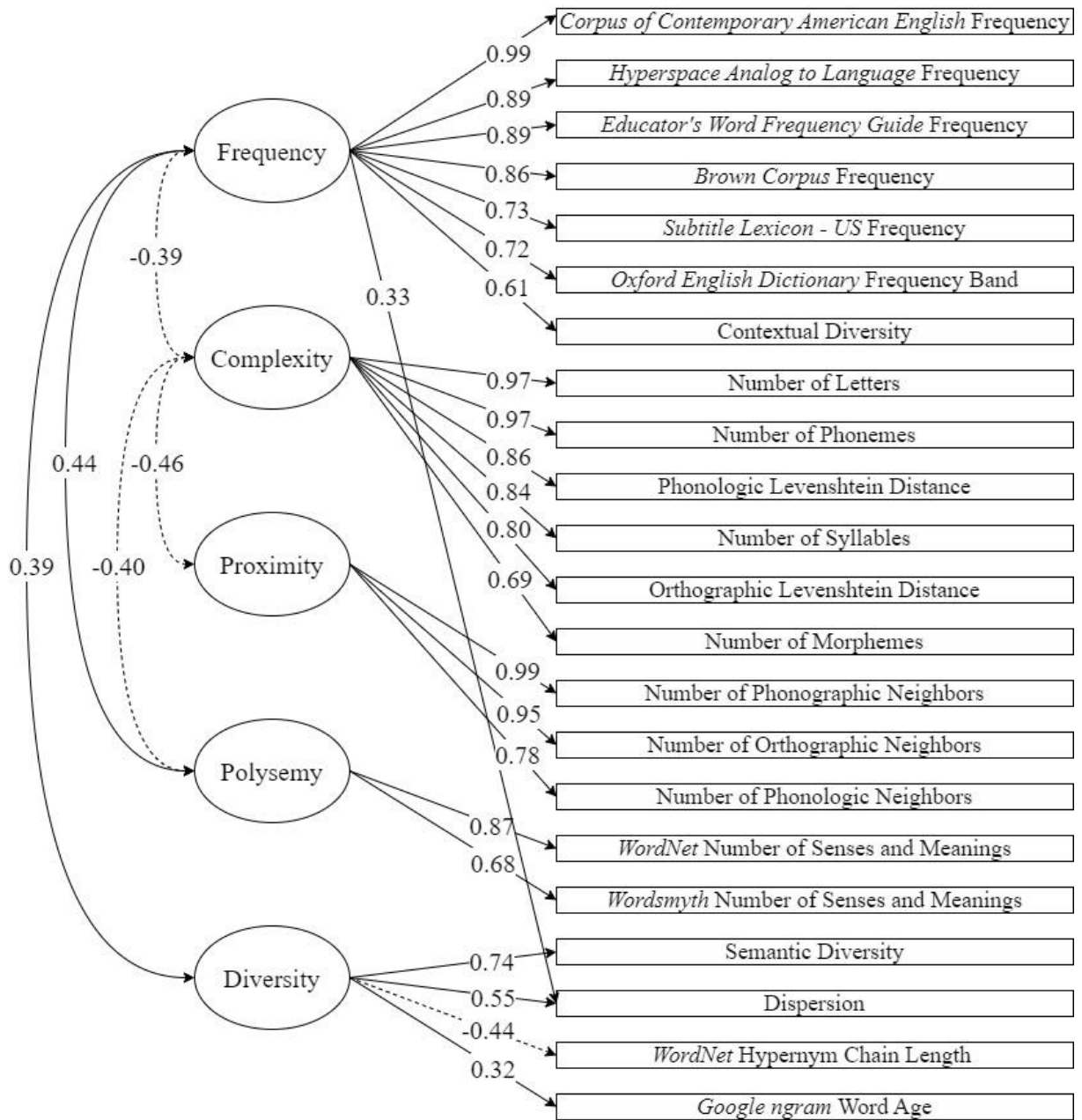


Figure 8. Results of the EFA for AWL words

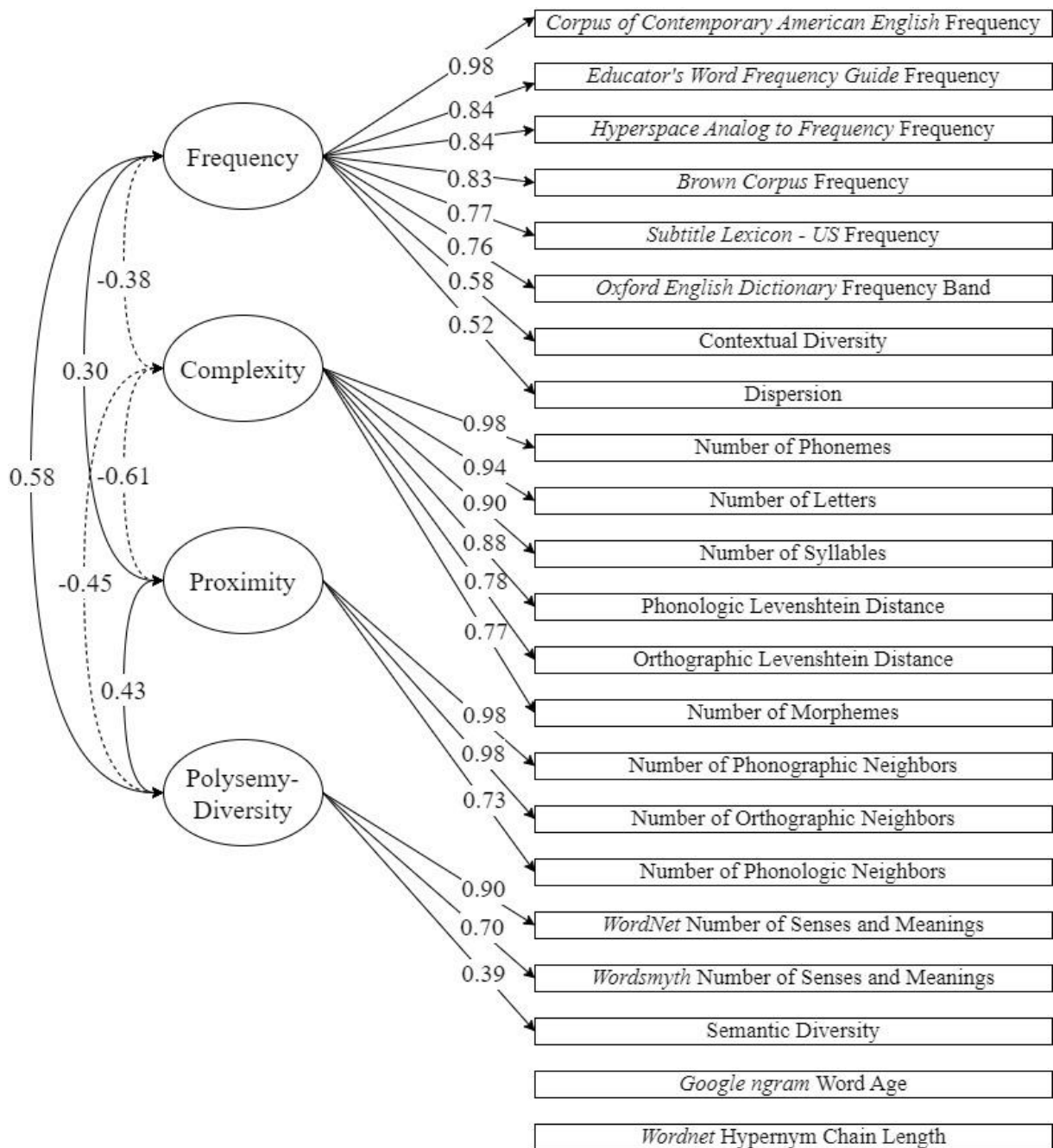


Figure 9. Results of the EFA for domain-specific words

Comments

Though the work in this study was systematic, the lexical dimensions (and, in particular, the measures included) are not an end-all-be-all list. It is impossible to include all possible measures that have been and will be—and I expect to revisit this work with more precise measures across vocabulary components in the future—resulting in more comprehensive and precise estimates of latent dimensions. Still, understanding that this study is not the end-all-be-all does not negate the findings: cutting-edge science is rarely (if ever) complete.

Along the same lines, the three exploratory factor analyses were based on the same lexical feature spaces, but the correlations between measures were contingent on the reference sample. Therefore, different relationships may exist for different samples—for example, function versus content words or Latin versus German roots. Hence, while we can estimate latent scores for any word using any reference model, estimates should be considered within the context of that model, especially when the word is not similar to words in the reference sample.

4.2 Paper 2: Lexical Relationships for the Monolingual Subsample

Paper II examines student differences in vocabulary knowledge as a function of both person characteristics and target-word characteristics for monolingual speakers on synonym tasks. Educational research often considers individual (person) differences in proficiency and ability (McNamara & Magliano, 2009; Hamilton, Freed, & Long, 2016; O'Reilly, Wang & Sabatini, 2019; Ellis & Shintani, 2013), and cognitive research often considers differences in target-word features (e.g., Joseph, Nation, & Liversedge, 2013; Dujardin & Mathey, 2022; Adelman & Brown, 2007; Beretta, Fioentino, Poepell, 2005) but minimal research bridges these two together, particularly via cross-classified interaction (though see for example, Kulesz, et al., 2016). We used explanatory Item Response Theory to simultaneously estimate individual students' vocabulary ability (i.e., predicting theta as a function of grade level, participation in special programs, and general reading comprehension) and individual items' difficulty levels (i.e., predicting the location of item characteristic curves as a function of target word Frequency, Complexity, Proximity, Polysemy, and Diversity; as estimated in Paper I). We also included cross-classified interactions between person-level reading comprehension scores and item-level lexical characteristics to examine if relationships between characteristics differed as a function of reading comprehension scores.

We found that students were more likely to correctly identify synonyms of target words with more senses and meanings (i.e., high values on the latent dimension Polysemy). Notably, the positive relationship of Polysemy overshadowed well-established relationships, such as the word frequency effect and contextual diversity effect (see reviews by Brysbaert, Mander, & Kueleers, 2018, and Johns, Dye, & Jones, 2022; respectively). When Frequency was the sole lexical characteristic in the model, we found a significant relationship; however, when Frequency and Polysemy were in the same model, only Polysemy remained statistically significant.

However, the positive relationship between Frequency and probability of correctly answering an item did vary as a function of reading comprehension scores, such that stronger

readers were more sensitive to word frequency. Conversely, the negative relationship between Complexity and probability of correctly answering an item also varied as a function of reading comprehension scores, wherein weaker readers were more sensitive to target word complexity.

Comments

It is crucial to remember that the dependent variable is the *probability* of answering an item correctly, which serves only as a proxy for word knowledge. Hopefully, this point has become clear throughout this dissertation, but at the time of writing Paper II, it was an important distinction. To say that “more students know words with many meanings” requires more than just the data presented in this paper because defining “know” is not as clear-cut as “correctly answering an item.” And as evidenced in Paper III, identifying a synonym is by no means the only way to “know” a word.

One originally perplexing finding was that the varying slopes for Frequency across levels of reading comprehension scores do not appear to follow the trend expected based on theory. Specifically, the slopes for the struggling reader (-1SD) and the advanced reader (+1SD) appear nearly parallel, while only the average reader (at the mean) shows a steeper slope. However, one critical observation is the limits on the y-axis: the probability of a correct answer. If the slope for the advanced reader were parallel to the average reader, then the advanced reader would have over a 100% probability of correctly identifying the synonym for a high-frequency (+1SD) target word. Thus, it may be that the slope for advanced readers is not as steep because it approaches the upper asymptote: greater than perfect is impossible.

Finally, we must remember that this study involved only a small sample of general academic words and middle-school students. It may be that relationships between reading comprehension, lexical features, and vocabulary knowledge differ when surveying different words and different students. For example, we may find that younger children just learning to read experience a strong relationship with Complexity because they are acquiring decoding skills. Of course, this is a testable hypothesis and worthy of exploration; but the data in this specific paper cannot and does not refute nor accept these hypotheses.

4.3 Paper 3: Lexical Relationships across English Language Learners

Paper III examines relationships between the estimated latent lexical dimensions from Paper I between different English Language Learner (ELL) classifications and across two vocabulary assessments. When assessed via the synonym and definition task on the same word, students found the synonym task easier overall. However, the preference for the synonym task

was not overwhelming; there were many instances where students correctly identified the definition but not the synonym. Moreover, when comparing students by their ELL classification, Limited English Proficient (LEP) students showed no preference, instead finding both tasks equally difficult.

We also found that relationships between lexical characteristics and student performance differed across assessments. For example, students were more likely to correctly identify synonyms when target words were more diverse and polysemous, but the effect of diversity did not hold for the definition task. Additionally, students across ELL classifications showed different relationships between lexical characteristics and performance, even after controlling for differences in reading comprehension scores. For example, LEP students did not find more frequent words as easy as their peers, but were also less hindered by complex words. These findings align with Paper II in that students with lower English proficiency were less sensitive to the word frequency effect.

Comments

An important consideration in Paper III is that the slopes come with the caveat “when all else is equal.” Given that the models in Paper III include a main effect for reading comprehension scores, all other estimates are read as “when the reading comprehension score is zero.” Although reading comprehension scores were standardized, and thus a score of zero is somewhat meaningful (it is the overall average), it is not “the average reader” for the different ELL classifications. Understanding this context is important for Paper III given that, for example, a score of 0 is an entire standard deviation above average for the LEP students ($M = -0.87$, $SD = 0.66$) and half of a standard deviation below average for the IFEP students ($M = 0.49$, $SD = 0.90$).

Second, there is still a lot to unpack within the models reported in Paper III. A subsequent simple slopes analysis may prove fruitful in understanding the complex relationships between vocabulary assessments, lexical characteristics, and different ELL students. Paper III discusses the model slopes, which answer the question “is this estimate significantly different from the reference group?” Conversely, simple slopes analysis would answer the question “is the slope for this lexical feature with this ELL group on this vocabulary assessment significantly different from 0?” These are two distinct questions worthy of future exploration.

Finally, it is worth highlighting that the results in this study are not a result of poor-performing items. If we examine the lowest-performing students, the LEP students performed better than chance, correctly identifying both the synonym and definition 14% of the time

(chance would be 6%) and at least the synonym or definition 57% of the time (chance would be 44%) on average. There is still more work to do in unpacking the complexity of vocabulary for diverse learners, but this does not negate the novel findings of these papers. Exploration of the patterns found here—particularly that only the LEP students performed differently from their peers—is interesting in its own right.

5

Discussion

In this final chapter of the extended abstract, I describe how the findings across the three papers provide methodological and theoretical contributions to the research field. As summarized previously, there are many components to word knowledge, and thus many ways to measure lexical characteristics. However, many measures are highly-correlated, such as measures of word frequency across different corpora, which can cause instability in statistical models. Paper I showed that these individual lexical characteristics can form latent dimensions from a data-driven approach, but that still align with current vocabulary learning theories; such as Nation's (2001) three components: form, meaning, and use.

With latent lexical dimension estimates available, Papers II and III were able to explore relationships in vocabulary learning using complex modelling, without sacrificing statistical power or theoretical clarity that would have occurred if using the original 22 individual lexical characteristics. Paper II indicated that item difficulty varied across latent dimensions, particularly that words with many senses and meanings were easier. Moreover, the relationships varied as a function of standardized reading comprehension scores. Stronger readers were more sensitive to the word frequency effect, and weaker readers were more

sensitive to word complexity (i.e., in relation to form, such as length, syllables, and morphemes) in a monolingual subsample. Paper III extended these findings across diverse English Language Learner (ELL) classifications and between two vocabulary assessments. Overall, students found the synonym task easier than the definition task, but when examined by ELL classification, Limited English Proficient (LEP) students did not demonstrate the same preference; instead finding the tasks equally difficult. Moreover, many students correctly identified definitions but *not* synonyms; indicating that the preference was not sequential. The relationships between lexical characteristics and item difficulty varied between the synonym and definition task as well. Overall, students found target words with many senses and meanings (Polysemy), and used across multiple contexts (Diversity) easier on the synonym task. However, on the definition task, students did not find words used across multiple contexts easier; but they did find long words (high scores on Complexity) easier.

Finally, the relationships between target word characteristics and item difficulty also varied as a function of ELL classification and between assessments. There were few significant differences in slopes between English monolinguals (EOs), Initially Fluent students (IFEPs), and Reclassified students (RFEPs), however the LEP students exhibited different relationships. LEP students were less sensitive to the positive effects of target word frequency and polysemy, but also less sensitive to the negative effects of target word complexity and proximity.

5.1 Methodological Contributions

This dissertation provides two significant methodological contributions to the field: estimated latent lexical dimension scores and examples of the utility of explanatory Item Response Theory. I will discuss these in turn.

5.1.1 Estimated Latent Lexical Dimensions are Valuable and Practical

In Paper I, we systematically compiled lexical feature data from multiple sources to create data-driven latent models using Exploratory Factor Analysis. The results of these can be used to select words for intervention, to match stimuli across assessment forms, as statistical control, or in models as variables of interest. Papers II and III illustrate one way that the estimated latent scores can be used to further vocabulary research. When we explored complex relationships based on five latent dimensions instead of 22 individual measures, we were able to leverage statistical power towards cross-classified interactions without sacrificing valuable information from individual measurements. For example, the Frequency measure includes word frequency in spoken *and* written corpora, so limiting analyses to one corpora means disregarding

information from the other. Even though they are highly correlated, we know that spoken and written English do vary (Brysbaert & New, 2009), thus researchers may want to consider the effects of language exposure as a whole, instead of written or spoken exposure alone.

It is worth noting that, as long as a word has complete data, latent lexical dimension scores can be estimated using the reported models in Paper I—even when the word is not a part of the original sample used to establish the factor models. For instance, “elevator” does not occur on any of the word lists used in Paper I, however, we can estimate the latent factor scores across all three reference models, as shown in Table 1:

Table 1. Estimated latent factor scores for the word "elevator".

Reference Model	Frequency	Complexity	Proximity	Polysemy	Diversity
GSL	-0.76	1.35	-0.73	-1.24	-2.90
AWL	0.21	-0.05	-0.29	-0.46	-1.62
AVL-DS	0.66	0.42	-0.53	-0.42	

5.1.2 Explanatory Item Response Theory is Complex but Advantageous

The analytical procedures in Papers II and III add to the methodological literature for explanatory Item Response Theory, which has been a cutting-edge statistical approach in educational research throughout the last decade (e.g. Kulesz et al., 2016; Francis et al., 2018; Spencer et al., 2019; Elleman et al, 2022). Papers II and III demonstrate the unique advantage of cross-classified doubly-explanatory Item Response Theory models in a way that parallels current theories about reading comprehension. It is not enough to consider individual differences at the person-level (i.e., *theta*), nor to consider individual differences at the text-level (i.e., *beta*). We know that reading comprehension and vocabulary knowledge are related to both person-characteristics and text-characteristics simultaneously, *and* that these characteristics interact with one another (Perfetti & Stafura, 2014; Ahmed et al., 2016, Kulesz et al., 2016); thus, analytical models should reflect the same level of complexity when possible. The statistical code used for Papers II and III are available as online supplemental material in line with the ethical considerations of data transparency; but also in the hopes that future researchers might adapt the code to new studies and take advantage of the procedures used in these papers.

5.2 Theoretical Contributions

This dissertation also contributes to our theoretical understanding of vocabulary words, particularly in what makes general academic vocabulary difficult across diverse learners. In essence, understanding vocabulary requires considering what vocabulary words we want to measure, which students we want to measure, and how we measure vocabulary knowledge.

5.2.1 *Lexical Characteristics can be Empirically Grouped into Dimensions*

Paper I showed that, despite the plethora of lexical characteristics, five latent dimensions remain persistent: Frequency, Complexity, Proximity, Polysemy, and Diversity. While previous studies have used factor analysis to combine multiple lexical characteristics (Brysbaert et al., 2019; Yap, Balota, & Ratcliff, 2012; Clark & Paivio, 2004), this study was the first to do so in a systematic way and across multiple samples of words.

The first novel finding in Paper I is that the non-behavioral lexical characteristics can be synthesized into five latent dimensions. The analyses in Paper I were not directly informed by theory (e.g., as would be the case with Confirmatory Factor Analyses), yet the findings support current theories of vocabulary knowledge. Nation's (2001) three components of vocabulary knowledge align well with the empirically-derived, data-driven factors in Paper I: Complexity and Proximity approximate the form component, while Polysemy parallels the meaning component, and Frequency and Diversity describe the use component.

A second theoretical contribution in Paper I is describing how the dimensions correlate: Frequent words tend to be shorter, more diverse, and have more senses and meanings; complex words tend to be proximal to other words, but also have fewer senses and meanings; and there is no relation between proximity to other words and diversity, for example. Additionally, the fact that dimensions were relatively consistent across the word lists indicated that, while mean scores on lexical features differ between word lists, relationships are relatively similar. We know that there are many characteristics to learn about an individual word, and many ways to learn the same word, so it is not obvious that *relationships* between characteristics would be relatively similar across different word samples. This may be one reason why Nation's (2022) model does not specify how dimensions relate to one another, just that they likely do relate in some way.

5.2.2 *General Academic Words are Not Equally Difficult*

Papers II and III explore what makes general academic words difficult. We know that the learning burden varies across words, for example, complex words take more time to decode

(New et al., 2006; Ehri, 2005; Ellis, 2002) and frequent encounters increase the likelihood that we know a particular word (Brysbaert, Mandera, & Keuleers, 2019; Stahl & Fairbanks, 1986; Swanborn & de Glopper, 1999). However, Paper II is one of the first to explore multiple lexical dimensions at once, along with interactions between each dimension and reading comprehension scores, while Paper III explores interactions between the lexical dimensions and English Language Learner classification across different tasks.

We found that students were more likely to know target words with many senses and meanings, even after controlling for other well-known effects, such as the word frequency effect (Brysbaert, Mandera, & Keuleers, 2019). Notably, the significant slope for word frequency dissipated when Polysemy was included in the model, indicating that much of the frequency effect can be explained by polysemy. This finding aligns with the Lexical Quality Hypothesis (Perfetti & Hart, 2002) and Nation's (2022) model of the development of vocabulary knowledge, because words with many meanings provide more opportunities to integrate usages, as opposed to repeated but identical usages, which in turn supports a stronger and richer lexical representation (González-Fernández & Schmitt, 2019; Nation, 2022, Adelman, Brown, & Quesada, 2006).

5.2.3 Academic Vocabulary Learning is not Linear

Findings from Paper III indicated that, while students overall found the synonym task easier, they did not always correctly identify synonyms before being able to correctly identify definitions. This finding suggests that we are not required to completely master a single component of a vocabulary word before we can begin to develop knowledge of other components. Instead, it is much more likely that, even though components of vocabulary knowledge are acquired at different trajectories, we learn multiple components simultaneously (Schmitt, 2019). This is potentially why Nation's (2022) model of the development of vocabulary knowledge suggests multiple components but makes no distinction about the order of acquisition.

Additionally, Paper III indicates that item difficulty is related to lexical dimensions in different ways across tasks. Students may approach the synonym and definition tasks differently, and therefore be able to leverage different skills in order to answer items correctly. For example, being exposed to a target word in multiple contexts may strengthen the lexical representation of that word, particularly in relation to implicit knowledge about that word, which could make identifying synonyms easier. However, the response options on the

definition task cover only one specific usage of the target word, and thus, the word being used in multiple contexts may not be particularly helpful.

