2019

Gaining Access to the Language of Science: A Research Partnership for Disciplined, Discursive Ways to Select and Assess Vocabulary Knowledge

Emily Hayden
Iowa State University, ehayden@iastate.edu

Anupma Singh
Iowa State University

Michelle Eades-Baird
Empire State College

Gaining Access to the Language of Science: A Research Partnership for Disciplined, Discursive Ways to Select and Assess Vocabulary Knowledge

Abstract
To equalize access to science learning across genders and demographic groups, access to the disciplinary language of science is one place to start. The language of science is highly challenging and specialized, and difficulties acquiring this language contribute to disparities in science achievement across diverse student groups. This study used a pre/post design to analyze effectiveness of a brief classroom science vocabulary assessment designed to assess receptive and productive vocabulary knowledge across multiple sections of one seventh-grade science teacher's class. Vocabulary was selected and analysis conducted by an interdisciplinary research partnership, including the science teacher, a literacy specialist, and a scientist. The resulting model presents an assessment that evaluates receptive knowledge and productive use of science language and reinforces vocabulary theory: learning words is incremental and multidimensional, and assessment should address this specialized skill in principled, disciplined ways.

Keywords
disciplinary literacy, vocabulary, science, assessment

Disciplines
Curriculum and Instruction | Educational Assessment, Evaluation, and Research | Educational Methods | Higher Education | Science and Mathematics Education | Vocational Education

Comments
This article is published as Hayden, H. E., Singh, A*, & Baird, M. E. (2019). Gaining access to language of science: A research partnership for disciplined, discursive ways to select and assess vocabulary knowledge. *Reading Horizons*. 2019, 58(2); 1-23. Posted with permission.

READING HORIZONS

Gaining Access to the Language of Science: A Research Partnership for Disciplined, Discursive Ways to Select and Assess Vocabulary Knowledge

H. Emily Hayden, ***Iowa State University***
Anupma Singh, ***Iowa State University***
Michelle Eades-Baird***, Empire State College***

Abstract

To equalize access to science learning across genders and demographic groups, access to the disciplinary language of science is one place to start. The language of science is highly challenging and specialized, and difficulties acquiring this language contribute to disparities in science achievement across diverse student groups. This study used a pre/post design to analyze effectiveness of a brief classroom science vocabulary assessment designed to assess receptive and productive vocabulary knowledge across multiple sections of one seventh-grade science teacher's class. Vocabulary was selected and analysis conducted by an interdisciplinary research partnership, including the science teacher, a literacy specialist, and a scientist. The resulting model presents an assessment that evaluates receptive knowledge and productive use of science language and reinforces vocabulary theory: learning words is incremental and multidimensional, and assessment should address this specialized skill in principled, disciplined ways.

Keywords: disciplinary literacy, vocabulary, science, assessment

Several years ago Brown, Ryoo, and Rodriguez (2010) posed a question and a challenge to researchers in science and literacy education: "If students are struggling to acquire the language of science, how can science education continue to neglect ... language instruction?" (p. 1490). Scientific language is replete with specialized vocabulary (Honig, 2012) to describe, compare, categorize, and explain; and while these features help organize content they also make learning in science highly challenging (Brown, et al., 2010). The complexity of science language and the difficulty of mastery contribute to disparities in science achievement across diverse student groups, and place unique demands on students to learn and use specific disciplinary language orally and in print forms (Gee, 2005). If we hope to equalize access to science content across diverse groups, providing access to the language is a critical starting point because, "oral and written language is the symbol system most often used by scientists ... [it] shapes science ideas and understanding" (Hand, et al., 2003, p. 608).

Perhaps recognizing this need, the U.S. Common Core State Standards (CCSS) (National Governors Association Center for Best Practices & Council of Chief State School Officers, 2010) renewed emphasis on vocabulary learning across disciplines, with science standards that require definitions, as well as ability to determine meanings of key words and domain-specific words and phrases used in specific science contexts. The Next Generation Science Standards (NGSS) (National Research Council, 2011) also emphasize language for communication and learning in science, for: asking questions, defining problems, constructing explanations, and engaging in argument from evidence (Lee, Quinn, & Valdes, 2013). These standards emphasize receptive and productive knowledge of science language: ability to understand science words when encountering them in reading and hearing them in speech, and ability to use science language in writing and speaking. Such multifaceted proficiency with words demonstrates ownership, or the ability to use words in all expressions of language: reading, writing, speaking, and listening (Kamil & Hiebert, 2005; Nagy & Townsend, 2012). How can teachers teach and assess science vocabulary so students can access and develop ownership of science language?

Purpose

This study analyzed a brief assessment designed to capture development of vocabulary knowledge in one science classroom. We chose the science classroom as the setting for developing this assessment because mastery of science language has been identified as highly challenging for many students, and lack of mastery of language in science is particularly detrimental to learning science concepts and content. Because standards emphasized both receptive and productive language, we assessed students' abilities to perceive word meanings receptively and use those word meanings productively by writing connected text (Kamil & Hiebert, 2005; Nagy & Townsend, 2012). Through examination of specific practices in vocabulary assessment within one middle-school science classroom, our goal was to develop an assessment that could capture students' developing ownership of important science language, and thus improve access to science learning. We drew on several decades of research that 1) suggests significant revisions to the ways we teach and assess vocabulary learning, and 2) calls for research that explores and models the varied dimensions of vocabulary learning and knowledge. We asked two research questions: 1) How effective is a brief classroom vocabulary assessment at capturing the development of receptive and productive vocabulary knowledge in one science classroom? and 2) How can varied dimensions of vocabulary learning and knowledge be assessed and modeled?

Framework

Because this research took place in an authentic classroom setting, we wanted to design an assessment that accounted for the variations and accommodations of classroom life. We viewed the classroom teacher as an expert on these particular theory-to-practice translations, and we utilized a methodological framework that specifically focused on working with teachers in real practice settings. Additionally, we considered the complexity of word knowledge when designing the vocabulary assessment, and thus drew on theory that helped us understand what it means to develop ownership of a word. Following are descriptions of both our methodological and theoretical frameworks.

Rigor and Realism

We used Practice Embedded Educational Research