5.2.4 Monolingual and Multilingual Learning is Qualitatively Different

Looking across the interaction effects in Papers II and III exemplify the fact that deficits in vocabulary and reading comprehension stem from many challenges. In their *Simple View of Reading*, Gough & Tunmer (1986) suggest that reading comprehension can fall apart because of decoding deficits or language deficits; which we see evidenced in Papers II and III, respectively.

In Paper II, the lowest-performing monolinguals (i.e., students with low reading comprehension scores) were more sensitive to target word complexity, while in Paper III, the lowest-performing English Language Learners (i.e., limited English proficient students) were less sensitive to target word complexity. The different results across the two papers illustrate one way in which monolingual and multilingual vocabulary learning differs. Multilingual learners often develop L2 vocabulary skills and L2 literacy simultaneously, while monolingual students usually develop a foundation of L1 vocabulary orally before developing literacy (Nation & Snowling, 2000; Chall, 1996). While multilingual learners can leverage decoding skills from their L1 (Cummins, 2007; Kuo & Anderson, 2006), monolingual learners cannot. At the same time, multilingual learners are more likely to exhibit limited L2 vocabulary compared to their peers (Spencer & Wagner, 2018), even after several years of L2 instruction (Farnia & Geva, 2011). This is not to say that monolingual students cannot struggle with vocabulary knowledge or that multilingual students cannot struggle with decoding; but that overall, struggling readers with diverse linguistic backgrounds face different challenges.

5.3 Implications and Future Directions

The methodological and theoretical contributions discussed in the proceeding sections imply that what words we measure, which students we measure, and how we measure their knowledge are critical. In an era where advanced technology allows for automatic generation of vocabulary assessments via artificial intelligence and large language models, understanding the ways in which vocabulary knowledge varies as a function of lexical characteristics, individual differences, and assessment types is critical in generating items and interpreting results. By the same token, the results of this dissertation can inform future vocabulary and reading comprehension intervention research to generate interventions tailored to different academic contexts.

The overarching aim of this dissertation was to bridge some of the gaps between cognitive research on different words and educational research on diverse learners, while leveraging advantages from cross-classified interactions in explanatory Item Response Theory. While I knew that vocabulary learning was a complex process (both as a result of reading empirical research, and from my own experience learning a second language as an adult while my bilingual toddler nearly surpasses me), I have learned that different learners are not only faced with specific challenges, but also approach vocabulary learning with different strengths. Similarly, different words pose specific challenges and strengths as well. When we are cognizant of these person and text differences, we can better leverage strengths to overcome challenges in vocabulary learning and reading comprehension in general.

6

References

- Aarnoutse, C., van Leeuwe, J., Voeten, M., & Oud, H. (2001). Development of decoding, reading comprehension, vocabulary and spelling during the elementary school years. *Reading and Writing, 14*(1), 61–89.
- Adelman, J. S., & Brown, G. D. A. (2007). Phonographic neighbors, not orthographic neighbors, determine word naming latencies. *Psychonomic Bulletin & Review, 14*(3), 455–459.
- 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.
- Adesope, O. O., Lavin, T., Thompson, T., & Ungerleider, C. (2010). A Systematic Review and Meta-Analysis of the Cognitive Correlates of Bilingualism. *Review of Educational Research, 80*(2), 207–245.
- Ahmed, Y., Francis, D. J., York, M., Fletcher, J. M., Barnes, M., & Kulesz, P. (2016). Validation of the direct and inferential mediation (DIME) model of reading comprehension in grades 7 through 12. *Contemporary Educational Psychology, 44-45*, 68–82.

- Alexander, P. (2012). Reading into the future: Competence for the 21st century. *Educational Psychologist*, 47(4), 259–280.
- Anderson, R. C., & Freebody, P. (1981). Vocabulary knowledge. In J. T. Guthrie (Ed.), *Comprehension and Teaching: Research Reviews* (pp. 77–117). International Reading Association.
- Andrews, S. (1996). Lexical retrieval and selection processes: Effects of transposed-letter confusability. *Journal of Memory and Language*, 35(6), 775–800.
- Andrews, S., & Hersch, J. (2010). Lexical precision in skilled readers: Individual differences in masked neighbor priming. *Journal of Experimental Psychology. General*, 139(2), 299–318.
- Anglin, J. M., Miller, G. A., & Wakefield, P. C. (1993). Vocabulary development: A morphological analysis. *Monographs of the Society for Research in Child Development*, 58(10), 1–186.
- Archer, A. L., Gleason, M. M., & Vachon, V. L. (2003). Decoding and fluency: Foundation skills for struggling older readers. *Learning Disability Quarterly: Journal of the Division for Children with Learning Disabilities*, 26(2), 89–101.
- 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.
- Baddeley, A. (2003). Working memory and language: An overview. *Journal of Communication Disorders*, 36(3), 189–208.
- Bar-Ilan, L., & Berman, R. A. (2007). Developing register differentiation: the Latinate-Germanic divide in English. *Linguistics*, 45(1), 1–35.
- Barcroft, J. (2009). Strategies and performance in intentional L2 vocabulary learning. *Language Awareness*, 18(1), 74–89.
- Baumann, J. F., & Graves, M. F. (2010). What is academic vocabulary? *Journal of Adolescent & Adult Literacy: A Journal from the International Reading Association*, 54(1), 4–12.
- Beck, I. L., McKeown, M. G., & Kucan, L. (2002). *Bringing Words to Life: Robust Vocabulary Instruction*. Guilford Press.
- Ben-Simon, A., Budescu, D. V., & Nevo, B. (1997). A Comparative study of measures of partial knowledge in multiple-choice tests. *Applied Psychological Measurement*, 21(1), 65–88.
- 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.
- Bhattacharya, A., & Ehri, L. C. (2004). Graphosyllabic analysis helps adolescent struggling readers read and spell words. *Journal of Learning Disabilities*, 37(4), 331–348.

- Bialystok, E., Peets, K. F., & Moreno, S. (2014). Producing bilinguals through immersion education: Development of metalinguistic awareness. *Applied Psycholinguistics*, 35(1), 177–191.
- Biemiller, A. (2012). Teaching vocabulary in the primary grades: Vocabulary instruction needed. In E. J. Kame'enui & J. F. Baumann (Eds.), *Vocabulary Instruction: Research to Practice* (2nd ed., pp. 34–50). Guilford Press.
- 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.
- Binder, K. S., Cote, N. G., Lee, C., Bessette, E., & Vu, H. (2017). Beyond breadth: The contributions of vocabulary depth to reading comprehension among skilled readers. *Journal of Research in Reading*, 40(3), 333–343.
- Bohn-Gettler, C. M., & Kendeou, P. (2014). The Interplay of reader goals, working memory, and text structure during reading. *Contemporary Educational Psychology*, 39(3), 206–219.
- Borodkin, K., Kenett, Y. N., Faust, M., & Mashal, N. (2016). When pumpkin is closer to onion than to squash: The Structure of the second language lexicon. *Cognition*, 156, 60–70.
- Borowsky, R., & Masson, M. E. J. (1996). Semantic ambiguity effects in word identification. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, 22(1), 63.
- Braze, D., Tabor, W., Shankweiler, D. P., & Mencl, W. E. (2007). Speaking up for vocabulary: Reading skill differences in young adults. *Journal of Learning Disabilities*, 40(3), 226–243.
- Brett, A., Rothlein, L., & Hurley, M. (1996). Vocabulary acquisition from listening to stories and explanations of target words. *The Elementary School Journal*, 96(4), 415–422.
- Brown, J. C., Frishkoff, G. A., & Eskenazi, M. (2005). Automatic Question Generation for Vocabulary Assessment. *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, 819–826.
- Brysbaert, M., Mander, P., & Keuleers, E. (2018). The Word frequency effect in word processing: An Updated review. *Current Directions in Psychological Science*, 27(1), 45–50.
- Brysbaert, M., Mander, P., McCormick, S. F., & Keuleers, E. (2019). Word prevalence norms for 62,000 English lemmas. *Behavior Research Methods*, 51(2), 467–479.
- Brysbaert, M., & New, B. (2009). Moving beyond Kucera 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.
- Brysbaert, M., Stevens, M., Mander, P., & Keuleers, E. (2016). How many words do we know? Practical estimates of vocabulary size dependent on word definition, the degree

- of language input and the participant's age. *Frontiers in Psychology*, 7, 1116.
- Buly, M. R., & Valencia, S. W. (2002). Below the bar: Profiles of students who fail state reading assessments. *Educational Evaluation and Policy Analysis*, 24(3), 219–239.
- Cai, L., & Thissen, D. (2014). Modern Approaches to Parameter Estimation in Item Response Theory. In S. P. Reise & D. A. Revicki (Eds.), *Handbook of Item Response Theory Modeling Applications to Typical Performance Assessment* (pp. 41–59). Routledge.
- Cain, K., & Oakhill, J. (2014). Reading comprehension and vocabulary: Is vocabulary more important for some aspects of comprehension? *L'annee Psychologique*, 114(4), 647–662.
- Cain, K., Oakhill, J., & Bryant, P. (2004). Children's reading comprehension ability: Concurrent prediction by working memory, verbal ability, and component skills. *Journal of Educational Psychology*, 96(1), 31–42.
- Cain, K., Oakhill, J., & Lemmon, K. (2004). Individual differences in the inference of word meanings from context: The Influence of reading comprehension, vocabulary knowledge, and memory capacity. *Journal of Educational Psychology*, 96(4), 671–681.
- Carlisle, J. F. (2000). Awareness of the structure and meaning of morphologically complex words: Impact on reading. *Reading and Writing*, 12(3), 169–190.
- Carlisle, J. F., & Beeman, M. M. (2000). The Effects of language of instruction on the reading and writing achievement of first-grade hispanic children. *Scientific Studies of Reading*, 4(4), 331–353.
- Carlisle, J. F., & Stone, C. A. (2005). Exploring the role of morphemes in word reading. *Reading Research Quarterly*, 40(4), 428–449.
- Carretti, B., Caldarola, N., Tencati, C., & Cornoldi, C. (2014). Improving reading comprehension in reading and listening settings: The Effect of two training programmes focusing on metacognition and working memory. *The British Journal of Educational Psychology*, 84(Pt 2), 194–210.
- 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.
- Cevoli, B., Watkins, C., & Rastle, K. (2021). What is semantic diversity and why does it facilitate visual word recognition? *Behavior Research Methods*, 53(1), 247–263.
- Chall, J. S., Jacobs, V. A., Baldwin, L. E. (1990). *The Reading Crisis: Why Poor Children Fall Behind*. Boston, MA: Harvard University Press.
- Clark, J. M., & Paivio, A. (2004). Extensions of the Paivio, Yuille, and Madigan (1968) norms. *Behavior Research Methods, Instruments, & Computers*, 36(3), 371–383.
- Cohen, J. (1983). The Cost of dichotomization. *Applied Psychological Measurement*, 7(3), 249–253.

- Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences*. Lawrence Erlbaum Associates Inc.
- Collins, A. M., & Quillian, M. R. (1969). Retrieval time from semantic memory. *Journal of Verbal Learning and Verbal Behavior*, 8(2), 240–247.
- Coltheart, M., Davelaar, E., Jonasson, J.T., & Besner, D. (1977). Access to the internal lexicon. In S. Dornic (Ed.), *Attention and Performance, VI* (pp. 535–555). Lawrence Erlbaum Associates.
- Cop, U., Keuleers, E., Drieghe, D., & Duyck, W. (2015). Frequency effects in monolingual and bilingual natural reading. *Psychonomic Bulletin & Review*, 22(5), 1216–1234.
- Copland, D. A., de Zubicaray, G. I., McMahon, K., Wilson, S. J., Eastburn, M., & Chenery, H. J. (2003). Brain activity during automatic semantic priming revealed by event-related functional magnetic resonance imaging. *NeuroImage*, 20(1), 302–310.
- Coxhead, A. (2000). An Introduction to the Academic Word List. *TESOL Quarterly*, 34(2), 213–238.
- Coxhead, A. (2011). The Academic Word List 10 years on: Research and teaching implications. *TESOL Quarterly*, 45(2), 355–362.
- Coxhead, A., & Hirsch, D. (2007). A pilot science-specific word list. *Revue Francaise de Linguistique Appliquee*, 2, 65–78.
- 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.
- Crossley, S., Salsbury, T., & McNamara, D. (2009). Measuring L2 lexical growth using hypernymic relationships. *Language Learning*, 59(2), 307–334.
- 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.
- Crosson, A. C., McKeown, M. G., & Ward, A. K. (2019). An Innovative approach to assessing depth of knowledge of academic words. *Language Assessment Quarterly*, 16(2), 196–216.
- Cummins, J. (1978). Bilingualism and the development of metalinguistic awareness. *Journal of Cross-Cultural Psychology*, 9(2), 131–149.
- Cummins, J. (2007). Rethinking monolingual instructional strategies in multilingual classrooms. *Canadian Journal of Applied Linguistics*, 12(2), 221–240.
- Cutting, L. E., & Scarborough, H. S. (2006). Prediction of reading comprehension: Relative contributions of word recognition, language proficiency, and other cognitive skills can depend on how comprehension is measured. *Scientific Studies of Reading*, 10(3), 277–

- Dale, E., & O'Rourke, J. (1986). *Vocabulary Building: A Process Approach* (W. B. Barbe (ed.)). Zaner-Bloser.
- Daller, H., & Xue, H. (2007). Lexical richness and the oral proficiency of Chinese EFL students. In H. Daller, M. J., & J. Treffers-Daller (Eds.), *Modelling and Assessing Vocabulary Knowledge* (pp. 93–115). Cambridge University Press.
- Dattalo, P. (2008). *Determining Sample Size: Balancing Power, Precision, and Practicality*. Oxford University Press, USA.
- 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.
- Davis, C. J. (2006). Orthographic input coding: A Review of behavioural data and current models. In S. Andrews (Ed.), *From Inkmarks to Ideas: Challenges and Controversies about Word Recognition and Reading* (pp. 180–206). Academic Press.
- de Ayala, R. J. (2009). *The Theory and Practice of Item Response Theory*. The Guilford Press.
- de Boeck, P., & Wilson, M. R. (2016). Explanatory Item Response Models. In W. J. van der Linden (Ed.), *Handbook of Item Response Theory: Volume 1: Models* (pp. 565–580). CRC Press.
- de Groot, A. M. B., & van Hell, J. G. (2005). The learning of foreign language vocabulary. In J. F. Kroll & A. M. B. de Groot (Eds.), *Handbook of bilingualism: Psycholinguistic approaches* (pp. 9–29). Oxford University Press.
- Deane, P., Lawless, R. R., Li, C., Sabatini, J., Bejar, I. I., & O'Reilly, T. (2014). Creating vocabulary item types that measure students' depth of semantic knowledge. *ETS Research Report Series, 2014*(1), 1–19.
- DeCoster, J., Gallucci, M., & Iselin, A.-M. R. (2011). best practices for using median splits, artificial categorization, and their continuous alternatives. *Journal of Experimental Psychopathology*, 2(2), 197–209.
- Degaetano-Ortlieb, S., Kermes, H., Lapshinova-Koltunski, E., & Teich, E. (2013). SciTex-a diachronic corpus for analyzing the development of scientific registers. *New Methods in Historical Corpus Linguistics*, 3, 93–104.
- DeRocher, J. E. (1973). *The Counting of Words: A Review of the History, Techniques and Theory of Word Counts with Annotated Bibliography*. Syracuse University Research Corp.
- Dijkstra, T., Wahl, A., Buytenhuijs, F., Van Halem, N., Al-Jibouri, Z., de Korte, M., & Rekké, S. (2019). Multilink: a Computational model for bilingual word recognition and word translation. *Bilingualism: Language and Cognition*, 22(4), 657–679.

- Dóczy, B. (2019). An overview of conceptual models and theories of lexical representation in the mental lexicon. *The Routledge Handbook of Vocabulary Studies*.
- Dujardin, E., & Mathey, S. (2022). The Neighbourhood frequency effect in naming is influenced by substituted-letter confusability and lexical skills. *Journal of Cognitive Psychology*, 34(8), 947–961.
- Duke, N. K., & Cartwright, K. B. (2021). The Science of reading progresses: Communicating advances beyond the Simple View of reading. *Reading Research Quarterly*, 56(S1). <https://doi.org/10.1002/rrq.411>
- Duñabeitia, J. A., Perea, M., & Carreiras, M. (2010). Masked translation priming effects with highly proficient simultaneous bilinguals. *Experimental Psychology*, 57(2), 98–107.
- Dupuy, H. P. (1974). *The Rationale, Development and Standardization of a Basic Word Vocabulary Test*. Washington, DC: U.S. Government Printing Office. (DHEW Publications No. HRA 74-1334)
- 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.
- Edwards, J., Beckman, M. E., & Munson, B. (2004). The Interaction between vocabulary size and phonotactic probability effects on children’s production accuracy and fluency in nonword repetition. *Journal of Speech, Language, and Hearing Research*, 47(2), 421–436.
- Ehri, L. C. (1995). Phases of development in learning to read words by sight. *Journal of Research in Reading*, 18(2), 116–125.
- Ehri, L. C. (2005). Learning to read words: Theory, findings, and issues. *Scientific Studies of Reading*, 9(2), 167–188.
- Ehri, L. C. (2014). Orthographic mapping in the acquisition of sight word reading, spelling memory, and vocabulary learning. *Scientific Studies of Reading*, 18(1), 5–21.
- 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.
- Elleman, A. M., Steacy, L. M., Gilbert, J. K., Cho, E., Miller, A. C., Coyne-Green, A., Pritchard, P., Fields, R. S., Schaeffer, S., & Compton, D. L. (2022). Exploring the role of knowledge in predicting reading and listening comprehension in fifth grade students. *Learning and Individual Differences*, 98, 102182.
- Ellis, N. C. (2002). Frequency effects in language processing: A Review with implications for theories of implicit and explicit language acquisition. *Studies in Second Language Acquisition*, 24(2), 143–188.

- Ellis, R. (2016). Focus on form: A Critical review. *Language Teaching Research*, 20(3), 405–428.
- Ellis, R., & Shintani, N. (2013). *Exploring Language Pedagogy through Second Language Acquisition Research*. Routledge.
- Embretson, S. E., & Reise, S. P. (2013). *Item Response Theory*. Psychology Press.
- European Parliament and Council of European Union (2016). *Regulation (EU) 2016/679*. Retrieved from https://lovdata.no/dokument/NL/lov/2018-06-15-38/KAPITTEL_gdpr-2#KAPITTEL_gdpr-2
- Fellbaum, C. (2010). WordNet. In R. Poli, M. Healy, & A. Kameas (Eds.), *Theory and Applications of Ontology: Computer Applications* (pp. 231–243). Springer Netherlands.
- Ferré, P., Sánchez-Casas, R., Comesaña, M., & Demestre, J. (2017). Masked translation priming with cognates and noncognates: Is There an effect of words' concreteness?. *Bilingualism: Language and Cognition*, 20(4), 770–782.
- Ferreira, F., & Henderson, A. M. (1991). Recovery from misanalyses of 'garden-path sentences. *Journal of Memory and Language*, 30, 725-745.
- Ferreira, F., Christianson, K., & Hollingworth, A. (2001). Misinterpretations of garden-path sentences: Implications for models of sentence processing and reanalysis. *Journal of Psycholinguistic Research*, 30(1), 3–20.
- Field, A. (2013). *Discovering Statistics Using IBM SPSS Statistics*. SAGE.
- Fletcher, J. M., Catts, H. W., & Tomblin, J. B. (2005). Dimensions affecting the assessment of reading comprehension, in S. G. Paris & S. A. Stahl (eds) *Children's Reading Comprehension and Assessment*. New York: Routledge.
<https://doi.org/10.4324/9781410612762-26/>
- Foorman, B. R., Petscher, Y., & Herrera, S. (2018). Unique and common effects of decoding and language factors in predicting reading comprehension in grades 1–10. *Learning and Individual Differences*, 63, 12–23.
- Forskningsetikk (2019). Guidelines for Research Ethics in the Social Sciences, Humanities, Law and Theology. Retrieved from <https://www.forskningsetikk.no/en/guidelines/social-sciences-humanities-law-and-theology/guidelines-for-research-ethics-in-the-social-sciences-humanities-law-and-theology/>
- Forster, K. I., Davis, C., Schoknecht, C., & Carter, R. (1987). Masked priming with graphemically related forms: Repetition or partial activation? *The Quarterly Journal of Experimental Psychology Section A*, 39(2), 211–251.
- Fowler, A., & Liberman, I. Y. (1995). The role of phonology and orthography in morphological awareness. In L. Feldman (Ed.), *Morphological Aspects of Language Processing* (pp. 189–209). Mahwah, NJ: Lawrence Erlbaum Associates, Inc

- Frances, C., Martin, C. D., & Duñabeitia, J. A. (2020). The Effects of contextual diversity on incidental vocabulary learning in the native and a foreign language. *Scientific Reports*, *10*(1), 13967.
- Galambos, S. J., & Goldin-Meadow, S. (1990). The Effects of learning two languages on levels of metalinguistic awareness. *Cognition*, *34*(1), 1–56.
- Gándara, P., & Remberger, R. W. (2007). *Resource Needs for California's English Learners*. Institute for Research on Education Policy & Practice. <https://cepa.stanford.edu/sites/default/files/Gandara.pdf>
- García, J. R., & Cain, K. (2014). Decoding and reading comprehension: A Meta-analysis to identify which reader and assessment characteristics influence the strength of the relationship in English. *Review of Educational Research*, *84*(1), 74–111.
- Gardner, D., & Davies, M. (2014). A New Academic Vocabulary List. *Applied Linguistics*, *35*(3), 305–327.
- German, D. J., & Newman, R. S. (2004). The Impact of lexical factors on children's word-finding errors. *Journal of Speech, Language, and Hearing Research*, *47*(3), 624–636.
- Gernsbacher, M. A., & Faust, M. E. (1991). The mechanism of suppression: A Component of general comprehension skill. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *17*(2), 245–262.
- Giesbrecht, B., Camblin, C. C., & Swaab, T. Y. (2004). Separable effects of semantic priming and imageability on word processing in human cortex. *Cerebral Cortex*, *14*(5), 521–529.
- Gilmore, A. (2007). Authentic materials and authenticity in foreign language learning. *Language Teaching*, *40*(2), 97–118.
- Gilner, L. (2011). A primer on the General Service List. *Reading in a Foreign Language*, *23*(1), 65–83.
- Gimenes, M., & New, B. (2016). Worldlex: Twitter and blog word frequencies for 66 languages. *Behavior Research Methods*, *48*(3), 963–972.
- González-Fernández, B., & Schmitt, N. (2019). Word Knowledge: Exploring the relationships and order of acquisition of vocabulary knowledge components. *Applied Linguistics*. <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.
- Goodwin, A. P., Gilbert, J. K., & Cho, S. J. (2013). Morphological contributions to adolescent word reading: An item response approach. *Reading Research Quarterly*, *48*(1), 39-60. <https://ila.onlinelibrary.wiley.com/doi/abs/10.1002/rrq.037>

- Gough, P. B., & Tunmer, W. E. (1986). Decoding, reading, and reading disability. *Remedial and Special Education*, 7(1), 6–10.
- Goulden, R., Nation, P., & Read, J. (1990). How large can a receptive vocabulary be? *Applied Linguistics*, 11(4), 341–363.
- Grabe, W. (2012). *Reading in a Second Language: Moving from Theory to Practice*. Cambridge University Press.
- Granger, S. (2003). The International Corpus of Learner English: A new resource for foreign language learning and teaching and second language acquisition research. *TESOL Quarterly*, 37(3), 538–546.
- Hacker, D. J. (1997). Comprehension monitoring of written discourse across early-to-middle adolescence. *Reading and Writing*, 9(3), 207–240.
- Haladyna, T. M. (2004). *Developing and Validating Multiple-choice Test Items*. Routledge.
- Haladyna, T. M., Downing, S. M., & Rodriguez, M. C. (2002). A Review of multiple-choice item-writing guidelines for classroom assessment. *Applied Measurement in Education*, 15(3), 309–333.
- Hamilton, S., Freed, E., & Long, D. L. (2016). Word-decoding skill interacts with working memory capacity to influence inference generation during reading. *Reading Research Quarterly*, 51(4), 391–402.
- Hart, B., & Risley, T. R. (1995). *Meaningful Differences in the Everyday Experience of Young American Children*. Paul H Brookes Publishing.
- Heckman, J. J. (1979). Sample selection bias as a specification error. *Econometrica*, 47(1), 153–161.
- Henning, G. H. (1973). Remembering foreign language vocabulary: Acoustic and semantic parameters. *Language Learning*, 23(2), 185–196.
- Hiebert, E. H., & Fisher, C. W. (2005). A Review of the National Reading Panel’s studies on fluency: The Role of text. *The Elementary School Journal*, 105(5), 443–460.
- 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.
- Hiebert, E. H., & Lubliner, S. (2008). The nature, learning, and instruction of general academic vocabulary. In S.J. Samuels & A. Farstrup (Eds.), *What Research has to Say about Vocabulary* (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 Sciences*, 9(1), 8.
- Hill, L. E. (2012). California’s English Learner Students. *California Counts / Public Policy Institute of California*. http://www.ppic.org/main/publication_quick.asp?i=1031

- Hill, L. E., Weston, M., & Hayes, J. M. (2014). *Reclassification of English learner students in California*. Retrieved from https://www.ppic.org/wp-content/uploads/content/pubs/report/R_114LHR.pdf
- Hinkel, E. (2022). Teaching and Learning Multiword Expressions. In *Handbook of Practical Second Language Teaching and Learning* (pp. 435–448). Routledge.
- 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.
- 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.
- Hoff, E. (2006). How social contexts support and shape language development. *Developmental Review*, 26(1), 55–88.
- 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.
- Hoover, W. A., & Gough, P. B. (1990). The Simple View of reading. *Reading and Writing*, 2(2), 127–160.
- Hu, M., & Nation, P. (2000). Unknown vocabulary density and reading comprehension. *Reading in a Foreign Language*, 13(1), 403–430.
- Huckin, T., & Coady, J. (1999). Incidental vocabulary acquisition in a second language: A Review. *Studies in Second Language Acquisition*, 21(2), 181–193.
- Humphreys, L. G., & Fleishman, A. (1974). Pseudo-orthogonal and other analysis of variance designs involving individual-differences variables. *Journal of Educational Psychology*, 66(4), 464–472.
- Hyland, K., & Tse, P. (2007). Is There an “academic vocabulary”? *TESOL Quarterly*, 41(2), 235–253.
- Hyönä, J., & Olson, R. K. (1995). Eye fixation patterns among dyslexic and normal readers: Effects of word length and word frequency. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, 21(6), 1430–1440.
- Jacobs, A., Fricke, M., & Kroll, J. F. (2016). Cross-language activation begins during speech planning and extends into second language speech. *Language Learning*, 66(2), 324–353.
- Jalbert, A., Neath, I., & Surprenant, A. M. (2011). Does length or neighborhood size cause the word length effect? *Memory & Cognition*, 39(7), 1198–1210.

- Jaradat, D., & Tollefson, N. (1988). The Impact of alternative scoring procedures for multiple-choice items on test reliability, validity, and grading. *Educational and Psychological Measurement*, 48(3), 627–635.
- Jarvis, S., & Pavlenko, A. (2008). *Crosslinguistic Influence in Language and Cognition*. Routledge.
- Jenkins, J. R., Fuchs, L. S., van den Broek, P., Espin, C., & Deno, S. L. (2003). Sources of individual differences in reading comprehension and reading fluency. *Journal of Educational Psychology*, 95(4), 719–729.
- Jeon, E. H., & Yamashita, J. (2014). L2 reading comprehension and its correlates: A Meta-analysis. *Language Learning*, 64(1), 160–212.
- Jiang, N. (2002). Form–meaning mapping in vocabulary acquisition in a second language. *Studies in Second Language Acquisition*, 24(4), 617–637.
- Johns, B. T., Dye, M., & Jones, M. N. (2016). The influence of contextual diversity on word learning. *Psychonomic Bulletin & Review*, 23(4), 1214–1220.
- Jones, M. N., Johns, B. T., & Recchia, G. (2012). The role of semantic diversity in lexical organization. *Canadian Journal of Experimental Psychology*, 66(2), 115–124.
- Joseph, H. S. S. L., & Liversedge, S. P. (2013). Children’s and adults’ on-line processing of syntactically ambiguous sentences during reading. *PloS One*, 8(1), e54141.
- Joseph, H. S. S. L., Nation, K., & Liversedge, S. P. (2013). Using eye movements to investigate word frequency effects in children’s sentence reading. *School Psychology Review*, 42(2), 207–222.
- Joyce, P. (2018). L2 vocabulary learning and testing: The Use of L1 translation versus L2 definition. *The Language Learning Journal*, 46(3), 217–227.
- Just, M. A., & Carpenter, P. A. (1980). A theory of reading: From eye fixations to comprehension. *Psychological Review*, 87(4), 329–354.
- Keenan, J. M., Betjemann, R. S., & Olson, R. K. (2008). Reading comprehension tests vary in the skills they assess: Differential dependence on decoding and oral comprehension. *Scientific Studies of Reading*, 12(3), 281–300.
- Keenan, J. M., & Meenan, C. E. (2014). Test differences in diagnosing reading comprehension deficits. *Journal of Learning Disabilities*, 47(2), 125–135.
- Khorramdel, L., von Davier, M., Gonzalez, E., & Yamamoto, K. (2020). Plausible values: Principles of Item Response Theory and Multiple Imputations. In D. B. Maehler & B. Rammstedt (Eds.), *Large-Scale Cognitive Assessment: Analyzing PIAAC Data* (pp. 27–47). Springer International Publishing.
- 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.

- Kintsch, W. (1988). The role of knowledge in discourse comprehension: A Construction-integration model. *Psychological Review*, 95(2), 163–182.
- Kintsch, W., & Rawson, K. A. (2005). Comprehension. In M. J. M. J. Snowling & C. Hulme (Eds.), *The Science of Reading: A Handbook* (pp. 209–226). Blackwell.
- Klauda, S. L., & Guthrie, J. T. (2008). Relationships of three components of reading fluency to reading comprehension. *Journal of Educational Psychology*, 100(2).
<https://doi.org/10.1037/0022-0663.100.2.310>
- Klein, D. E., & Murphy, G. L. (2001). The Representation of polysemous words. *Journal of Memory and Language*, 45(2), 259–282.
- Klepousniotou, E., & Baum, S. R. (2007). Disambiguating the ambiguity advantage effect in word recognition: An advantage for polysemous but not homonymous words. *Journal of Neurolinguistics*, 20(1), 1–24.
- Klepousniotou, E., Titone, D., & Romero, C. (2008). Making sense of word senses: The Comprehension of polysemy depends on sense overlap. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, 34(6), 1534–1543.
- Kornai, A. (2002). How many words are there? *Glottometrics*, 4, 61–86.
- Kotz, S. A., Cappa, S. F., von Cramon, D. Y., & Friederici, A. D. (2002). Modulation of the lexical-semantic network by auditory semantic priming: An event-related functional MRI study. *NeuroImage*, 17(4), 1761–1772.
- Kroll, J. F., Michael, E., Tokowicz, N., & Dufour, R. (2002). The Development of lexical fluency in a second language. *Second Language Research*, 18(2), 137–171.
- Kroll, J. F., van Hell, J. G., Tokowicz, N., & Green, D. W. (2010). The Revised Hierarchical Model: A Critical review and assessment. *Bilingualism*, 13(3), 373–381.
- Kuo, L.-J., & Anderson, R. C. (2006). Morphological awareness and learning to read: A Cross-language perspective. *Educational Psychologist*, 41(3), 161–180.
- Landauer, T. K., Kireyev, K., & Panaccione, C. (2011). Word naturity: A New metric for word knowledge. *Scientific Studies of Reading*, 15(1), 92–108.
- Lau, P. N. K., Lau, S. H., Hong, K. S., & Usop, H. (2011). Guessing, partial knowledge, and misconceptions in multiple-choice tests. *Journal of Educational Technology & Society*, 14(4), 99–110.
- Laufer, B. (1989). What percentage of text-lexis is essential for comprehension? In C. Lauren & M. Nordman (Eds.), *Special Language: From Humans to Thinking Machines* (pp. 316–323). Multilingual Matters.
- Laufer, B. (1992). How Much Lexis is Necessary for Reading Comprehension? In P. J. L. Arnaud & H. Béjoint (Eds.), *Vocabulary and Applied Linguistics* (pp. 126–132).

Palgrave Macmillan.

- Laufer, B. (2003). Vocabulary acquisition in a second language: Do Learners really acquire most vocabulary by reading? Some empirical evidence. *The Canadian Modern Language Review*, 59(4), 567–587.
- Laufer, B., & Goldstein, Z. (2004). Testing vocabulary knowledge: Size, strength, and computer adaptiveness. *Language Learning*, 54(3), 399–436.
- Laufer, B., & Ravenhorst-Kalovski, G. C. (2010). Lexical threshold revisited: Lexical coverage, learners' vocabulary size and reading comprehension. *Reading in a Foreign Language* 22, 15–30.
- Lawrence, J., Capotosto, L., Branum-Martin, L., White, C. & Snow, C. (2012). Language proficiency, home-language status and English vocabulary development: A longitudinal follow-up of the Word Generation program. *Bilingualism: Language and Cognition*. 15(3). 437–451.
- 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.
- Lawrence, J. F., Maher, B., & Snow, C. E. (2013). Research in Vocabulary: Word power for content-area learning. In J. Ippolito, J. F. Lawrence, & C. Zaller (Eds.), *Adolescent Literacy in the Era of the Common Core: From Research into Practice*. Harvard Education Press.
- Leach, J. M., Scarborough, H. S., & Rescorla, L. (2003). Late-emerging reading disabilities. *Journal of Educational Psychology*, 95(2), 211–224.
- Leech, G. N. (1992). 100 million words of English: The British National Corpus (BNC). *Language Research*, 28(1), 1–13.
- Lesaux, N. K., Kieffer, M. J., Faller, S. E., & Kelley, J. G. (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.
- Li, M., & Kirby, J. R. (2015). The Effects of vocabulary breadth and depth on english reading. *Applied Linguistics*, 36(5), 611–634.
- Lord, F. M. (1980). *Applications of Item Response Theory to Practical Testing Problems*. Routledge.
- Lucas, T., & Katz, A. (1994). Reframing the debate: The Roles of native languages in English-only programs for language minority students. *TESOL Quarterly*, 28(3), 537–561.
- Luce, P. A., & Pisoni, D. B. (1998). Recognizing spoken words: The Neighborhood activation model. *Ear and Hearing*, 19(1), 1–36.

- Ludo 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.
- MacGinitie, W. H., MacGinitie, R. K., Maria, K., Dreyer, L., & Hughes, K. E. (2007). *GMRT manual for scoring and interpretation*. Rolling Meadows, IL: Riverside Publishing.
- MacGinitie, W. H., MacGinitie, R. K., Maria, K., & Dreyer, L. G. (2000). *Gates-MacGinitie reading tests: Directions for administration* (4th ed.). Itasca, IL: Riverside.
- MacWhinney, B. (2000). *The CHILDES Project: The Database*. Psychology Press.
- Mahowald, K., Fedorenko, E., Piantadosi, S. T., & Gibson, E. (2013). Info/information theory: Speakers choose shorter words in predictive contexts. *Cognition, 126*(2), 313–318.
- Malatesha Joshi, R. (2005). Vocabulary: A Critical component of comprehension. *Reading & Writing Quarterly: Overcoming Learning Difficulties, 21*(3), 209–219.
- Marsh, H. W., Liem, G. A. D., Martin, A. J., Morin, A. J. S., & Nagengast, B. (2011). Methodological measurement fruitfulness of exploratory structural equation modeling (ESEM): New approaches to key substantive issues in Motivation and engagement. *Journal of Psychoeducational Assessment, 29*(4), 322–346.
- Marulis, L. M., & Neuman, S. B. (2010). The Effects of vocabulary intervention on young children's word learning: A Meta-analysis. *Review of Educational Research, 80*(3), 300–335.
- Marzano, R. J., & Simms, J. A. (2011). *Vocabulary for the Common Core*. Solution Tree Press.
- Matsumoto, A., Iidaka, T., Haneda, K., Okada, T., & Sadato, N. (2005). Linking semantic priming effect in functional MRI and event-related potentials. *NeuroImage, 24*(3), 624–634.
- 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.
- McKeown, M. G., Crosson, A. C., Moore, D. W., & Beck, I. L. (2018). Word knowledge and comprehension effects of an academic vocabulary intervention for middle school students. *American Educational Research Journal, 55*(3), 572–616.
- McKeown, M. G., Deane, P. D., & Lawless, R. R. (2017). *Vocabulary Assessment to Support Instruction: Building Rich Word-Learning Experiences*. Guilford Publications.
- McNamara, D. S., Graesser, A. C., & Louwerse, M. M. (2012). Sources of text difficulty: Across genres and grades. In J. Sabatini, E. Albro, & T. O'Reilly (Eds.), *Measuring up: Advances in how we assess reading ability* (pp. 89–116). R&L Education.

- McNamara, D. S., & Magliano, J. (2009). Toward a comprehensive model of comprehension. In B. Ross (Ed.), *The Psychology of Learning and Motivation* (Vol. 51, pp. 297–384). New York, NY: Elsevier Science.
- McNeish, D. M., & Stapleton, L. M. (2016). The Effect of small sample size on two-level model estimates: A Review and illustration. *Educational Psychology Review*, 28(2), 295–314.
- Meade, G., Grainger, J., Midgley, K. J., Emmorey, K., & Holcomb, P. J. (2018). From sublexical facilitation to lexical competition: ERP effects of masked neighbor priming. *Brain Research*, 1685, 29–41.
- Meade, G., Midgley, K. J., Dijkstra, T., & Holcomb, P. J. (2018). Cross-language neighborhood effects in learners indicative of an integrated lexicon. *Journal of Cognitive Neuroscience*, 30(1), 70–85.
- Meara, P. (2005). Designing vocabulary tests for English, Spanish and other languages. In C. Butler, M. Gómez-González, & S. Doval Suárez (Eds.), *The Dynamics of Language Use* (pp. 271–286). John Benjamins Publishing Company.
- Meara, P. (2009). *Connected Words: Word Associations and Second Language Vocabulary Acquisition*. John Benjamins Publishing Company.
- Meara, P., & Wolter, B. (2004). V_LINKS: Beyond vocabulary depth. In D. Albrechtsen, K. Haastrup, & B. Henriksen (Eds.), *Angles on the English speaking world 4* (pp. 85–96). Copenhagen, Denmark: Museum Tusulanum Press.
- Melby-Lervåg, M., & Lervåg, A. (2014). Reading comprehension and its underlying components in second-language learners: A meta-analysis of studies comparing first- and second-language learners. *Psychological Bulletin*, 140(2), 409–433.
- Melby-Lervåg, M., Lyster, S.-A. H., & Hulme, C. (2012). Phonological skills and their role in learning to read: A Meta-analytic review. *Psychological Bulletin*, 138(2), 322–352.
- Monsell, S., Doyle, M. C., & Haggard, P. N. (1989). Effects of frequency on visual word recognition tasks: Where are they? *Journal of Experimental Psychology: General*, 118(1), 43–71.
- Myers, J. L., & O'Brien, E. J. (1998). Accessing the discourse representation during reading. *Discourse Processes*, 26(2), 131–157.
- 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). Guilford Publications.
- Nagy, W. E., & Anderson, R. C. (1984). How many words are there in printed school English? *Reading Research Quarterly*, 19(3), 304–330.
- Nagy, W. E., Anderson, R. C., & Herman, P. A. (1987). Learning word meanings from context during normal reading. *American Educational Research Journal*, 24(2), 237–

- 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). Lawrence Erlbaum Associates.
- Nagy, W., & Townsend, D. (2012). Words as tools: Learning academic vocabulary as language acquisition. *Reading Research Quarterly*, 47(1), 91–108.
- Nakayama, M., Lupker, S. J., & Itaguchi, Y. (2018). An examination of L2-L1 noncognate translation priming in the lexical decision task: Insights from distributional and frequency-based analyses. *Bilingualism: Language and Cognition*, 21(2), 265–277.
- Nation, I. S. P. (2001). *Learning Vocabulary in Another Language* (1st ed.). Cambridge University Press.
- Nation, I. S. P. (2004). A study of the most frequent word families in the British National Corpus. *Vocabulary in a Second Language*, 3–13.
- Nation, I. S. P. (2006). How large a vocabulary is needed for reading and listening? *The Canadian Modern Language Review*, 63(1), 59–81.
- Nation, I. S. P. (2022). *Learning Vocabulary in Another Language* (3rd ed.). Cambridge University Press.
- Nation, I. S. P., & Coxhead, A. (2021). *Measuring Native-Speaker Vocabulary Size*. John Benjamins.
- Nation, K., & Snowling, M. (1997). Assessing reading difficulties: The validity and utility of current measures of reading skill. *British Journal of Educational Psychology*, 67, 359–370. doi:10.1111/j.2044-8279.1997.tb01250.x
- Nation, K., & Snowling, M. J. (1999). Developmental differences in sensitivity to semantic relations among good and poor comprehenders: Evidence from semantic priming. *Cognition*, 70(1), B1–B13.
- Nation, K., & Snowling, M. J. (2004). Beyond phonological skills: Broader language skills contribute to the development of reading. *Journal of Research in Reading*, 27(4), 342–356.
- Nation, K., Snowling, M. J., & Clarke, P. (2007). Dissecting the relationship between language skills and learning to read: Semantic and phonological contributions to new vocabulary learning in children with poor reading comprehension. *Advances in Speech Language Pathology*, 9(2), 131–139.
- 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.
- Nippold, M. A., & Sun, L. (2008). Knowledge of morphologically complex words: A Developmental study of older children and young adolescents. *Language, Speech, and*

Hearing Services in Schools, 39(3), 365–373.

- O'Malley, J. M., & Chamot, A. U. (1990). *Learning Strategies in Second Language Acquisition*. Cambridge University Press.
- O'Reilly, T., Wang, Z., & Sabatini, J. (2019). How much knowledge is too little? When a lack of knowledge becomes a barrier to comprehension. *Psychological Science*, 30(9), 1344–1351.
- Ouellette, G., & Beers, A. (2010). A not-so-simple view of reading: How oral vocabulary and visual-word recognition complicate the story. *Reading and Writing*, 23(2), 189–208.
- Ouellette, G. P. (2006). What's meaning got to do with it: The role of vocabulary in word reading and reading comprehension. *Journal of Educational Psychology*, 98(3), 554–566.
- Paradis, J. (2011). Individual differences in child English second language acquisition: Comparing child-internal and child-external factors. *Linguistic Approaches to Bilingualism*, 1(3), 213–237.
- Parks, R., Ray, J., & Bland, S. (1998). *Wordsmyth English Dictionary Thesaurus*. University of Chicago, <http://www.wordsmyth.net/> [Electronic version].
- Pearson, P. D., Hiebert, E. H., & Kamil, M. L. (2007). Vocabulary assessment: What we know and what we need to learn. *Reading Research Quarterly*, 42(2), 282–296.
- Peereman, R., & Content, A. (1997). Orthographic and phonological neighborhoods in naming: Not all neighbors are equally influential in orthographic space. *Journal of Memory & Language*, 37, 382–410.
- Pek, J., Wong, O., & Wong, A. C. (2017). Data transformations for inference with linear regression: Clarifications and recommendations. *Practical Assessment, Research, and Evaluation*, 22(9). <https://doi.org/10.7275/2w3n-0f07>
- Perea, M., & Rosa, E. (2002). The Effects of associative and semantic priming in the lexical decision task. *Psychological Research*, 66(3), 180–194.
- Perfetti, C. (2007). Reading ability: Lexical quality to comprehension. *Scientific Studies of Reading*, 11(4), 357–383.
- Perfetti, C. A. (1999). Comprehending written language: A Blueprint of the reader. In C. M. Brown & P. Hagoort (Eds.), *The Neurocognition of Language* (pp. 167–208). Oxford University Press.
- Perfetti, C., & Adlof, S. M. (2012). Reading comprehension: A Conceptual framework from word meaning to text meaning. In J. Sabatini, E. Albro, & T. O'Reilly (Eds.), *Measuring up: Advances in how we Assess Reading Ability* (pp. 3–20). Rowman & Littlefield Publishers, Inc.
- Perfetti, C., & Hart, L. (2002). The Lexical Quality Hypothesis. In L. Verhoeven, C. Elbro, & P. Reitsma (Eds.), *Precursors of Functional Literacy* (pp. 189–213). John Benajmin's

Publishing.

- Perfetti, C., & Stafura, J. (2014). Word knowledge in a theory of reading comprehension. *Scientific Studies of Reading, 18*(1), 22–37.
- Peters, C. C., & Van Voorhis, W. R. (1940). *Statistical procedures and their mathematical bases*. New York: McGraw-Hill.
- Piantadosi, S. T., Tily, H., & Gibson, E. (2011). Word lengths are optimized for efficient communication. *Proceedings of the National Academy of Sciences of the United States of America, 108*(9), 3526–3529.
- Preston, K. A. (1935). The Speed of word perception and its relation to reading ability. *The Journal of General Psychology, 13*(1), 199–203.
- Prevoo, M. J. L., Malda, M., Mesman, J., & van IJzendoorn, M. H. (2016). Within- and cross-language relations between oral language proficiency and school outcomes in bilingual children with an immigrant background: A Meta-analytical study. *Review of Educational Research, 86*(1), 237–276.
- Qian, D. D. (2002). Investigating the relationship between vocabulary knowledge and academic reading performance: An assessment perspective. *Language Learning, 52*(3), 513–536.
- 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.
- Ravin, Y., & Leacock, C. (2000). *Polysemy: Theoretical and Computational Approaches*. Oxford: Oxford University Press.
- Read, J. (2004). Plumbing the depths: How should the construct of vocabulary knowledge be defined? In P. Bogaards & L. B. (Eds.), *Vocabulary in a Second Language* (pp. 209–227). John Benjamins.
- Read, J., & Dang, T. N. Y. (2022). Measuring depth of academic vocabulary knowledge. *Language Teaching Research, <https://doi.org/10.1177/13621688221105913>*
- Recht, D. R., & Leslie, L. (1988). Effect of prior knowledge on good and poor readers' memory of text. *Journal of Educational Psychology, 80*(1), 16–20.
- Reder, F., Marec-Breton, N., Gombert, J.-E., & Demont, E. (2013). Second-language learners' advantage in metalinguistic awareness: a question of languages' characteristics. *The British Journal of Educational Psychology, 83*(4), 686–702.
- Ringbom, H. (1987). *The Role of L1 in Foreign Language Learning*. Multilingual Matters.
- Ringbom, H. (1992). On L1 transfer in L2 comprehension and L2 production. *Language Learning, 42*(1), 85–112.

- Ringbom, H. (2007). *Crosslinguistic Similarity in Foreign Language Learning*. Multilingual Matters.
- Rissman, J., Eliassen, J. C., & Blumstein, S. E. (2003). An event-related fMRI investigation of implicit semantic priming. *Journal of Cognitive Neuroscience*, *15*(8), 1160–1175.
- 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.
- Rodd, J. M., Gaskell, M. G., & Marslen-Wilson, W. D. (2004). Modelling the effects of semantic ambiguity in word recognition. *Cognitive Science*, *28*(1), 89–104.
- Rossell, S. L., Price, C. J., & Nobre, A. C. (2003). The Anatomy and time course of semantic priming investigated by fMRI and ERPs. *Neuropsychologia*, *41*(5), 550–564.
- Rowe, M. L., Pan, B. A., & Ayoub, C. (2005). Predictors of variation in maternal talk to children: A Longitudinal study of low-income families. *Parenting, Science and Practice*, *5*(3), 259–283.
- Sabatini, J. P., Sawaki, Y., Shore, J. R., & Scarborough, H. S. (2010). Relationships among reading skills of adults with low literacy. *Journal of Learning Disabilities*, *43*(2), 122–138.
- Saunders, W. M., & Marcelletti, D. J. (2013). The Gap that can't go away: The Catch-22 of reclassification in monitoring the progress of English learners. *Educational Evaluation and Policy Analysis*, *35*(2), 139–156.
- Scarborough, H. S. (2001). Connecting Early Language and Literacy to Later Reading (Dis)Abilities: Evidence, Theory, and Practice. In S. B. Neuman & D. K. Dickinson (Eds.), *Handbook of Early Literacy Research* (pp. 97–125). Guilford Press.
- Schatschneider, C., Fletcher, J. M., Francis, D. J., Carlson, C. D., & Foorman, B. R. (2004). Kindergarten prediction of reading skills: A Longitudinal comparative analysis. *Journal of Educational Psychology*, *96*(2), 265–282.
- Schleppegrell, M. J. (2004). *The Language of Schooling: A Functional Linguistics Perspective*. Lawrence Erlbaum Associates.
- Schmitt, N. (2008). Review article: Instructed second language vocabulary learning. *Language Teaching Research*, *12*(3), 329–363.
- Schmitt, N. (2010). *Researching Vocabulary: A Vocabulary Research Manual*. Palgrave Macmillan.
- Schmitt, N. (2014). Size and depth of vocabulary knowledge: What the research shows. *Language Learning*, *64*(4), 913–951.
- Schmitt, N. (2019). Understanding vocabulary acquisition, instruction, and assessment: A research agenda. *Language Teaching*, *52*(2), 261–274.

- Schmitt, N., Cobb, T., Horst, M., & Schmitt, D. (2017). How much vocabulary is needed to use English? Replication of van Zeeland & Schmitt (2012), Nation (2006) and Cobb (2007). *Language Teaching*, 50(2), 212–226.
- 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.
- Schmitt, N., & Zimmerman, C. B. (2002). Derivative word forms: What do learners know? *TESOL Quarterly*, 36(2), 145.
- Schmitt, T. A. (2011). Current methodological considerations in Exploratory and Confirmatory Factor Analysis. *Journal of Psychoeducational Assessment*, 29(4), 304–321.
- Schwanenflugel, P. J., Harnishfeger, K. K., & Stowe, R. W. (1988). Context availability and lexical decisions for abstract and concrete words. *Journal of Memory and Language*, 27(5), 499–520.
- Scott, J. A., Lubliner, S., & Hiebert, E. H. (2006). Constructs underlying word selection and assessment tasks in the archival research on vocabulary instruction. In *55th Yearbook of the National Reading Conference* (pp. 264–275). NRC.
- Sénéchal, M. (2006). Testing the home literacy model: parent involvement in kindergarten is differentially related to grade 4 reading comprehension, fluency, spelling, and reading for pleasure. *Scientific Studies of Reading*, 10(1), 59–87.
- Slama, R. B. (2012). A longitudinal analysis of academic English proficiency outcomes for adolescent English language learners in the United States. *Journal of Educational Psychology*, 104(2), 265–285.
- Snow, C. E., & Kim, Y. (2007). Large problem spaces: The challenge of vocabulary for English language learners. In R. K. Wagner, A. E. Muse, & K. R. Tannenbaum (Eds.), *Vocabulary acquisition: Implications for Reading Comprehension* (pp. 123–139). Guilford.
- Snow, C. E., Lawrence, J. F., & White, C. (2009). Generating knowledge of academic language among urban middle school students. *Journal of Research on Educational Effectiveness*, 2(4), 325–344.
- Snow, C. E., & Uccelli, P. (2009). The Challenge of Academic Language. In D. R. Olson & N. Torrance (Eds.), *The Cambridge Handbook of Literacy* (pp. 112–133). Cambridge University Press.
- Soderstrom, N. C., & Bjork, R. A. (2015). Learning versus performance: An integrative review. *Perspectives on Psychological Science*, 10(2), 176–199.
- Spencer, L. H., & Hanley, J. R. (2003). Effects of orthographic transparency on reading and phoneme awareness in children learning to read in Wales. *British Journal of Psychology*, 94(1), 1–28.

- Spencer, M., Gilmour, A. F., Miller, A. C., Emerson, A. M., Saha, N. M., & Cutting, L. E. (2019). Understanding the influence of text complexity and question type on reading outcomes. *Reading and Writing, 32*(3), 603–637.
- Spencer, M., & Wagner, R. K. (2017). The Comprehension problems for second-language learners with poor reading comprehension despite adequate decoding: A Meta-analysis. *Journal of Research in Reading, 40*(2), 199–217.
- Srinivasan, M., & Snedeker, J. (2011). Judging a book by its cover and its contents: The Representation of polysemous and homophonous meanings in four-year-old children. *Cognitive Psychology, 62*(4), 245–272.
- Stahl, S. A. (2003). Words are learned incrementally over multiple exposures. *American Educator, 27*(1), 18–22.
- 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.
- Stanovich, K. (2000). *Progress in Understanding Reading: Scientific Foundations and New Frontiers*. The Guilford Press.
- Stanovich, K. E. (1980). Toward an interactive-compensatory model of individual differences in the development of reading fluency. *Reading Research Quarterly, 16*(1), 32–71.
- Stanovich, K. E. (1988). Explaining the differences between the dyslexic and the garden-variety poor reader: The Phonological-core variable-difference model. *Journal of Learning Disabilities, 21*(10), 590–604.
- Stemler, S. E., & Naples, A. (2021). Rasch measurement v. Item Response Theory: Knowing when to cross the line. *Practical Assessment, Research, and Evaluation, 26*(11), 1–16.
- Storkel, H. L. (2001). Learning New Words: Phonotactic probability in language development. *Journal of Speech, Language, and Hearing Research, 44*, 1321–1337.
- Storkel, H. L. (2004). Do children acquire dense neighborhoods? An investigation of similarity neighborhoods in lexical acquisition. *Applied Psycholinguistics, 25*(2), 201–221.
- Strange, W., & Shafter, V. L. (2008). Speech perception in second language learners: The re-education of selective perception. In J. G. H. Edwards & M. L. Zampini (Eds.), *Phonology and Second Language Acquisition* (pp. 153–191). John Benjamins Publishing Company.
- Suhr, D. (2006). Exploratory or Confirmatory Factor Analysis. *SAS Users Group International Conference* (pp. 1 - 17). Cary: SAS Institute, Inc.
- Swanborn, M. S. L., & de Glopper, K. (1999). Incidental word learning while reading: A Meta-analysis. *Review of Educational Research, 69*(3), 261–285.
- Swart, N. M., Muijselaar, M. M. L., Steenbeek-Planting, E. G., Droop, M., de Jong, P. F., & Verhoeven, L. (2017). Differential lexical predictors of reading comprehension in fourth

- graders. *Reading and Writing*, 30(3), 489–507.
- 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.
- Townsend, D., & Collins, P. (2009). Academic vocabulary and middle school English learners: An Intervention study. *Reading and Writing*, 22(9), 993–1019.
- 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.
- Tukey, J. W. (1980). We need both exploratory and confirmatory. *The American Statistician*, 34(1), 23–25.
- van Buuren, S. (2007). Multiple imputation of discrete and continuous data by fully conditional specification. *Statistical Methods in Medical Research*, 16(3), 219–242.
- Van Buuren, S., Brand, J. P. L., Groothuis-Oudshoorn, C. G. M., & Rubin, D. B. (2006). Fully conditional specification in multivariate imputation. *Journal of Statistical Computation and Simulation*, 76(12), 1049–1064.
- van Dijk, T. A., & Kintsch, W. (1983). *Strategies of Discourse Comprehension*. Academic Press, Inc.
- van Gelderen, A., Schoonen, R., Stoel, R. D., de Glopper, K., & Hulstijn, J. (2007). Development of adolescent reading comprehension in language 1 and language 2: A Longitudinal analysis of constituent components. *Journal of Educational Psychology*, 99(3), 477–491.
- Veldre, A., & Andrews, S. (2014). Lexical quality and eye movements: Individual differences in the perceptual span of skilled adult readers. *Quarterly Journal of Experimental Psychology*, 67(4), 703–727.
- Verhoeven, L., & van Leeuwe, J. (2008). Prediction of the development of reading comprehension: A Longitudinal study. *Applied Cognitive Psychology*, 22(3), 407–423.
- Wainer, H. (1983). Pyramid power: Searching for an error in test scoring with 830,000 helpers. *The American Statistician*, 37(1), 87–91.
- Walczyk, J. J., Marsiglia, C. S., Johns, A. K., & Bryan, K. S. (2004). Children's compensations for poorly automated reading skills. *Discourse Processes*, 37(1), 47–66.
- Walley, A. C., Metsala, J. L., & Garlock, V. M. (2003). Spoken vocabulary growth: Its role in the development of phoneme awareness and early reading ability. *Reading and Writing*, 16(1), 5–20.
- Wang, H.-C., Nickels, L., Nation, K., & Castles, A. (2013). Predictors of orthographic learning of regular and irregular words. *Scientific Studies of Reading*, 17(5), 369–384.

- Wang, Z., Sabatini, J., O'Reilly, T., & Weeks, J. (2019). Decoding and reading comprehension: A test of the decoding threshold hypothesis. *Journal of Educational Psychology, 111*(3), 387–401.
- Webb, S. (2008). Receptive and productive vocabulary sizes of L2 learners. *Studies in Second Language Acquisition, 30*(1), 79–95.
- Webb, S. A., & Chang, A. C. (2012). Second language vocabulary growth. *RELC Journal, 43*(1), 113–126.
- Webb, S., & Nation, P. (2017). *How Vocabulary is Learned* (Vol. 23). Harvard Education Press.
- Webster, N. (1981). *Webster's Third New International Dictionary of the English Language, Unabridged* (Vol. 1). Merriam-Webster.
- Weisberg, M. (2011). Student attitudes and behaviors towards digital textbooks. *Publishing Research Quarterly, 27*(2), 188–196.
- Wen, Y., & van Heuven, W. J. B. (2017). Non-cognate translation priming in masked priming lexical decision experiments: A meta-analysis. *Psychonomic Bulletin & Review, 24*(3), 879–886.
- West, M. (1953). *A General Service List of English Words*. Longmans, Green and Co.
- White, I. R., Royston, P., & Wood, A. M. (2011). Multiple imputation using chained equations: Issues and guidance for practice. *Statistics in Medicine, 30*(4), 377–399.
- Wible, C. G., Han, S. D., Spencer, M. H., Kubicki, M., Niznikiewicz, M. H., Jolesz, F. A., McCarley, R. W., & Nestor, P. (2006). Connectivity among semantic associates: An fMRI study of semantic priming. *Brain and Language, 97*(3), 294–305.
- Willett, J. B., Singer, J. D., & Martin, N. C. (1998). The design and analysis of longitudinal studies of development and psychopathology in context: Statistical models and methodological recommendations. *Development and Psychopathology, 10*(2), 395–426.
- Wilson, M., & De Boeck, P. (2004). Descriptive and explanatory item response models. In P. De Boeck & M. Wilson (Eds.), *Explanatory Item Response Models: A Generalized Linear and Nonlinear Approach* (pp. 43–74). Springer New York.
- Witzel, N. (2019). Can masked synonym priming replicate masked translation priming? *Quarterly Journal of Experimental Psychology, 72*(10), 2554–2562.
- 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.
- 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.

- 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.
- Youn, H., Sutton, L., Smith, E., Moore, C., Wilkins, J. F., Maddieson, I., Croft, W., & Bhattacharya, T. (2016). On the universal structure of human lexical semantics. *Proceedings of the National Academy of Sciences of the United States of America*, *113*(7), 1766–1771.

Part II

Papers

The Dimensionality of Lexical Features in General, Academic, and Disciplinary Vocabulary

Abstract

Purpose: There are many aspects of words that can influence our lexical processing, and the words we are exposed to influence our opportunities for language and reading development. The purpose of this study is to establish a more comprehensive understanding of the lexical challenges and opportunities students face. **Method:** We explore the latent relationships of word features across three established word lists: the General Service List, Academic Word List, and discipline-specific word lists from the Academic Vocabulary List. We fit exploratory factor models using 22 non-behavioral, empirical measures to three sets of vocabulary words: 2,060 high-frequency words, 1,051 general academic words, and 3,413 domain-specific words. **Results:** We found Frequency, Complexity, Proximity, Polysemy, and Diversity were largely stable factors across the sets of high-frequency and general academic words, but that the challenge facing learners is structurally different for domain-specific words. **Conclusion:** Despite substantial stability, there are important differences in the latent lexical features that learners encounter. We discuss these results and provide our latent factor estimates for words in our sample.

Introduction

Oral and linguistic exposure influences learners' opportunities for verbal and reading development, and advances in research methods have driven an explosion of discrete lexical measures. To date, there have been no attempts to establish the latent dimensions of these lexical characteristics or to understand relationships between dimensions. In this study, we created a comprehensive data set of empirical lexical measures for three well-known word lists, and explored the latent relationships within each. These results allow us to specify the latent factors across groups and their interrelationships for the first time. Frequency, Complexity, Proximity, Polysemy, and Diversity are largely stable factors across the sets of basic and general academic words, but the challenge facing learners is structurally different for domain-specific words. We share our latent estimates so researchers can use them in analyses that explore, or wish to control for, lexical characteristics. In the next section, we review some word-learning processes and lexical features. We then describe related work and the word lists we use, before discussing our research methods.

Word features

The variety of words children encounter shifts as they immerse themselves in age-appropriate language situations or texts and receive tailored linguistic input from caregivers and teachers (Snow, 1972; Hiebert, Goodwin, & Cervetti, 2018). At the same time, the words children learn change predictably (Biemiller & Slonim, 2001). Most monolingual children start talking at around twelve months and experience a vocabulary spurt around 18-24 months (Bates, Bretherton, & Snyder, 1991; Fenson et al., 1994; Goldfield & Reznick, 1990). Young children attend to word families and near neighbors (words that share letters or phonemes with other words) through rhymes and word games, which help develop phonological awareness, leading to better reading acquisition (Kjeldsen, Niemi, & Olofsson, 2003; Bryant

& Goswami, 1987). Most, but certainly not all, words learned in early childhood are phonologically simple.

Children apply the alphabetic system to basic texts with words they already know, although they also encounter rare words in texts even in early grades (Hiebert & Fisher, 2005). Phonological awareness, decoding ability, and morphological parsing skills determine how well students master reading basic words (Bhattacharya & Ehri, 2004; Carlisle, 2000; Singson, Mahony, & Mann, 2000). Hence, word similarities continue to play a role in language development. For example, “face” and “place” are phonologic neighbors, “face” and “fact” are orthographic neighbors, and “face” and “fade” are both (phonographic neighbors). Readers recognize words with many neighbors in a lexical decision task quickly (Laxon, Coltheart, & Keating, 1988), acquire them earlier (Storkel, 2004; 2009), and retain them better (Vitevitch, Storkel, Francisco, Evans, & Goldstein, 2014). The Levenshtein distance (Levenshtein, 1966) measures the similarity of a word to its nearest neighbors by calculating the total number of insertions, deletions, or substitutions necessary to get from one word to another (Yarkoni, Balota, & Yap, 2008). This distance is measured orthographically or phonologically—for example, the orthographic Levenshtein distance (OLD) between “shell” and “tell” is two, but the phonographic Levenshtein distance (PLD) is one. The mean Levenshtein distance between a word and its 20 closest neighbors (OLD20/PLD20) is used to determine neighborhood density; however, previous research has found these are more related to complexity measures than density. For example, in English, short words can be easily transposed to others in the same word family, but complex words tend to have few near neighbors (Yap, Balota, Sibley, & Ratcliff, 2012; Yarkoni et al., 2008).

In upper elementary grades, children learn derivational forms of known words (Anglin, Miller, & Wakefield, 1993), which tend to be multimorphemic and orthographically complex. Children encounter relatively more new words while reading. With each exposure to

a word, a learner can establish a more complete and stable representation of it (Perfetti & Hart, 2002). Since 5th-11th graders have a 15% probability of learning a novel word from an incidental encounter, the likelihood of learning a word correlates with estimates of text exposure (see meta-analysis by Swanborn & de Glopper, 1999). Unsurprisingly, large-scale correlational studies have found a strong relationship between estimated word frequency and when children learn a word. For example, the Living Word Vocabulary study (Dale & O'Rourke, 1981) tested 44,000 individual words with 4th-12th graders on target words to determine when at least 67% of students knew the word. These grade-level estimates of acquisition ratings correlate with frequency estimates from the Brown corpus ($r = -0.690$; see Kuperman, Stadthagen-Gonzalez, & Brysbaert, 2012).

In upper-grade classrooms, school texts tend to incorporate more academic language. Academic language is “able to convey abstract, technical, and nuanced ideas... not typically examined in... social and/or casual conversation” (Nagy & Townsend, 2012). One of the features of general academic words is they tend to be lexically ambiguous. Lexical ambiguity applies when a word has several interpretations or meanings, a common and frequent feature of natural language (Klepousniotou, 2002). Most words in English have etymologically related senses, while relatively few have distinct and etymologically unrelated meanings (Rodd, Gaskell, & Marslen-Wilson, 2004). For example, “bark” has two distinct meanings (dog-bark; tree-bark). Dog-bark has four related senses (dog-bark; noise like dog-bark; making barking sounds; unfriendly tone), and tree-bark has two related senses (wood-bark; covering with bark; Miller, 1990).

The number of meanings and senses a word has influences learning and processing. Sullivan (2007) found even second-grade participants could identify multiple senses of words. Other researchers have found the number of meanings is related to the ease with which a word is learned (Miller & Lee, 1993; Cervetti, Hiebert, Pearson, & McClung, 2015). Studies of

older participants have demonstrated that polysemous words are processed more efficiently (Azuma & Van Orden, 1997; Borowsky & Masson, 1996; Hino & Lupker, 1996), although homophones are processed less efficiently in lexical decision tasks (Beretta, Fiorentine, & Poeppel, 2005; Rodd, Gaskell, & Marslen-Wilson, 2002) and semantic categorization tasks (Hino, Lupker, & Pexman, 2002).

While high school students begin to master higher-frequency general academic words, they are required to focus more on lower-frequency words only useful in specific domains, words such as “mitochondria”. Generally, domain-specific words tend to be less ambiguous and more restrictive in usage across fewer texts. Local (sentence-level) diversity can be measured using latent semantic analysis, which estimates the semantic differences in the contexts where a word appears (Hoffman, Lambon Ralph, & Rogers, 2013). For example, “perjury” usually co-occurs with words like “witness,” while “predicament” has a similar overall frequency but appears next to a broader set of words. Global (document-level) diversity can be measured with contextual diversity. For example, Adelman, Brown, and Quesada (2006) counted the number of documents where each word appeared in the British National Corpus. They found that “HIV” and “lively” have similar total frequency; however, “HIV” is concentrated in a few texts, whereas “lively” appears sparsely across many documents (Leech & Rayson, 2014). Nevertheless, contextual diversity still counts word occurrences and correlates highly with frequency (Brysbaert, Mandera, McCormick, & Keuleers, 2019).

Reading comprehension is determined, at a minimum, by student skill and the text under consideration. Examining the relationships between lexical features of the language encountered in different contexts can help us understand the diverse linguistic challenges we face and advance our understanding of language and reading development.

Relationships between dimensions

Four previous studies have modeled English linguistic features into dimensions, although none made the models an explicit focus in their study. Paivio (1968) examined a set of 96 nouns for experiments on associative reaction times and learning. Clark and Paivio (2004) then expanded to 925 selected nouns with non-behavioral measures, e.g., the number of letters, meanings, and new word frequency measures. Brysbaert et al. (2019) examined the same 925 nouns against 51 word features, including the orthographic and phonological Levenshtein distances. Finally, Yap et al. (2012) used 28,803 words from the English Lexicon Project to reduce their ten lexical variables into broader components.

Across these studies, three factors remained relatively stable: Frequency, Complexity, and Proximity. Yap et al. (2012) found that the number of letters, syllables, morphemes, and Levenshtein distances formed “Structural Properties”. Clark and Paivio (2004) found that the number of letters and syllables and the mean rating for the number of rhyming words, similar-looking words, ease of pronounceability, and age of acquisition formed the “Length” factor. Further, Brysbaert et al. (2019) modeled “Similarity” as the number of rhyming words, the number of words with the same initial letters, neighborhood sizes, and the Levenshtein distances; while Yap et al. (2012) only included the orthographic and phonological neighborhood sizes.

None of these studies discussed how measures fit within the model. Most models also did not allow factors to correlate, despite current recommendations that factor analyses should, by default, not restrict factors to be uncorrelated (Loewen & Gonulal, 2015; Field, 2013). The strong relationship between Complexity and Proximity was still apparent, as variables tended to cross-load onto both factors, providing further evidence for the need for oblique rotation. Previous models included some behavioral measures and ratings, which depend on the participants who created the ratings, such as introductory psychology students,

and can be influenced by non-behavioral measures in ways we find difficult to measure or do not currently understand. None of these studies systematically sampled list of words purposely to understand latent dimensions and relations.

Word lists

Linguists and researchers have created word lists using corpus linguistics to help educators and interventionists target instructional words, and help researchers more easily identify words that may be of particular interest to different profiles of learners. Many such lists are created with specialized corpora, using increasingly sophisticated methods. We wanted to extend what is known about the relationships between lexical dimensions and so identified lists that were sufficiently unique from each other, clearly documented, and well-used in the research community.

The General Service List (GSL) identifies 2,000 high-frequency headwords and derivations from analyzing five million running words (West, 1953). Learners who have mastered only these words can expect approximately 80% coverage of written English (DeRocher, 1973). Words range from high-frequency words like “one” to less frequent words like “congratulations.” The GSL has been cited more than 3,000 times.

The Academic Word list (AWL) is derived from an analysis of a 3.5-million-word corpus containing over 400 texts categorized as Arts, Commerce, Law, and Science (Coxhead, 2000). The AWL excludes the GSL words and those words that occurred less than 100 times in the corpus; the resulting academic words in this list are in the middle range of frequency. Coxhead also 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 comparison bands of words based on frequency alone. As a result, this list has been referenced in influential instructional texts (Beck, McKeown, & Kucan, 2002), used in the creation of vocabulary interventions for middle school students

(Lawrence, Francis, Paré-Blagoev, & Snow, 2016; Lesaux, Kieffer, Kelley, & Harris, 2014), and cited more than 5,000 times.

The new Academic Vocabulary List (Gardner & Davies, 2014) is derived from the 125-million-word sub-corpus for the Corpus of Contemporary American English (Davies, 2012). The entire list includes 8,300 words. Each word occurs more than three times the expected frequency in at least one of nine disciplines, but not more than three: Education, Humanities, History, Social Science, Philosophy/Religion/Psychology, Law/Political Science, Science/Technology, Medicine/Health, or Business/Finance. This corpus has been cited nearly 900 times.

The need for latent estimates

There are distinct advantages to using latent estimates of word characteristics. Grouping word features can alleviate multicollinearity, which can “cause regression coefficients to fluctuate in magnitude and direction, leading to estimates of individual regression coefficients that are unreliable due to large standard errors” (Yap et al., 2012, p. 60). Groupings can also reduce data requirements for advanced modeling, increase statistical power, and improve clarity. Future researchers can also use groupings based on non-behavioral data to explore the relationship with behavioral measures, such as reaction time, age of acquisition, or item difficulty, at the word- or item-level. Similarly, researchers can rely on latent estimates to select equivalent stimuli across many dimensions instead of relying on a single measure. Estimates for each word are available for non-commercial use at <http://www.xxx.edu/>.

Research questions

To date, no one has systematically explored relationships across lexical dimensions in different sets of words to better articulate learners' linguistic environments and challenges. We believe establishing a more comprehensive and credible understanding of the differences

in the challenges and opportunities students face is essential to advancing our scientific knowledge of language and reading development. Therefore, our research questions are:

1. What are the factor structures for the lexical characteristics of words in the General Service List, Academic Word List, and Domain-Specific Academic Vocabulary List?
2. How do these different factor spaces compare to one another?

Methods

We compiled a list of possible word features and extracted data across all possible letter strings. We removed non-relevant letter strings and words with missing data. We then conducted exploratory factor analyses using maximum likelihood with oblique rotations for three different word samples: basic, general academic, and domain-specific. We repeated the analysis for each word sample so models could differ, if appropriate.

Sample

We sampled words from three existing word lists that others have created with explicit documentation and used widely in research: the General Service List (GSL; West, 1953), Academic Word List (AWL; Coxhead, 2000) and the Domain-Specific subset of the Academic Vocabulary List (AVL-DS; Gardner & Davies, 2014). To create each sample of orthographically unique letter strings, we included headwords, lemmas, and derivations (e.g. “die” includes “dies” and “died”) explicitly provided by the original authors (for the GSL and AVL-DS) or in the Oxford American Dictionary (for the AWL). As a result, our sample included 2,284 orthographically unique letter strings for the GSL, 2,958 for the AWL, and 8,300 for the AVL-DS.

Measures

We included all word features from the four previous factor analyses and searched for additional word features in peer-reviewed articles citing either Brysbaert et al. (2019) or Yap

et al. (2012). We then excluded any feature with data for less than 1,000 words, based on human ratings or behavioral measures, and any feature measured before 1950 or after 2020. We recognize this list is not exhaustive; however, we believe it covers a diverse, representative, and systematic sample of possible word features available at the time of publication. We next describe each word feature in alphabetical order.

cd (contextual diversity) is the number of documents in which a word appears (Adelman et al., 2006) in the TASA corpus (Touchstone Applied Science Associates), containing approximately 120,000 paragraphs taken from 38,000 academic texts.

cocazipf is the Zipfian-transformedⁱ word frequenciesⁱⁱ from the Corpus of Contemporary American English (COCA; Davies, 2008), containing approximately 560 million words from T.V., radio, newspapers, fiction, academic papers, and popular magazines.

d (dispersion) is the number of subject areas in which a word appears in The Educator's Word Frequency Guide (Zeno, Ivens, Millard, & Duvvuri, 1995).

freqband is the frequency groupingⁱⁱⁱ from the Oxford English Dictionary (OED) based on the raw frequencies from Google Ngrams version 2 (Lin et al., 2012).

length is the number of letters in the word.

log_freq_hal is the log-transformed word frequencies from the HAL corpus (Hyperspace Analogue to Language; Lund & Burgess, 1996), containing approximately 131 million words from 3,000 Usenet newsgroups; collected from the English Lexicon Project website (Balota, Yap, Hutchison, Cortese, & Klessner, et al., 2007).

log_freq_kf is the log-transformed word frequencies from the Brown corpus (Kučera & Francis, 1967), containing approximately 1 million words from American English texts; collected from the English Lexicon Project website.

nmorph is the number of morphemes in the word.

nphon is the number of phonemes in the word.

nsyll is the number of syllables in the word.

og_n is the raw number of phonographic neighbors (i.e., the number of words that are one letter *and* one phoneme away from the word, e.g., “stove” and “stone”), excluding homophones.^{iv}

old20 is the mean Levenshtein distance of the 20 closest orthographic neighbors (Yarkoni et al., 2008).

ortho_n is the raw number of orthographic neighbors (i.e., the number of words that are one letter away from the word, e.g., “lost” and “lose”), excluding homophones.

phono_n is the raw number of phonologic neighbors (i.e., the number of words that are one phoneme away from the word, e.g., “hear” and “hare”), excluding homophones.

pld20 is the mean Levenshtein distance of the 20 closest phonographic neighbors (Yarkoni et al., 2008).

semd (semantic diversity) is the mean cosine of the latent semantic analysis vectors for all pairwise combinations of contexts containing the word (Hoffman et al., 2013). Information comes from the British National Corpus, containing approximately 100 million words from T.V., radio, newspapers, fiction, academic papers, and popular magazines.

subzipf refers to the Zipfian-transformed word frequencies from the SubtlexUS corpus (Subtitle Lexicon- U.S. version; Brysbaert & New, 2009), containing approximately 51 million words from American subtitles.

wordage is the number of years^v since a word was first used (as of 2000), as reported by Google Ngram, based on 450 million words scanned from Google Books (Lin et al., 2012).

wordnet_lnapossam is the log-transformed number of senses and meanings a word has across all possible parts of speech scraped from the WordNet lexical database (Miller, 1990).

wordsmyth_Inapossam is the log-transformed number of senses and meanings a word has across all possible parts of speech scraped from the Wordsmyth integrated dictionary and thesaurus, compiled of 50,000 headwords (Parks, Ray, & Bland, 1998).

z_sem_prec is the z-transformed depth score^{vi} scraped from WordNet (Fellbaum, 2005). Words with multiple definitions received multiple scores, which were averaged

zenozipf is the Zipfian-transformed word frequency from The Educator's Word Frequency Guide (Zeno et al., 1995), containing 17 million words from kindergarten- to college-level texts.

Data merging and cleaning

We collected data for all possible strings of letters, regardless of type (e.g., lemma, inflection, derivative, abbreviation, suffix, etc.). To combine datasets from varying sources, we merged datasets and collapsed measures that differed between parts of speech into a single entry per word (see above footnotes). We then merged onto datasets without part of speech for a total of 407,510 unique letter strings. Last, we omitted all entries without complete data on all twenty-two measures.^{vii} This process eliminated nonwords (e.g., “2-Feb,” “-ed,” “NASA,”) but also valid words with missing data.

The entire process reduced the dataset from 407,510 unique letter strings to 10,744 words with complete data. We retained 2,060 (90.19%) basic, 1,051 (35.53%) general academic, 3,413 domain-specific (41.12%), and 4,978 words not present in any of the three samples; many words overlapped between samples (see Figure 1). For example, “medical” appears in all three samples, 774 words appear in at least two, and 5,267 appear in only one.

Analyses

To determine the factor structure for word characteristics from different word samples (i.e., RQ1), we conducted separate maximum likelihood EFAs with each word sample as the reference. Each model factored the correlation matrix using only words with complete data

from the relevant sample and maximum likelihood estimation of the factors, along with a direct oblimin rotation via the *psych* package for R (Revelle, 2020). The final models met multivariate assumptions, correlational matrix adequacy, and sampling adequacy. We computed factor scores for all words based on each model to address how different factor spaces compare (i.e., RQ2), then examined the distributions of factor scores for the different populations of words when scored according to the three different reference spaces.

Results

Table 1 includes descriptive information about word features from each sample, with features in alphabetical order and word lists moving from basic to discipline-specific. For example, the fifth row shows that the average length of basic words is 5.84 letters, but for general academic words is 8.57 and 7.31 for domain-specific words. The 16th row shows that basic and general academic words are semantically dispersed (mean *semd*=1.80 and 1.79, respectively), but domain-specific words are an entire standard deviation less dispersed (mean *semd*=1.44, SD=0.30).

RQ1. Factor structure for GSL, AWL, and AVL-DS words

Model fit

We considered five methods for determining the number of factors for each word sample using the *nFactors* (Raiche, 2010) and *psych* (Revelle, 2020) packages in R, which consistently suggested four- or five-factor solutions, which we assessed for all samples (Table 2). We discuss the final solutions here.

For the GSL, the four-factor model fit was poor and combined the Frequency and Diversity factors, making the five-factor model preferable. The model had overall good fit, with the RMSEA indicating moderate fit (.083), the RMSR indicating excellent fit (.02), and the CFI and TLI also indicating excellent fit (.962 and .934, respectively; Table 2), and explained 75% of the variance in word features.

The four-factor model had poor fit and combined Frequency and Diversity factors for the AVL, also. The five-factor model had overall good fit, with the RMSEA indicating moderate fit (.088), the RMSR indicating excellent fit (.02), and the CFI and TLI also indicating excellent fit (.948 and .907, respectively; Table 2), and explained 69% of the variance.

For the AVL-DS, the five-factor model had good fit but contained a factor with *pld* alone. The four-factor model still had overall good fit, with the RMSEA indicating moderate fit (.092), the RMSR indicating excellent fit at (.03), and the CFI and TLI also indicating good fit (.933 and .896, respectively; Table 2), and explained 67% of the variance.

Factor loadings

Table 3 contains standardized factor loadings for the final model of each word list. Measures are in order of factor loadings on the GSL-reference model so that groupings are easier to see. For example, the COCA frequency had the strongest loading on the Frequency factor for all word lists. Table 3 also shows each factor's explained variance, eigenvalue, and standardized α within the model for the specified word list.

For the GSL-reference model, the latent factor Frequency included all word frequency measures in the diverse corpora (the COCA, HAL, Educator's Word Frequency Guide, Brown, Oxford English Dictionary, and Subtlex) with reasonably high loadings (from .99 for the COCA to .75 for the frequency band). However, Frequency also included contextual diversity and word age—albeit at lower loadings (.61 and .30, respectively). Frequency had a large eigenvalue (5.65), high reliability ($\alpha=.94$), and explained 26% of the variance in word features.

For the AVL-reference model, Frequency also explained the most variance (24%, eigenvalue=5.24) and was also highly reliable ($\alpha=.93$). It included all word frequency measures in diverse corpora, with loadings ranging from .99 for the COCA to .61 for Subtlex-

US. Two other measures loaded onto the Frequency factor: contextual diversity and dispersion, although relatively weakly (.61 to .32).

Frequency also explained the most variance for the discipline-specific-reference model (24%, eigenvalue=5.21, α =.93). COCA frequency was again the highest-loading factor (0.98), followed by frequency in the Brown and HAL corpora, Educator's Word Frequency Guide, and frequency band (0.76-0.84). The lowest loadings were for frequency based on the Subtlex corpus, contextual diversity, and dispersion (0.52-0.76).

The second latent factor, Complexity, measured various linguistic elements such as the number of letters, syllables, morphemes, and phonemes, as well as Levenshtein distances. It exhibited high loadings ranging from .98-.72, with phonemes and old20 having the highest and lowest loadings, respectively. Additionally, Complexity had a high eigenvalue and reliability coefficient (4.91; α =.96), explaining 22% of the variance. Similar results were observed for the AWL-reference model, with Complexity being a strong and reliable factor that explained 21% of the variance (eigenvalue=4.65, α =.95). Letters and phonemes had the strongest loadings (.96 and .97, respectively), while the number of morphemes, syllables, and Levenshtein distances had relatively lower—but still strong—loadings (.69-.86). Similarly, Complexity explained 22% of the variance for the AVL-DS-reference model and was the most internally-stable factor (eigenvalue=4.93, α =.96). The number of letters, syllables, and phonemes were the strongest loading measures (.90-.98), followed by phonologic Levenshtein distance (.88). Orthographic Levenshtein distance and the number of morphemes had weaker loadings at .77.

Factor 3, Proximity, included the size of orthographic, phonologic, and phonographic neighborhoods. This factor contained high loadings, ranging from .96-.64 (orthographic versus phonographic neighborhood, respectively) for the GSL-reference model. However, the reliability (α =.93), eigenvalue (2.55), and explained variance (12%) were lower than the

previous two factors. Findings for both the AWL-reference and AVL-DS-reference models were similar: Proximity explained 12% of the variance in word characteristics and had a reliability of .94-.95, respectively. However, Proximity loadings were also high: .99 for phonographic, .95 for orthographic, and .78 for the phonologic neighborhood size, for the AWL; .98 for orthographic and phonographic, and .73 for the phonologic neighborhood for the AVL-DS.

Factor 4, Polysemy, included the two measures of senses and meanings from WordNet (loading=.96) and Wordsmyth (loading=.79). Even with only two items, this factor retained acceptable reliability ($\alpha=.90$), while the eigenvalue (1.72) and explained variance (8%) were lower than previous factors. For the AWL-reference model, Polysemy also had reduced explained variance (7%, eigenvalue=1.49) and reliability ($\alpha=.79$). The loading for the number of senses and meanings from WordNet was stronger than the loading based on Wordsmyth (.87 and .68, respectively).

The latent factor Polysemy is a mix of polysemy and diversity measures for the discipline-specific model. Polysemy/Diversity explained the least amount of variance and was less reliable than previous factors (9%, eigenvalue=1.93, $\alpha=.75$). The number of senses and meanings from various dictionaries loaded strongest (WordNet and Wordsmyth, at .90 and .70, respectively), followed by a relatively weaker loading for semantic diversity (.39).

The GSL- and AWL-reference models included semantic dispersion and precision as Diversity. Loadings were more varied on this factor (0.98 for semantic dispersion to -0.42 for semantic precision), and the eigenvalue (1.59), explained variance (7%), and reliability ($\alpha=.68$) were considerably lower than for other factors on the GSL-reference model. For the AWL-reference model, Diversity and Polysemy explained a similar amount of variance (6% vs. 7%, eigenvalue=1.31) but with lower reliability ($\alpha=.62$ vs. .79); and included semantic diversity, precision, and word age, along with the cross-loaded dispersion.

RQ2. Comparison of factor spaces

To compare the different factor spaces to one another, we examined the correlations among factors and the distributions of factor scores by scoring the words in each reference sample using the factor score regressions from the three separate analyses. To examine factor correlations and densities, we present scatterplot matrices in Figures 2-4 for GSL-, AWL-, and AVL-DS-reference models. For example, Figure 2 is based on the model created by analyzing only GSL words but includes red plots for estimated factor scores on AVL-DS words based on the GSL-reference model. Figure 2 plots density curves for each word sample on the diagonal, along with the factor correlation above the diagonal and a scatterplot below the diagonal. Across Figures 2-4, red plots consistently show the estimated factor scores for AVL-DS words, green plots show AWL words, blue plots show GSL words, and purple plots show all words in any list. These estimates change across figures because the scoring coefficients differ depending on the reference sample used in the analysis.

Correlations between factors

In the GSL-reference model, significant correlations ($p < .001$) were observed among factors, as shown in Figure 2. The strongest negative correlation was between Complexity and Proximity (-.67), indicating that more complex words had fewer neighboring words. Frequency and Diversity were positively correlated at .54, suggesting that frequently used words appear in various contexts. The mid-range correlations (ranging from -.47 to .41) were all related to Polysemy, indicating that more complex words tend to have fewer meanings and that words with more meanings tend to be used more frequently and have more neighbors. Polysemy and Diversity had a weak but still significant correlation at .30. Frequency showed the weakest correlations, with Complexity being negatively correlated at -.28 and Proximity positively correlated at .20.

Comparing the estimated correlations using the AWL and AVL-DS words and the GSL-reference model, there are a few apparent differences across reference word lists. The

most striking finding is that the estimated correlations among factors are generally larger when based on all words across all lists, except for the correlation between Proximity and Complexity. The next striking finding is that correlations are usually somewhat weaker when calculated from AWL-sample estimates compared to GSL-sample or AVL-DS-sample estimates against the GSL-reference model.

Similar to the GSL-reference model, nearly all correlations between factors were significant at $p < .001$ for the AWL-reference model (Figure 3). Complexity and Proximity again correlated the strongest (-.48), indicating that less complex words tend to have more words in their neighborhood. Frequency and Polysemy then correlated moderately at .47, indicating that words used more frequently have more meanings. Frequency also correlated moderately with Diversity (.45) and with Complexity (-.39). Polysemy correlated moderately with Complexity (-.42) and weakly with Proximity (.25). The weakest correlations included Frequency and Proximity (.21), Diversity and Polysemy (.20), and Diversity with Complexity (-.16). Our previous observation that correlation estimates based on the AWL sample tend to be somewhat weaker than GSL- and AVL-DS-sample estimates mostly hold for the AWL-reference model.

For the AVL-DS-reference model, all correlations between factors were significant at $p < .001$ (Figure 4). Frequency and Polysemy/Diversity correlated the strongest, closely followed by Complexity and Proximity (.63 and -.62, respectively). The Polysemy/Diversity factor then correlated moderately with Complexity and Proximity (-.47 and .45, respectively). The weakest correlations were still quite strong for Frequency with Proximity and Frequency with Complexity (.39 and -.36, respectively). Comparing the estimates for different word lists using the AVL-DS-reference, we again see that correlations are somewhat weaker for the AWL-sample estimates and tend to be strongest for the entire word sample estimates. The consistency of the latter finding across all three scoring models suggests that the three specific

word lists somewhat restrict the range of observations, such that when the restriction of range is removed, correlations are stronger.

Comparing the correlations across the separate analyses reveals the correlations were reasonably consistent. Although exploratory and descriptive, these comparisons are consistent with the notion that the estimated factors are the same, regardless of which word list is the reference. That is, the characteristics of words seem to define a common set of dimensions regardless of the reference word list used to define the space. What changes between analyses is the reference space and the distribution of factor scores within that reference space, but not the factors themselves.

Comparing factor score estimates across models and samples

As mentioned above, we estimated factor scores for each word sample (and all word samples together) based on separate models for each target population. Thus, Figures 2–4 also compare the factor score distributions in the different reference word lists and across all words. For example, we can see from the density plots for Frequency in Figure 2 that general academic words (AWL) are less frequent because the green Frequency density plot is further to the left than the blue (GSL). The same holds for the density plots in Figures 3 and 4, where general academic words (AWL) and domain-specific words (AVL-DS) serve as the model reference.

Figure 5 displays scaled density plots for each factor across the three scoring models and four word samples (GSL, AWL, AVL-DS, All Words). Each row represents one scoring model, while each column is the density plot for a specific factor. The color of the density plot still indexes the word sample used for estimation. Hence, the first row uses the GSL-reference model to estimate factor scores, while the first column shows the distribution of the Frequency factor across all three reference models. Thus, the first cell shows the distribution of scores for the Frequency factor using the GSL-reference model. The least frequent words

are the domain-specific words (red), then general academic words (green), and finally, basic words (blue), with the entire range represented in purple. The choice of reference model has a negligible impact on the factor distribution; what matters is which word *sample* is used to estimate the distribution. We reach the same conclusions regardless of model examined, except for Polysemy/Diversity, which is one factor in the AVL-DS scoring model and separate factors in the GSL and AWL models. We host animations of the scaled density plots to demonstrate how the different samples of words compare across various models at XXX. These animations show more clearly the slight variations induced by shifting the reference distribution for a factor as the scoring model shifts from one reference sample to another.

Discussion

To some extent, learners' language contexts define the skills they need to develop and the opportunities to do so. Yet, few studies systematically parameterize the latent features of the diverse language environments that learners experience. This study focused on words as one critical language unit and asked what exploratory factor structures emerge for a systematic collection of word features and how those structures differ across purposely selected word lists. We searched the literature for empirical measures of words and included non-behavioral measures after 1950 for more than 1,000 words. We combined all the word features into a large dataset of 22 measures and 10,744 unique words with complete data and conducted analyses on data from three different word lists. We found that English word features grouped into a similar five-factor structure regardless of word list: Frequency, Complexity, Proximity, Polysemy, and Diversity, although the emerging factor structure for domain-specific words combined Polysemy and Diversity into a single factor. While we cannot explicitly test the equivalence of factor structures using the current exploratory factor analytic methods, we were able to compare our three models descriptively. The differences between factor structures were minor, suggesting that word features identify the same latent

dimensions regardless of the reference word list. Below we discuss the similarities and differences in the factor structures, the implications for this work, and its limitations.

Comparing models with different reference samples

Analyses revealed some factors were stable in all models while others were less so.

Universal word factors

Frequency. The latent Frequency construct describes a word's occurrence rate and is considered a proxy for relative exposure level; words that are more frequent in text and speech are more likely to be encountered more often. Since encounters with words provide opportunities to learn them, it is unsurprising that frequency has been a significant predictor of which words children know (Goodman, Dale, & Li, 2008; Swanborn & De Glopper, 1999), the efficiency with which learners process words (Brysbaert, Mandera, & Keuleers, 2018; Monsell, Doyle, & Haggard, 1989), and how well learners know a word (Ellis, 2002).

The obtained Frequency factor in our study includes all word frequency measures for each reference model. Corpus frequency measures are highly correlated (Breland, 1996), and all estimate how frequently a word is used by counting occurrences in corpora from different sources. Our latent factor incorporates frequency scores from various corpora and is thus more representative than a frequency measure derived from any single corpus. As a result, researchers not interested in word frequency in specific modalities or formats may wish to use our factor scores that account for word frequency across modalities and corpora.

The Frequency factor also includes a few measures that do not directly measure raw frequency in a corpus: contextual diversity, dispersion, and word age. That being said, contextual diversity and dispersion *do* measure frequency at a larger grain size. Adelman et al. (2006) operationalize contextual diversity as the number of *documents* in which a word appears in a corpus. Zeno et al. (1995) operationalize dispersion as the number of *content areas* in which a word appears in a corpus. Thus, both are corpus-derived frequency

measures, and our results suggest that these measures reflect a latent Frequency dimension. Researchers intending to control for frequency effects might want to consider using our latent score that accounts for these related measures rather than only raw frequency measures.

Raw frequency from the Corpus of Contemporary Academic English (COCA) was the strongest-loading measure on our latent Frequency factor. We had expected that frequency measures based on conversational corpora (e.g., Subtlex) would be stronger for basic words. However, the COCA contains almost ten times as many words as Subtlex; our results highlight the large corpora's dominating utility.

Complexity. The obtained Complexity factor relates to the orthographic and phonological difficulty of a word, which relates to the ease or difficulty of learning to say (Ehri, 2014), read (Carlisle, 2000), or process (Ehri, 2005) a word. Words that take longer to process or are difficult to decode tend to make reading more challenging (Ehri, 1992; Carlisle, 2000). On the other hand, information theory supports that longer words are more likely to contain more meaningful information than shorter words (Piantadosi, Tily, & Gibson, 2011; Mahowald, Fedorenko, Piantadosi, & Gibson, 2013). For example, “unbreakable” has three pieces of information: “un-break-able,” making it a denser and abstract word than “break” alone. The measures that load on the Complexity factor describe these different but related ways a word could be challenging to decode, encode, and process. For example, a word can be difficult to process due to a complex orthography or phonology, but these do not correlate perfectly (e.g., “cough”).

Levenshtein distances also loaded onto the Complexity factor. We expected these to load onto Proximity, yet, scores on neighborhood sizes (the Proximity factor) and Levenshtein distances vary systematically but not *linearly*. For example, words with a score of one on the old20 measure can have anywhere between 11-35 close neighbors (words that are exactly one change away from the original word). Furthermore, there is a considerable *variation* in

neighborhood size when Levenshtein distances are small but minimal variation when Levenshtein distances are large. Thus, it is unsurprising that other researchers found the Levenshtein distances to load onto Complexity.

Measures on Complexity are fixed-analytic computations and corpus-free (i.e., the number of letters or syllables in a word is the same regardless of where you read it, with dialectic exceptions). Our latent Complexity factor gives researchers a measure of orthographic and phonological complexity that accounts for information from related measures while mitigating concerns about multicollinearity

Proximity. The latent Proximity construct measures how many words are closely related to this word visually and aurally. Words in dense neighborhoods tend to be learned earlier, especially as we engage in phonological awareness training at a relatively young age. We recognize words with many neighbors more quickly, although which type of neighborhood (phonological or orthographical) is most useful is still debated (Adelman & Brown, 2007). Further, neighborhood size could be the driving factor behind the word length effect on recall (Jalbert, Neath, & Surprenant, 2011).

The Proximity factor included measurements of neighborhood size. The phonographic neighbors were consistently the strongest-loading measure. The clear distinction of the Proximity factor from other factors stems from the shape of the distributions of the three measures of neighborhood size. These distributions are highly positively skewed, with a large concentration around zero. Many multi-syllable words cannot transpose into any other word with only one change, such as “straightforward,” while few words reside in large neighborhoods, such as “cat,” with 32 phonologic neighbors. Similar to Complexity, Proximity contains distinct but highly related measures. Consequently, using the factor scores broadly represents a word’s proximity to other words while mitigating concerns about multicollinearity from using multiple measures.

Consistent word factors

The remaining two factors were distinct and weakly correlated for the basic- and general academic-reference models but combined into a single factor for the domain-specific model. We label these factors consistent as they were similarly defined across the different reference spaces, although consolidated into a single dimension for domain-specific words.

Polysemy. The latent Polysemy construct relates to how many distinct meanings and related senses a word has. Polysemy is an essential feature of all languages and there seem to be similarities in how different languages extend the senses of words to related concepts (Youn, Sutton, Smith, Moore, Wilkins, et al., 2016). Words with many related senses are processed more efficiently (Azuma & Van Orden, 1997; Borowsky & Masson, 1996; Hino & Lupker, 1996); second language learners may not enjoy the same advantages in learning polysemous words as their peers do. Some words have alternative senses that can be used in a wide variety of documents or contexts (“grasp” a cup or “grasp” an idea). Other words have senses that are more constrained by the document or disciplinary genre (jail “cell” versus biological “cell.”)

Diversity. The latent Diversity construct describes the number of contexts in which a word can be used and encompasses global and local contexts. The global context is at the discipline or document level, such as contextual diversity, which counts the number of documents among a large corpus in which a word occurs. When a word is used in more documents or contexts, it may provide more learning opportunities, which explains why the contextual diversity measure explains lexical processing efficiency so well (Adelman et al., 2006; Jones, Johns, & Recchia, 2012). At a more global level, the documents that include a word can be categorized by academic discipline resulting in a variable called dispersion (Zeno et al., 1995). Our latent factor accounts for both these measures and a measure of diversity at the sentence (local) level. Semantic diversity considers the words used next to or near a target

word across documents in a corpus. The relationship between Diversity and Polysemy is easy to understand when considering that a word with more meanings can usefully be employed in more diverse contexts. Semantic precision also relates to diversity (negatively) as it describes how far down a word is down a hypernym chain (Fellbaum, 2005).

Considerations for domain-specific words

Findings were slightly different for domain-specific words. The criterion used to identify domain-specific words ensured that these words are used in a limited number of contexts. As document-level variability for domain-specific words is constrained, so is the utility of global diversity measures such as contextual diversity or dispersion, which measure use across documents and disciplines. Conversely, word features that measure local variability relate to the number of senses and meanings a word has and thus loads onto the Polysemy/Diversity factor, as shown in Table 3. Instead of one factor for global/local diversity and one for polysemy, we also found that global diversity measures loaded with Frequency, and the local diversity measure loaded with Polysemy in the analysis of domain-specific words. Given these constraints, it is sensible that global diversity measures are related to the overall frequency of the word, as we see in Table 3: dispersion and contextual diversity load onto the Frequency factor.

Factor correlations

Our models used oblique rotations so that factors could correlate if appropriate. The correlations between factors generally followed the same direction and level of statistical significance for all models. However, the magnitude varied somewhat across word lists, possibly partly due to sampling variability and parameter differences. Complex words consistently had fewer neighbors; frequent words were used more diversely and had more senses and meanings, regardless of word set. Diverse words had little relation with neighborhood size or complexity.

The correlation pattern between the basic and general academic word models was similar (Figures 2 and 3). Nearly all factors correlated statistically significantly at $p < .001$, suggesting that the oblique rotation was necessary. Moreover, magnitudes ranged from .16 to .67, emphasizing that selecting five factors was suitable.

The relationships between factors remained stable across reference models, despite domain-specific words collapsing into four factors. One notable difference was a nonsignificant relationship between Diversity and Complexity for basic words, although still positive. This may be because the words sampled for basic words are less complex than the general academic and domain-specific words. Proximity and Complexity also correlated more weakly for general academic words than for others.

Still, the stability of the relationships between factors across models is noteworthy. For example, the correlation between Proximity and Complexity is consistently either the strongest or second strongest correlation. The correlation between Frequency and the Polysemy/Diversity combination was the other strongest correlation for all domain-specific models; however, correlations of Frequency with separate Polysemy and Diversity were moderate for both basic- and academic-reference models.

Previous data-driven models

Our work advances the field beyond previous studies in two ways. First, we included words from all parts of speech. Secondly, we excluded measures based on human ratings and behaviors. Third, our statistical models allowed factors to correlate, thereby reducing mathematical constraints that are not driven by linguistic data. Despite these differences, our findings were generally similar to those of prior authors. For example, Clark & Paivio's (2004) model with 925 nouns also shows word frequency measures loading onto a Frequency factor and the number of letters and syllables loading onto a Complexity-like factor. Although Clark and Paivio (2004) restricted the models to uncorrelated factors, they acknowledged the

issue of cross-loading, “implicating a multi-dimensional underlying structure for these variables” (p. 376). Similarly, Yap et al. (2012) used principal components analysis on ten measures included in our models. In this analysis, the Length and Neighborhood components are identical to our Complexity and Proximity factors, while the Frequency/Semantic component contained one measure from our Frequency, Polysemy, and Diversity factors each.

Brysbaert et al.’s (2019) model is arguably most aligned with our models. This model included the most measures in our model and an oblique rotation. We found this change of particular importance, as the individual measures are not necessarily highly correlated because they measure a similar construct, but because the constructs *themselves* are strongly related. Brysbaert et al. (2019) identified similar Frequency and Complexity-like factors, with variables loading according to our model’s factor pattern—including contextual diversity onto Frequency. The main difference is that the orthographic and phonologic Levenshtein distances for the 20 closest neighbors (old20 and pld20) loaded onto *both* the Complexity and Proximity factors, unlike our models and Yap et al.’s (2012) model, where old20 and pld20 only loaded onto Complexity. However, it is worth noting that old20, pld20, and neighborhood density measures would have cross-loaded onto Complexity and Proximity in Yap et al.’s (2012) model if they had used a .30 cutoff for factor loadings, as in Brysbaert et al. and the current study. Paivio’s (1968) and Clark & Paivio’s (2004) models do not include old20, pld20, or any neighbor measures. Brysbaert et al. (2019) also found a similar pattern to our correlation matrix for general academic words.

Limitations

The study has important limitations. First, though conceptually different, the three wordlists used in this study are not completely distinct at the word level, and alternatives could have been used. We believe the consistency of findings across these lists suggests that these results will generalize to other lists representing more specialized contexts. It would be

particularly interesting to see if these findings replicate with lists of words used frequently in child directed speech. Second, although the present study includes many word features, future research will produce additional measures.

Research applications

The current study indicated that five main latent factors underlie the empirical non-behavioral lexical measures, namely, Frequency, Complexity, Proximity, Polysemy, and Diversity, that may prove useful to understand how learners learn new words, select equivalent words for assessment or stimuli, or statistically control for differences in said stimuli. For example, Lawrence and colleagues (2022) used these five latent factors to explore interactions between lexical characteristics and reading performance.^{viii} In their study, factor scores obviated the need to make difficult decisions about specific measures to include while still accounting for the maximum effects of word characteristics on item difficulty. Future research can also use latent factor scores to identify sets of matched words when designing innovative intervention studies or vocabulary knowledge measures, for example, matching on Frequency as a holistic dimension, as opposed to a single corpus frequency measure. Given the advantages of the latent estimates, we therefore provide estimates for all 400K+ unique letter strings on all three reference models at XXX.edu for noncommercial use.

References

- Adelman, J. S. & Brown, G. D. (2007). Phonographic neighbors, not orthographic neighbors, determine word naming latencies. *Psychonomic Bulletin & Review*, *14*(3), 455-459.
- 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*, 814-823.
- Anglin, J. M., Miller, G. A., & Wakefield, P. C. (1993). Vocabulary development: A morphological analysis. *Monographs of the Society for Research in Child Development*, *58*(10), i-186.
- 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.
- Balota, D. A., Yap, M. J., Hutchison, K. A., Cortese, M. J., Kessler, B., Loftis, B., Neely, J.H., Nelson, D.L., Simpson, G.B., & Treiman, R. (2007). The English lexicon project. *Behavior Research Methods*, *39*(3), 445-459.
- Bates, E., Bretherton, I., & Snyder, L. S. (1991). *From first words to grammar: Individual differences and dissociable mechanisms* (Vol. 20). Cambridge: Cambridge University Press.
- Beck, I. L., McKeown, M. G., & Kucan, L. (2002). *Bringing words to life: Robust vocabulary instruction*. Guilford Press.
- Beretta, A., Fiorentine, R., & Poeppel, D. (2005). The effects of homonymy and polysemy on lexical access: An MEG study. *Brain Research: Cognitive Brain Research*, *24*(1) 57-65.
- Bhattacharya, A., & Ehri, L. C. (2004). "Graphosyllabic Analysis Helps Adolescent Struggling Readers Read and Spell Words." *Journal of Learning Disabilities*, *37*(4), 331.

- 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.
- Borowsky, R., & Masson, M. E. (1996). Semantic ambiguity effects in word identification. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 22*(1), 63-85.
- Breland, H. M. (1996). Word frequency and word difficulty: A comparison of counts in four corpora. *Psychological Science, 7*(2), 96-99.
- Brezina, V., & Gablasova, D. (2015). Is there a core general vocabulary? Introducing the New General Service List. *Applied Linguistics, 36*(1), 1-22.
- Bryant, P., Goswami, U. (1987). Beyond grapheme-phoneme correspondence. *Cahiers de Psychologie Cognitive/Current Psychology of Cognition, 7*(5), 439–443.
- 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.
- Brysbaert, M., Mandera, P., & Keuleers, E. (2018). The word frequency effect in word processing: An updated review. *Current Directions in Psychological Science, 27*(1), 45-50
- Brysbaert, M., Mandera, P., McCormick, S. F., & Keuleers, E. (2019). Word prevalence norms for 62,000 English Lemmas. *Behavior Research Methods, 51*(2), 467-479.
- Carlisle, J. F. (2000). awareness of the structure and meaning of morphologically complex words: Impact on reading. *Reading and Writing, 12*(3), 169-190.
- 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), 153–85.
- Clark, J. M., & Paivio, A. (2004). Extensions of the Paivio, Yuille, and Madigan (1968) norms. *Behavior Research Methods, Instruments, & Computers, 36*(3), 371-383.
- Coxhead, A. (2000). A new academic word list. *TESOL Quarterly, 34*(2), 213-238.

- Dale, E., & O'Rourke, J. (1981). *The Living Word Vocabulary*. Chicago: World Book/Childcraft International.
- Davies, M. (2008). The Corpus of Contemporary American English (COCA): One billion million words, 1990-2019.
- DeRocher, J. (1973). *The Counting of Words: A Review of the History, Techniques and Theory of Word Counts with Annotated Bibliography*. New York, NY: Syracuse University Research Corporation.
- Ehri, L. C. (2014) Orthographic Mapping in the Acquisition of Sight Word Reading, Spelling Memory, and Vocabulary Learning, *Scientific Studies of Reading*, 18(1), 5-21, DOI: 10.1080/10888438.2013.819356
- Ehri, L. C. (2005). Learning to read words: Theory, findings, and issues. *Studies of Reading*, 9(2), 167-188, DOI: 10.1207/s1532799xssr0902_4
- Ehri, L. C. (1992). Reconceptualizing the development of sight word reading and its relationship to recoding. In P. B. Gough, L. C. Ehri, & R. Treiman (Eds.), *Reading acquisition* (pp. 107–143). Lawrence Erlbaum Associates, Inc.
- Ellis, N. C. (2002). Frequency effects in language processing: A review with implications for theories of implicit and explicit language acquisition. *Studies in Second Language Acquisition*, 24, 143–188.
- Fellbaum, C. (2005). *WordNet and wordnets*.
- Fenson, L., Dale, P., Reznick, J. S., Bates, E., Thal, D., & Pethick, S. (1994). Variability in early communicative development. *Monographs for the Society for Research in Child Development*, 59(5)
- Field, A. (2013). *Discovering statistics using IBM SPSS statistics*. Sage Publishing.
- Gardner, D., & Davies, M. (2014). A new academic vocabulary list. *Applied Linguistics*, 35(3), 305-327.

- Goldfield, B. A., & Reznick, J. S. (1990). Early lexical acquisition: Rate, content, and the vocabulary spurt. *Journal of Child Language*, *17* (1), 171-183.
- Goodman, J. C., Dale, P. S., & Li, P. (2008). Does frequency count? Parental input and the acquisition of vocabulary. *Journal of Child Language*, *35*(3), 515-531
- Hiebert, E. H., & Fisher, C. W. (2005). A review of the national reading panel's studies on fluency: The role of text. *The Elementary School Journal*, *105*(5), 443–460.
- 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.
- 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.
- 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.
- 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.
- Jalbert, A., Neath, I., & Surprenant, A. M. (2011). Does length or neighborhood size cause the word length effect? *Memory and Cognition*, *39*, 1198-1210.
- Jones, M. N., Johns, B. T., & Recchia, G. (2012). The role of semantic diversity in lexical organization. *Canadian Journal of Experimental Psychology*, *66*, 115–124.
- Kjeldsen, A. C., Niemi, P., & Olofsson, Å. (2003). Training phonological awareness in kindergarten level children: Consistency is more important than quantity. *Learning and Instruction*, *13*(4), 349-365.

- Klepousniotou, E. (2002). The processing of lexical ambiguity: Homonymy and polysemy in the mental lexicon. *Brain and Language, 81*(1), 205-223.
- Kučera, H., & Francis, W. N. (1967). *Computational analysis of present-day American English*. Dartmouth.
- Kuperman, V., Stadthagen-Gonzalez, H., & Brysbaert, M. (2012). Age-of-Acquisition ratings for 30,000 English words. *Behavior Research Methods, 44*(4), 978-990
- Landauer, T. K., Kireyev, K., & Panaccione, C. (2011). Word maturity: A new metric for word knowledge. *Scientific Studies of Reading, 15*(1), 92-108.
- Lawrence, J. F., Knoph, R., McIlraith, A., Kulesz, P. A., & Francis, D. J. (2022). Reading comprehension and academic vocabulary: Exploring relations of item features and reading proficiency. *Reading Research Quarterly, 57*(2), 669-690
<https://doi.org/10.1002/rrq.434>
- 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>
- Laxon, V. J., Coltheart, V., & Keating, C. (1988). Children find friendly words friendly too: Words with many orthographic neighbours are easier to read and spell. *British Journal of Educational Psychology, 58*(1), 103-119.
- Leech, G., & Rayson, P. (2014). *Word Frequencies in Written and Spoken English: Based on the British National Corpus*. Routledge.
- 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>

- Levenshtein, V. I. (1966). Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics Doklady*, 10(8), 707-710.
- Lin, Y., Michel, J. B., Lieberman, E. A., Orwant, J., Brockman, W., & Petrov, S. (2012, July). Syntactic annotations for the google books ngram corpus. In *Proceedings of the ACL 2012 system demonstrations* (pp. 169-174).
- Loewen, S., & Gonulal, T. (2015). Exploratory factor analysis and principal components analysis. In *Advancing quantitative methods in second language research* (pp. 182-212). Routledge.
- Lund, K., & Burgess, C. (1996). Producing high-dimensional semantic spaces from lexical co-occurrence. *Behavior Research Methods, Instruments, & Computers*, 28(2), 203-208.
- Mahowald, K., Fedorenko, E., Piantadosi, S. T., & Gibson, E., (2013). Info/information theory: Speakers choose shorter words in predictive contexts, *Cognition*, 126, 313-318.
- 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), 438-441
- Miller, G. A. (1990). WordNet: An online lexical database. *International Journal of Lexicography*, 3(4), 235–312.
- Monsell, S., Doyle, M. C., & Haggard, P. N. (1989). Effects of frequency on visual word recognition tasks: Where are they? *Journal of Experimental Psychology: General*, 118, 43–71
- Nagy, W., & Townsend, D. (2012). Words as tools: Learning academic vocabulary as language acquisition. *Reading Research Quarterly*, 47(1), 91-108.
- Paivio, A. (1968). A factor-analytic study of word attributes and verbal learning. *Journal of Verbal Learning and Verbal Behavior*, 7, 41-49.

- Parks, R., Ray, J., & Bland, S. (1998). Wordsmyth English Dictionary-Thesaurus [Electronic version]. Chicago, IL: University of Chicago. Retrieved from www.wordsmyth.net
- Perfetti, C., & Hart, L. (2002). The lexical quality hypothesis. In L. Verhoeven, C. Elbro, & P. Reitsma (Eds.), *Precursors of functional literacy* (pp. 189–213). John Benjamins Publishing Company.
- Piantadosi, S. T., Tily, H., & Gibson, E. (2011). Word lengths are optimized for efficient communication, *Proceedings of the National Academic of Sciences of the United States of America*, 108(9), 3526-3529. <https://doi.org/10.1073/pnas.1012551108>
- Raiche, G. (2010). *an R package for parallel analysis and non graphical solutions to the Cattell scree test*. R package version 2.3.3.1., <https://CRAN.R-project.org/package=nFactors>.
- Ravin, Y., & Leacock, C. (2000). *Polysemy: Theoretical and computational approaches*. OUP Oxford.
- Revelle, W. (2020). *psych: Procedures for Psychological, Psychometric, and Personality Research*. Northwestern University, Evanston, Illinois. R package version 2.0.12, <https://CRAN.R-project.org/package=psych>.
- Rodd, J. M., Gaskell, G., & Marslen-Wilson, W. (2004). Modelling the effects of semantic ambiguity in word recognition. *Cognitive Science*, 28(1), 89-104.
- Rodd, J. M., 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.
- Singson, M., Mahony, D., & Mann, V. (2000). The relation between reading ability and morphological skills: Evidence from derivational suffixes. *Reading and Writing*, 12 (3), 219-252.

- Snow, C. E. (1972). Mothers' speech to children learning language. *Child Development*, 43(2), 549–565.
- Storkel, H. L. (2004). Do children acquire dense neighborhoods? An investigation of similarity neighborhoods in lexical acquisition. *Applied Psycholinguistics*, 25, 201–221.
- Storkel, H. L. (2009). Developmental differences in the effects of phonological, lexical, and semantic variables on word learning by infants. *Journal of Child Language*, 36(2), 291–321.
- Sullivan, J. (2007). “Developing Knowledge of Polysemous Vocabulary.” University of Waterloo. <https://uwspace.uwaterloo.ca/handle/10012/2637>.
- Swanborn, M. S., & De Glopper, K. (1999). Incidental word learning while reading: A meta-analysis. *Review of Educational Research*, 69(3), 261–285.
- Touchstone Applied Science Associates; <http://lsa.colorado.edu/spaces.html>
- Van Heuven, W. J. B., Mandera, P., Keuleers, E., Brysbaert, M. (2014). SUBTLEX-UK: A new and improved word frequency database for British English. *The Quarterly Journal of Experimental Psychology*, 67(6), 1176–1190.
- Vitevitch, M. S., Storkel, H. L., Francisco, A. C., Evans, K. J., & Goldstein, R. (2014). The influence of known-word frequency on the acquisition of new neighbours in adults: Evidence for exemplar representations in word learning. *Language, Cognition and Neuroscience*, 29(10), 1311–1316.
- West, M. (1953). *A General Service List of English Words*. Longmans, Green and Co.
- 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.
- 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

Youn, H., Sutton, L., Smith, E., Moore, C., Wilkins, J. F., Maddieson, I., Croft, W., &

Bhattacharya, T. (2016). On the universal structure of human lexical semantics.

Proceedings of the National Academy of Sciences of the United States of America,

113(7), 1766–1771. <https://doi.org/10.1073/pnas.1520752113>

Zeno, S., Ivens, S. H., Millard, R. T., & Duvvuri, R. (1995). *The Educator's Word Frequency*

Guide. Touchstone Applied Science Associates.

ⁱ Using the raw frequency can be problematic in model estimation because of Zipf's law: the frequency of a word is inversely proportional to its ranking. A few high-ranking words take up a significant portion of corpora (e.g. “the”, “and”, “a”), many low-ranking words take up a small portion of corpora (e.g. “projectile”, “calendar”), and frequency and rankings are not linearly related. For this reason, linear models tend to instead be based on some transformation of the raw frequency—either a log transformation or zipfian transformation. The zipfian transformation accounts for the word frequency effect based on Zipf's law (Van Heuven, Mandera, Keuleers, & Brysbaert, 2014) and is calculated as: $\log_{10} \left(\frac{\text{raw frequency} + 1}{\text{corpus size in millions} + \text{word types in millions}} \right) + 3$

ⁱⁱ Because the COCA splits by part of speech, we totaled word frequency for all parts of speech before taking the Zipfian transformation.

ⁱⁱⁱ Because the OED is split by part of speech, we used the highest occurring frequency band for each word.

^{iv} Neighborhood sizes were collected from the English Lexicon Project (Yap et al., 2012), however, no specific corpus is disclosed.

^v Because Ngram data splits by part of speech, we used the oldest occurrence for word age.

^{vi} Because WordNet splits by part of speech, we took the average score for each word.

^{vii} Other measures were considered for the factor analysis, but were too highly correlated with other measures ($r > .98$; Standardized Frequency Index (SFI) from Subtlex with zenozipf, and Contextual Diversity and Word Frequency from Subtlex with subzipf) or did not have enough variability to warrant inclusion for any word set (MSA $< .60$; mean bigram from English Lexicon Project and word age from Oxford English Dictionary).

^{viii} This paper uses the general academic word (AWL-reference) model to estimate factor scores on a specific set of vocabulary from the Word Generation trials. Factor score estimates for these words differ in the current paper when the words are scaled based on the GSL-reference, AWL-reference, and AVL-DS-reference as opposed to scaled amongst themselves.

Table 1

Description of Word Feature Measures

Measure	Name	Description	Corpus	Description	Citation
cd	contextual diversity	the number of documents in which a word appears	Touchstone Applied Science Associates (TASA)	Approx. 120,000 paragraphs taken from 38,000 academic texts	Adelman et al., 2006
cocazipf	COCA frequency	Zipfian-transformed word frequencies	Corpus of Contemporary American English (COCA)	Approx. 560 million words from T.V., radio, newspapers, fiction, academic papers, and popular magazines	Davies, 2008
d	dispersion	number of subject areas in which a word appears; scores range from 0 (only one area) to 1 (all areas)	The Educator's Word Frequency Guide	Approx. 17 million words from kindergarten- to college-level texts	Zeno, Ivens, Millar, & Duvvuri, 1995
freqband	frequency band	frequency band grouping from the Oxford English Dictionary; bands run from 1 (infrequent) to 8 (frequent) a	Google Ngrams, version 2	Approx. 450 million words scanned from Google Books	Lin et al., 2012
length	length	number of letters; collected from ELP		NA	Balota et al., 2007
log_freq_hal	HAL frequency	log-transformed word frequencies; collected from the English Lexicon Project (ELP; Balota et al., 2007)	Hyperspace Analogue to Language	Approx. 131 million words from 3,000 Usenet newsgroups	Lund & Burgess, 1996
log_freq_kf	Brown frequency	log-transformed word frequencies; collected from the English Lexicon Project (ELP; Balota et al., 2007)	Brown Corpus of Standard American English	Approx. 1 million words from American English texts	Kučera & Francis, 1967
nmorph	morphemes	number of morphemes; collected from ELP		NA	Balota et al., 2007
nphon	phonemes	number of phonemes; collected from ELP		NA	Balota et al., 2007
nsyll	syllables	number of syllables; collected from ELP		NA	Balota et al., 2007
og_n	phonographic neighbors	raw number of phonographic neighbors (i.e., the number of words that are one letter <i>and</i> one phoneme away from the word, e.g., "stove" and "stone"), excluding homophones	English Lexicon Project	Standardized behavioral and descriptive data set for 40,481 words (and 40,481 nonwords)	Balota, Yap, Hutchison, Cortese, & Klesser, et al., 2007

Measure	Name	Description	Corpus	Description	Citation
old20	orthographic Levenshtein distance	mean Levenshtein distance of the 20 closest orthographic neighbors	English Lexicon Project	Standardized behavioral and descriptive data set for 40,481 words (and 40,481 nonwords)	Yarkoni et al., 2008
ortho_n	orthographic neighbors	raw number of orthographic neighbors (i.e., the number of words that are one letter away from the word, e.g., “lost” and “lose”), excluding homophones	English Lexicon Project	Standardized behavioral and descriptive data set for 40,481 words (and 40,481 nonwords)	Balota et al., 2007
phono_n	phonologic neighbors	raw number of phonologic neighbors (i.e., the number of words that are one phoneme away from the word, e.g., “hear” and “hare”), excluding homophones	English Lexicon Project	Standardized behavioral and descriptive data set for 40,481 words (and 40,481 nonwords)	Balota et al., 2007
pld20	phonologic Levenshtein distance	mean Levenshtein distance of the 20 closest phonologic neighbors	English Lexicon Project	Standardized behavioral and descriptive data set for 40,481 words (and 40,481 nonwords)	Yarkoni et al., 2008
semd	semantic diversity	mean cosine of the latent semantic analysis vectors for all pairwise combinations of contexts containing the word	British National Corpus	Approx. 100 million words from T.V., radio, newspapers, fiction, academic papers, and popular magazines	Hoffman et al., 2013
subzipf	Subtlex frequency	Zipfian-transformed word frequencies	SubtlexUS corpus (Subtitle Lexicon-U.S. version)	Approx. 51 million words from American subtitles	Brybaert & New, 2009
wordage	word age	number of years since a word was first used (as of 2000), based on oldest occurrence across parts of speech	Google Ngrams, version 2	Approx. 450 million words scanned from Google Books	Lin et al., 2012
wordnet_lnaposam	Wordnet senses and meanings	log-transformed number of senses and meanings a word has across all parts of speech	WordNet lexical database	Database of 155,327 words organized in 175,979 synonym sets and hypernym chains	Fellbaum, 2005
wordsmyth_lnaposam	Wordsmyth senses and meanings	log-transformed number of senses and meanings a word has across all parts of speech	Wordsmyth integrated dictionary and thesaurus	Advanced dictionary and integrated thesaurus for 60,000 headwords	Parks, Ray, & Bland, 1998
z_sem_prec	semantic precision	z-transformed depth scores averaged across parts of speech; scores range from 0 (shallow/vague) to 10 (deep/precise)	WordNet lexical database	Database of 155,327 words organized in 175,979 synonym sets and hypernym chains	Fellbaum, 2005

Measure	Name	Description	Corpus	Description	Citation
zenozipf	Zeno frequency	Zipfian-transformed word frequencies	The Educator's Word Frequency Guide	Approx. 17 million words from kindergarten- to college-level texts	Zeno, Ivens, Millar, & Duvvuri, 1995

Table 2*Descriptive Statistics for Word Features by Word List*

Word Feature	General Service List (GSL) n = 2060		Academic Word List (AWL) n = 1051		Academic Vocabulary List - Domain Specific (AVL-DS) n = 3413	
	M	SD	M	SD	M	SD
cd	1036.52	2453.23	175.54	300.74	142.06	424.10
cocazipf	4.62	0.64	4.15	0.57	3.82	0.59
d	0.80	0.14	0.71	0.19	0.56	0.21
freqband	6.09	0.67	5.83	0.62	5.31	0.60
length	5.84	2.01	8.57	2.33	7.31	2.45
log_freq_hal	9.72	1.60	8.70	1.45	7.71	1.52
log_freq_kf	1.62	0.63	1.16	0.55	0.82	0.58
nmorph	1.32	0.58	2.19	0.83	1.73	0.78
nphon	4.75	1.81	7.40	2.15	6.19	2.27
nsyll	1.75	0.86	3.02	1.05	2.45	1.13
og_n	3.01	4.54	0.36	1.33	1.54	3.30
old	2.02	0.72	2.84	0.79	2.59	0.94
ortho_n	5.00	6.38	0.65	1.84	2.52	4.59
phono_n	11.87	14.39	1.39	4.31	5.84	10.59
pld	1.87	0.82	2.94	0.98	2.57	1.15
semd	1.80	0.27	1.79	0.25	1.44	0.30
subzipf	4.44	0.75	3.41	0.67	3.38	0.77
word_age	742.83	243.53	499.54	185.46	542.02	251.93
wordnet_lnapossam	1.73	0.82	1.26	0.68	1.09	0.79
wordsmyth_lnapossam	1.70	0.73	1.06	0.61	1.07	0.72
z_sem_prec	-0.03	0.73	-0.14	0.87	0.24	0.86
zenozipf	4.67	0.64	3.91	0.60	3.71	0.68

Note. This table includes only the final 22 features used in the models.

Table 3*Factor Analysis fit by Word List Reference*

	General Service List (GSL)	Academic Word List (AWL)	Academic Vocabulary List - Domain Specific (AVL-DS)
Bartlett's test			
df	231	231	231
X ²	50136.55	20743.13	72063.03
p	< .0001	< .0001	< .0001
Plot suggestions			
Parallel Analysis (fa.parallel)	5	5	5
Parallel Analysis (nFactors)	4	5	4
Optimal Coordinates (nFactors)	4	5	4
Eigenvalues > Mean	5	5	4
Eigenvalues > 1	5	5	4
Acceleration Factor	1	1	1
5-Factor Model Fit			
% Variance Explained	75 %	69 %	68 %
Overall Sampling Adequacy (MSA)	0.92	0.90	0.92
Comparative Fit Index (CFI)	0.962	0.948	0.956
Tucker-Lewis Index (TLI)	0.934	0.907	0.923
Root Mean Square of Residuals (RMSR)	0.02	0.02	0.03
Root Mean Square Error of Approximation (RMSEA)	0.083	0.088	0.084
Lower Bound	0.080	0.084	0.081
Upper Bound	0.087	0.093	0.086
4-Factor Model Fit			
% Variance Explained	71 %	65 %	67 %
Overall Sampling Adequacy (MSA)	0.92	0.90	0.92
Comparative Fit Index (CFI)	0.943	0.923	0.933
Tucker-Lewis Index (TLI)	0.912	0.88	0.896
Root Mean Square of Residuals (RMSR)	0.03	0.04	0.03
Root Mean Square Error of Approximation (RMSEA)	0.096	0.010	0.097
Lower Bound	0.093	0.096	0.095
Upper Bound	0.099	0.105	0.100

Note. Higher values (at least above .90) indicate adequate model fit for MSA, CFI, and TLI. Lower values (at least below .10) indicate adequate model fit for RMSR and RMSEA.

Table 4

Factor Loadings by Wordlist

	Frequency			Complexity			Proximity			Polysemy (Polysemy/Diversity) ^a			Diversity	
	GSL	AWL	AVL-DS	GSL	AWL	AVL-DS	GSL	AWL	AVL-DS	GSL	AWL	AVL-DS	GSL	AWL
	cocazipf	1.00	0.99	0.98	0.02	0.02	0.02	-0.01	0.01	0.01	0.01	-0.04	-0.03	-0.02
log_freq_kf	0.96	0.86	0.83	0.08	0.07	0.09	0.00	0.00	-0.03	-0.01	0.03	0.06	0.03	0.06
zenozipf	0.95	0.89	0.84	-0.07	-0.01	-0.07	-0.01	0.01	0.02	0.02	0.02	-0.01	-0.06	0.06
log_freq_hal	0.90	0.89	0.84	0.02	-0.06	-0.03	0.02	0.00	0.02	0.02	0.02	0.03	0.04	-0.07
subzipf	0.84	0.73	0.77	-0.17	-0.11	-0.21	0.02	0.03	0.04	0.01	0.09	-0.02	-0.05	0.07
freqband	0.75	0.72	0.76	0.08	-0.07	0.10	-0.01	-0.08	-0.06	0.02	-0.01	-0.01	0.12	-0.07
cd	0.61	0.61	0.58	-0.06	0.04	0.05	0.09	0.12	0.07	-0.17	-0.02	0.01	0.15	0.04
wordage	0.30	0.29	0.29	0.06	0.06	0.02	0.00	-0.02	-0.02	0.11	0.16	0.29	0.16	0.32
nphon	-0.02	-0.02	-0.01	0.98	0.97	0.98	0.01	0.01	0.00	0.06	0.05	0.03	-0.01	0.01
nsyll	0.04	0.05	-0.01	0.90	0.84	0.90	0.02	-0.02	0.01	-0.06	-0.06	-0.03	0.01	-0.11
length	-0.05	-0.03	-0.01	0.89	0.97	0.94	-0.09	0.01	-0.05	0.03	0.07	0.07	-0.02	0.01
pld	0.03	0.02	0.00	0.87	0.86	0.88	-0.06	-0.02	0.00	-0.07	-0.08	-0.11	-0.02	0.00
nmorph	-0.07	-0.09	-0.02	0.78	0.69	0.77	0.16	0.04	0.08	-0.04	0.00	0.04	0.03	0.01
old	0.00	0.01	0.00	0.73	0.80	0.78	-0.22	-0.08	-0.11	-0.09	-0.12	-0.12	0.00	0.05
og_n	-0.01	0.01	-0.02	0.05	0.06	0.06	0.96	0.99	0.98	0.03	0.01	0.02	-0.02	0.00
ortho_n	0.03	0.00	0.02	-0.03	-0.02	-0.01	0.95	0.95	0.98	0.01	-0.01	-0.02	0.00	-0.01
phono_n	0.01	0.02	0.02	-0.27	-0.11	-0.17	0.64	0.78	0.73	0.01	-0.01	0.02	0.02	0.01
wordnet_lnapossam	0.00	-0.03	0.01	0.04	0.00	0.00	0.02	0.02	0.01	0.96	0.87	0.90	0.02	0.05
wordsmyth_lnapossam	0.05	0.10	0.03	-0.11	-0.08	-0.11	0.00	-0.02	0.04	0.79	0.68	0.70	0.00	-0.07
semid	0.00	0.01	0.24	-0.01	-0.01	0.06	-0.01	0.00	-0.01	0.01	0.05	0.39	0.98	0.74
d	0.27	0.33	0.52	0.03	-0.04	0.00	-0.01	0.00	-0.01	0.10	0.02	0.19	0.46	0.55
z_sem_prec	0.07	0.21	0.04	0.07	0.09	-0.14	0.04	0.02	0.00	0.03	0.24	-0.26	-0.42	-0.44
Explained Variance	26 %	24 %	24 %	22 %	21 %	22 %	12 %	12 %	12 %	8 %	7 %	9 %	7 %	6 %
Eigenvalue	5.65	5.24	5.21	4.91	4.65	4.93	2.55	2.58	2.62	1.77	1.49	1.93	1.59	1.31
Standardized Alpha	0.94	0.93	0.93	0.96	0.95	0.96	0.93	0.94	0.95	0.90	0.79	0.75	0.68	0.62

Note. GSL - General Service List, AWL - Academic Word List, AVL-DS - Academic Vocabulary List (Domain-Specific). Loadings greater than .30 are bolded for ease.

^a For the 4-factor solutions, results for the Polysemy-Diversity factor are reported under Polysemy.

Table 5

Example Words and Factor Estimates based on Each Model

word	Word List	GSL (General Service List) Reference Model					AWL (Academic Word List) Reference Model					AVLDS (Domain-Specific Word) Reference Model				
		Frequency	Complexity	Proximity	Polysemy	Diversity	Freq	Comp	Prox	Poly	Div	Freq	Comp	Prox	Poly-Div	
accept	GSL	0.41	0.49	-0.64	0.66	1.04	1.47	-0.89	-0.07	1.52	1.12	2.04	-0.22	-0.38	1.62	
honest	GSL	-0.05	0.09	-0.75	0.15	0.22	0.93	-1.16	-0.27	0.92	0.81	1.48	-0.53	-0.53	1.11	
jaw	GSL	-0.88	-1.08	1.91	0.07	-1.21	-0.10	-2.31	8.36	0.86	-0.78	0.46	-1.56	3.15	0.89	
library	GSL	0.34	1.18	-0.78	-0.43	-1.66	1.42	-0.27	-0.31	0.25	-0.36	1.86	0.31	-0.56	0.37	
meet	GSL	0.91	-0.91	0.41	0.98	1.34	2.09	-2.13	3.89	1.93	1.45	2.62	-1.36	1.08	1.97	
nice	GSL	0.68	-0.93	0.68	0.17	-0.60	1.82	-2.08	4.83	1.03	-0.30	2.27	-1.36	1.44	1.08	
race	GSL	0.65	-1.05	1.68	0.92	-1.13	1.69	-2.22	8.42	1.86	-0.44	2.18	-1.44	2.80	1.89	
size	GSL	0.91	-0.96	0.09	0.63	1.01	2.18	-2.15	3.10	1.51	1.17	2.62	-1.38	0.61	1.68	
usual	GSL	0.32	0.39	-0.77	-1.12	1.22	1.33	-0.94	-0.28	-0.57	1.61	1.81	-0.29	-0.55	0.00	
wood	GSL	0.30	-0.89	0.93	0.52	-0.44	1.43	-2.08	4.96	1.43	-0.02	1.86	-1.34	1.84	1.40	
abandon	AWL	-0.82	0.99	-0.77	0.16	1.06	-0.01	-0.42	-0.31	0.97	0.50	0.62	0.18	-0.54	1.09	
compile	AWL	-1.97	0.55	-0.70	-0.74	0.03	-1.17	-0.77	-0.28	-0.16	-0.62	-0.51	-0.17	-0.52	0.11	
evidence	AWL	0.93	1.22	-0.77	0.04	0.70	2.02	-0.22	-0.30	0.89	0.84	2.55	0.34	-0.57	1.07	
illustrate	AWL	-0.82	1.85	-0.77	-0.82	0.49	0.01	0.39	-0.32	-0.14	0.74	0.58	0.84	-0.57	0.24	
maximize	AWL	-1.53	1.46	-0.71	-1.46	-0.73	-0.82	0.03	-0.32	-1.00	-0.87	-0.13	0.58	-0.55	-0.57	
nonetheless	AWL	-0.74	2.65	-0.82	-2.24	0.96	-0.03	1.02	-0.35	-1.86	1.17	0.60	1.35	-0.60	-1.06	
process	AWL	1.18	0.56	-0.55	0.97	0.89	2.41	-0.78	0.13	1.91	1.12	2.92	-0.15	-0.23	1.97	
ratio	AWL	-0.39	0.20	-0.43	-1.26	-0.52	0.47	-1.11	0.54	-0.71	-0.33	0.98	-0.47	-0.16	-0.32	
significant	AWL	0.60	3.10	-0.77	-0.71	1.34	1.50	1.36	-0.34	-0.09	1.37	2.14	1.84	-0.60	0.45	
underlying	AWL	-0.62	1.88	-0.71	-0.26	0.59	0.16	0.38	-0.32	0.47	0.46	0.81	0.87	-0.55	0.85	
absenteeism	AVLDS	-2.76	3.53	-0.76	-2.59	-1.44	-2.23	1.68	-0.38	-2.25	-1.58	-1.51	1.99	-0.58	-1.68	
autopsy	AVLDS	-1.85	0.89	-0.76	-1.44	-2.14	-1.19	-0.47	-0.30	-0.95	-2.18	-0.49	0.07	-0.55	-0.69	
carbon	AVLDS	-0.19	0.13	-0.42	-0.49	-2.29	0.84	-1.06	0.36	0.23	-1.50	1.26	-0.41	-0.06	0.31	
habitable	AVLDS	-2.47	1.74	-0.68	-2.18	-1.47	-1.91	0.32	-0.31	-1.76	-1.03	-1.32	0.78	-0.53	-1.16	
membrane	AVLDS	-1.25	1.21	-0.80	-1.37	-2.50	-0.33	-0.13	-0.31	-0.79	-1.63	0.12	0.36	-0.57	-0.65	
linguistic	AVLDS	-1.27	2.50	-0.74	-1.55	-2.86	-0.57	0.93	-0.35	-1.09	-1.49	0.08	1.40	-0.58	-0.65	
positivism	AVLDS	-2.86	2.90	-0.56	-1.81	-3.68	-2.34	1.21	-0.16	-1.36	-3.16	-1.69	1.67	-0.39	-1.03	
sacrificial	AVLDS	-2.37	2.88	-0.75	-2.36	-2.01	-1.79	1.22	-0.36	-1.98	-1.86	-1.12	1.58	-0.57	-1.41	
tuberculosis	AVLDS	-1.50	4.60	-0.89	-2.51	-1.72	-0.76	2.58	-0.43	-2.06	-0.87	-0.26	2.85	-0.63	-1.73	
wavelength	AVLDS	-1.75	2.04	-0.86	-1.61	-2.43	-0.97	0.53	-0.35	-1.04	-2.70	-0.44	0.89	-0.59	-0.93	

Note. Bolded estimates indicate when words were part of the sample used to create that model. Not bolded estimates were estimated separately, after creating models.

For brevity, Freq – Frequency, Comp – Complexity, Prox – Proximity, Poly – Polysemy, Div – Diversity, and Poly-Div – Polysemy/Diversity factors.

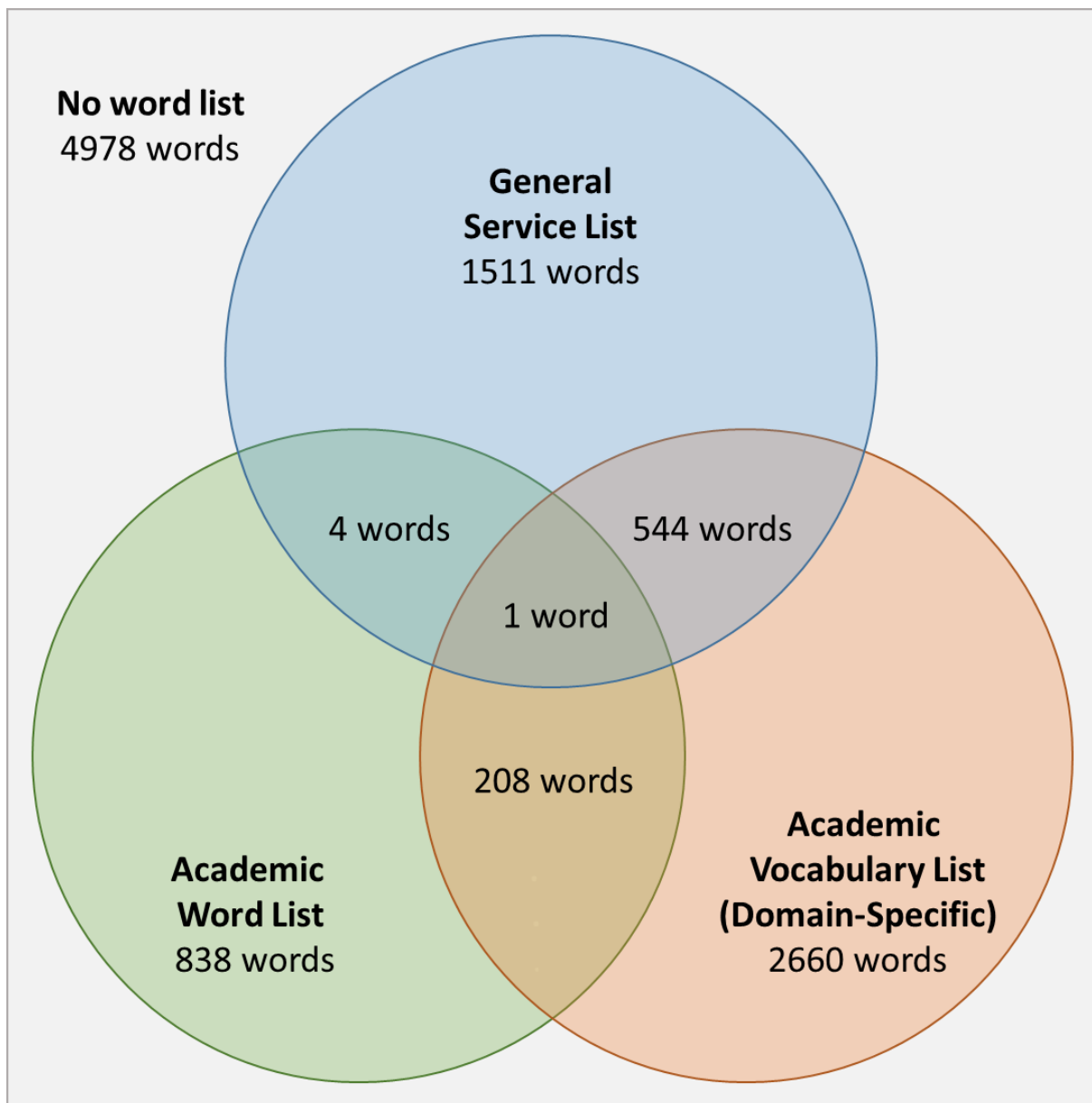


Figure 1. Overlap between word lists for unique words with complete data (n = 19,744 words).

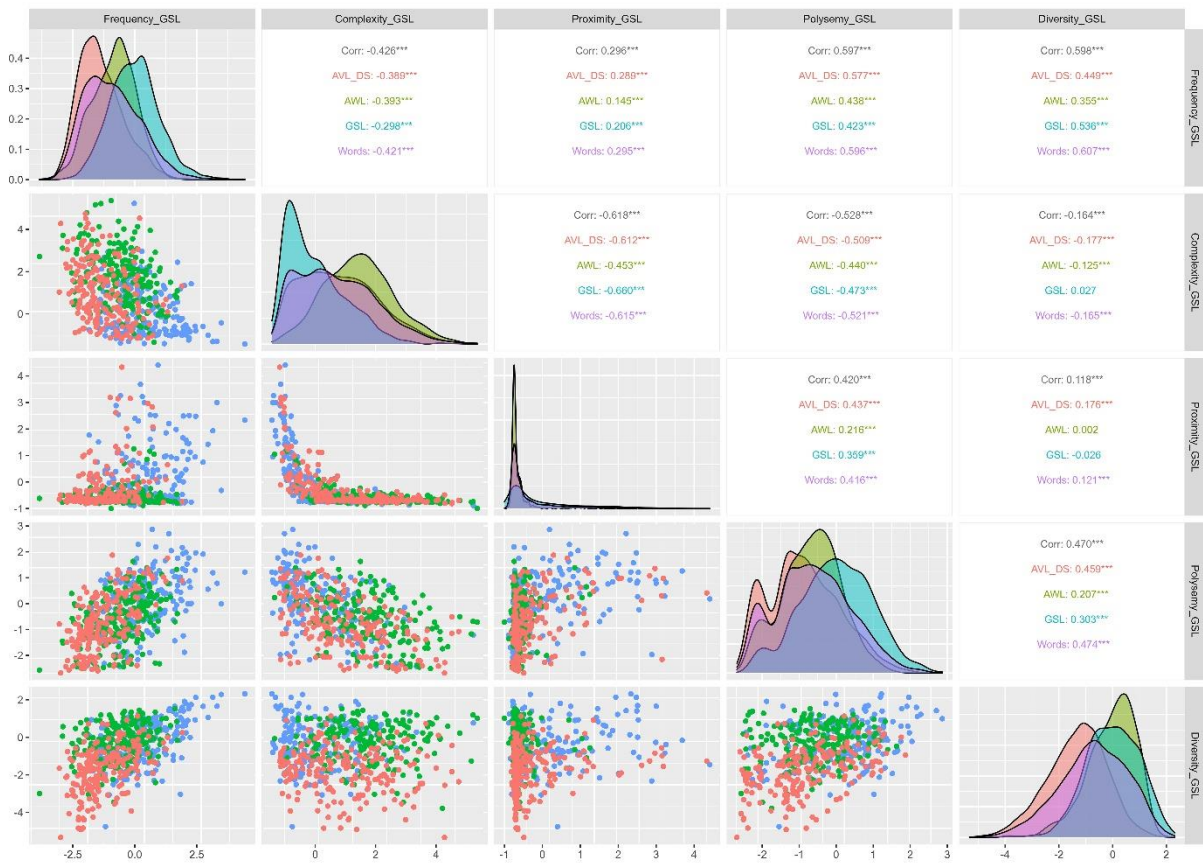


Figure 2. Factor correlations for the GSL-reference model estimated on each word sample. Note: Scatterplots below the diagonal contain random 200-word samples while density plots and correlations on and above the diagonal are based on entire word samples. Word samples include AVL_DS (red) Academic Vocabulary List Domain-Specific; AWL (green) Academic Word List; GSL (blue) General Service List; Words (purple) in any list.



Figure 3. Factor correlations for the AWL-reference model estimated on each word sample. Note: Scatterplots below the diagonal contain random 200-word samples while density plots and correlations on and above the diagonal are based on entire word samples. Word samples include AVL_DS (red) Academic Vocabulary List Domain-Specific; AWL (green) Academic Word List; GSL (blue) General Service List; Words (purple) in any list.

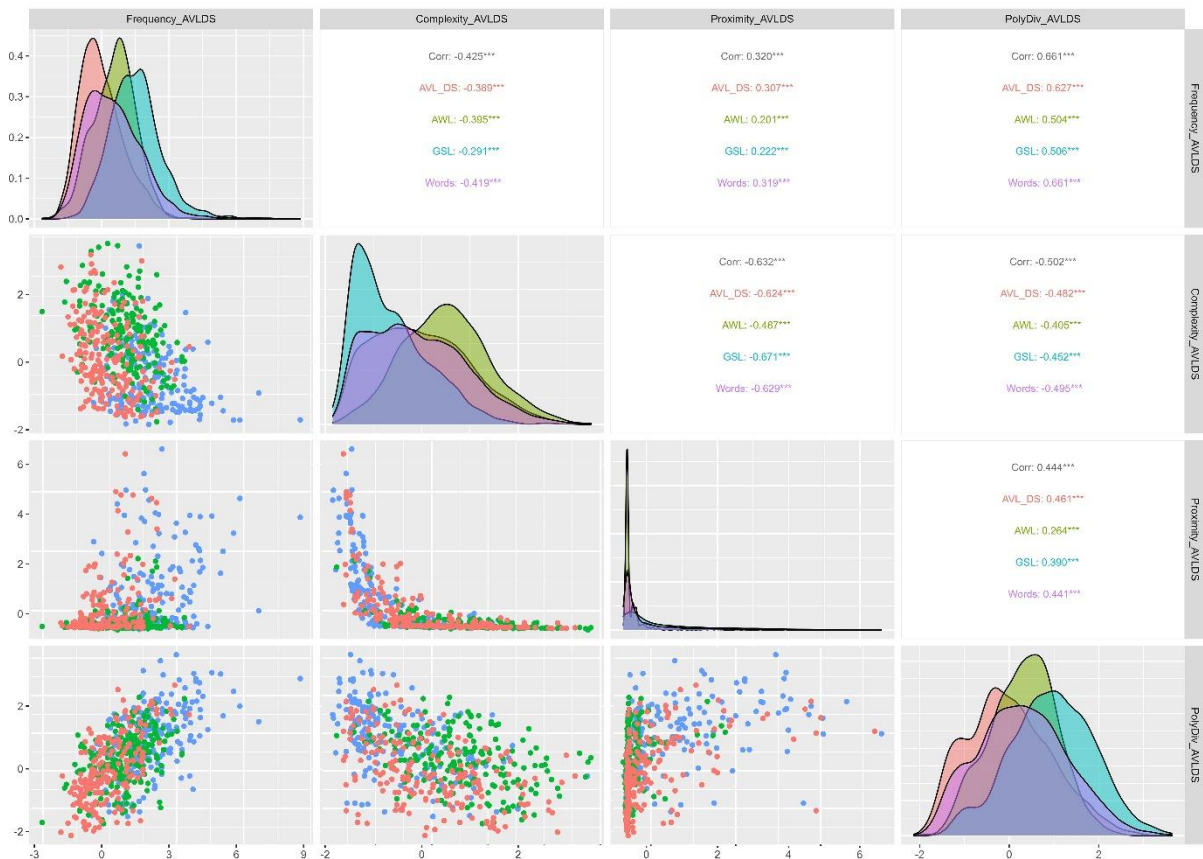


Figure 4. Factor correlations for the AVLDS-reference model estimated on each word sample. Note: Scatterplots below the diagonal contain random 200-word samples while density plots and correlations on and above the diagonal are based on entire word samples. Word samples include AVL_DS (red) Academic Vocabulary List Domain-Specific; AWL (green) Academic Word List; GSL (blue) General Service List; Words (purple) in any list.

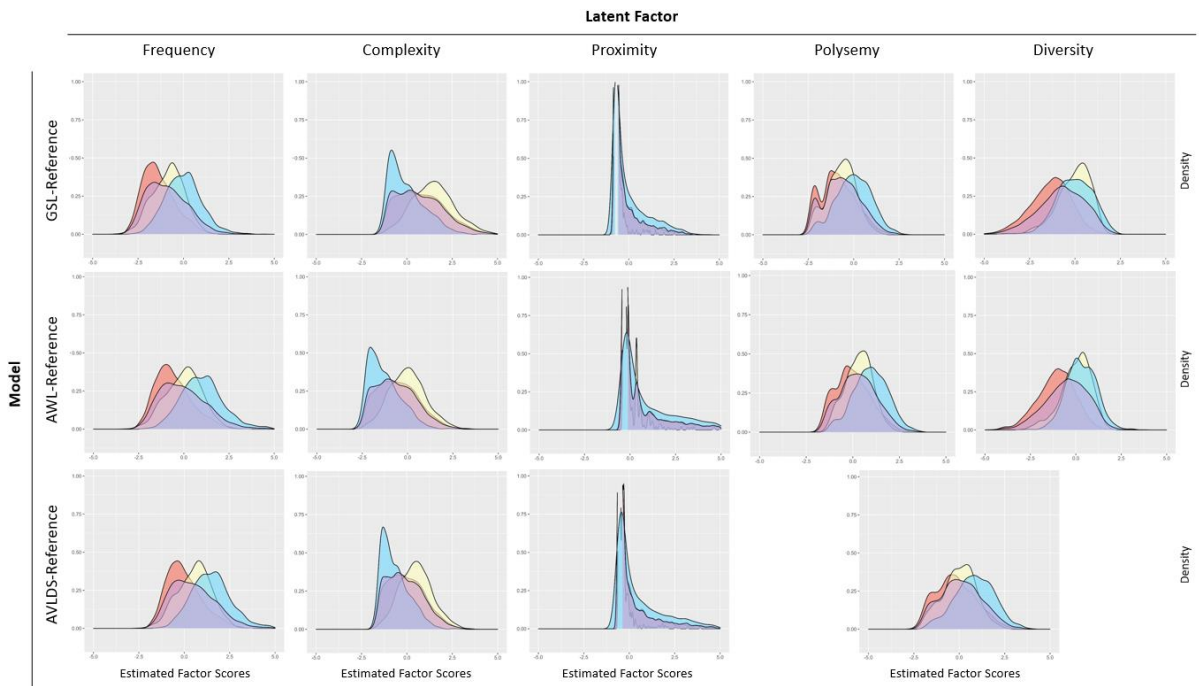


Figure 5. Scaled density plots by reference model and latent factor.

Note: Word list samples are: GSL (blue) General Service List; AWL (yellow) Academic Word List; AVLDS (red) Academic Vocabulary List Domain-Specific; Words (purple) in any word list

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.

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 $\alpha = .91$). 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 <i>M</i> (<i>SD</i>)	Academic vocabulary score ^a <i>M</i> (<i>SD</i>)	GATE <i>M</i> (<i>SD</i>)	SES <i>M</i> (<i>SD</i>)
6 (<i>n</i> = 492)	514.3 (40.1)	32.7 (9.8)	.45 (.5)	.37 (.5)
7 (<i>n</i> = 661)	537.4 (40.5)	35.4 (10.4)	.48 (.5)	.39 (.5)
8 (<i>n</i> = 594)	550.1 (43.0)	37.7 (9.6)	.45 (.5)	.35 (.5)
Total (<i>N</i> = 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.

^aThe maximum score is 50.

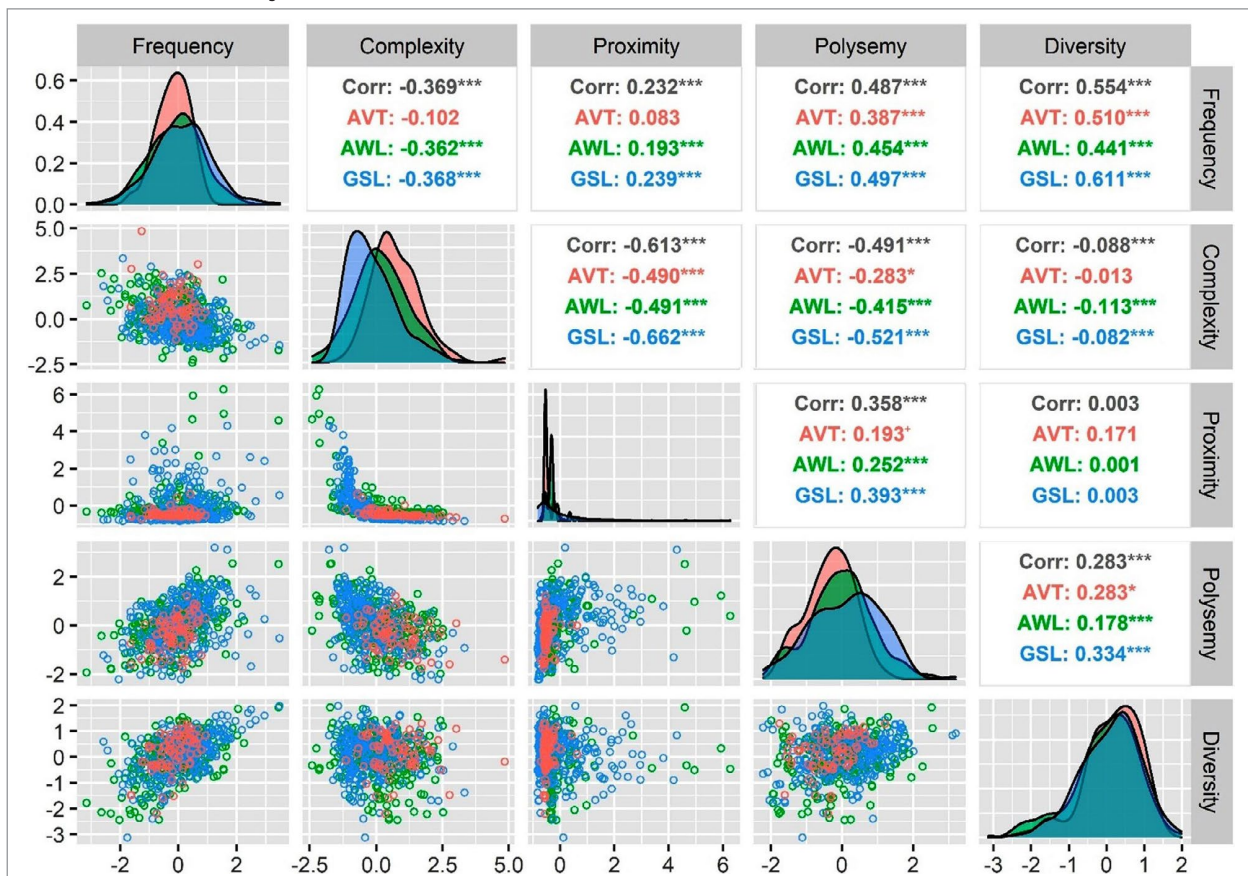
from the AWL (Coxhead, 2000) and seem to represent the class of words on the AWL with respect to word characteristics, as we subsequently discuss in detail. Within-sample internal consistency reliabilities for grades 6–8 ranged from .81 to .93. All Academic Vocabulary Test forms that were developed by 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

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

TABLE 3
Fixed Effects for the Main Effects Models

Fixed effect	Model 1		Model 2		Model 3		Model 4		Model 5		Model 6	
	<i>b</i>	<i>SE</i>	<i>b</i>	<i>SE</i>	<i>b</i>	<i>SE</i>	<i>b</i>	<i>SE</i>	<i>b</i>	<i>SE</i>	<i>b</i>	<i>SE</i>
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

Fixed effect	Model 7		Model 8		Model 9		Model 10		Model 11		Model 12	
	<i>b</i>	<i>SE</i>	<i>b</i>	<i>SE</i>	<i>b</i>	<i>SE</i>	<i>b</i>	<i>SE</i>	<i>b</i>	<i>SE</i>	<i>b</i>	<i>SE</i>
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

Model	AIC	BIC	Deviance	Person side		Item side	
				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 ($\beta = 0.41$, standard error [SE] = 0.08, $p < .001$) and 8 ($\beta = 0.75$, 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 ($\beta = 1.17$, 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 ($\beta = 0.35, SE = 0.13, p < .01$). 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

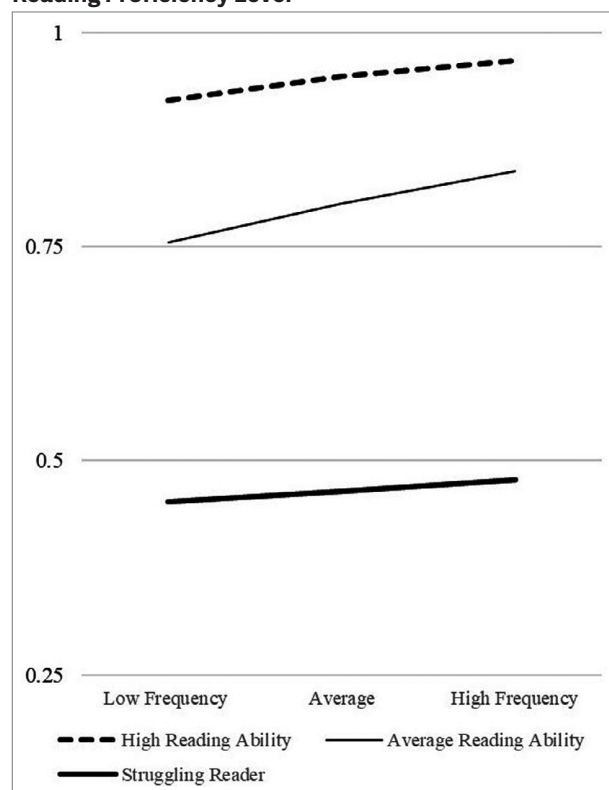
As expected, the combined model findings in model 6 for person characteristics and word features were identical to the results reported for these features separately in models 4 and 5, respectively, because person and word characteristics are not correlated in the design. That is, effects of student 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 ($\beta = 0.07, SE = 0.01, p < .001$), (b) polysemy ($\beta = 0.03, SE = 0.01, p = .002$), (c) complexity ($\beta = -0.08, SE = 0.01, p < .001$), and (d) proximity ($\beta = 0.06, SE = 0.01, p < .001$), 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 ($\beta = 0.07, SE = 0.01, p < .001$) and complexity ($\beta = -0.07, SE = 0.01, p < .001$) 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.

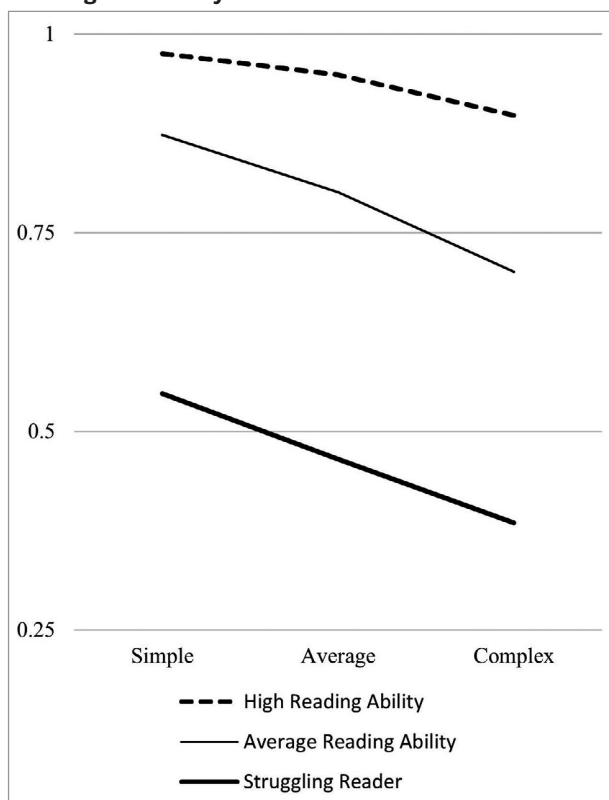
Interaction effects were generally small in their magnitudes. We can conceptualize this difference between variance accounted for in the two sides of the model as indicating that readers who are higher in ability tend to know more words regardless of the features of the words being tested. Although polysemy 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

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



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.

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 between 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 \times Complexity \times Reading Ability \times 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. Dornic (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 (CVRi). *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](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/10.3102/002831215579485>
- 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/10.3102/0028312024002237>
- 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.serp.institute.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., <i>the = th + he</i>)	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 <i>The Educator's Word Frequency Guide</i>	Zeno, Ivens, Millard, and Duvvuri (1995)

(continued)

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

Variable name	Description	Citation
freqband	Frequency groupings from the <i>Oxford Online Dictionary</i> 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., <i>stove/stone</i>)	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., <i>love/dove</i>)	Balota et al. (2007)
phono_n	Raw number of phonologic neighbors (e.g., <i>hear/hare</i>)	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 <i>Oxford Online Dictionary</i>	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 <i>The Educator's Word Frequency Guide</i>	Zeno et al. (1995)

TABLE A2
Factor Score Estimation Beta Weights for Word Features

Variable name	Beta weights for factor score estimates				
	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_lnaposam	0.011	-0.010	-0.002	0.200	-0.048
wordsmyth_lnaposam	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

**Understanding Vocabulary:
Making Sense of What We Measure, Who We Measure, and How We Measure**

Rebecca Knoph

Errata List

Location	Change	Amended version
Page 1 Line 4	Add "to"	...interesting to researchers.
Page 1 Line 5	Add "ed"	...knowledge is assessed...
Page 8 Line 15	Switch "the" and "individual"	...know the individual words...
Page 11 Line 5	Add "not"	...are not time-invariant.
Page 12 Line 7	Add "to"	...related to multiple languages.
Page 18 Line 17	Change to "graders"	...study of 2 nd -6 th graders who...
Page 18 Line 18	Add ")"	..."hilly") more accurately...
Page 22 Line 18	Add "in"	...represented in Figure 3...
Page 22 Line 22	Omit "1"	... Stahl (2003) proposed...
Page 24 Line 12	Change to "L1"	...initial development in L1...
Page 24 Line 12	Change to "L1"	...word to L1 translation...
Page 26 Line 5	Add "of"	...selection of issues...
Page 28 Line 3	Change to "as"	...meanings, as seen in...
Page 30 Line 12	Change to "scoring can be"	...Binary scoring can be...
Page 30 Line 22	Change to "is"	...response is that the...
Page 33 Line 16	Change to 2016	...De Boeck & Wilson, 2016)...
Page 33 Line 18	Omit "person"	...estimate for residual variance...
Page 37 Line 13	Change to "difficulty"	...predict item difficulty for...
Page 38 Line 13	Change to "if"	...estimates (if any at all)...
Page 39 Line 29	Change to ", " and "set"	...as Norway, must... ...protection set by...
Page 43 Line 10	Change to "scraped"	...scraped data from WordNet...
Page 46 Line 15	Change to "though"	...(though see for example...
Page 50 Line 4	Change to "to measure"	...to measure lexical characteristics...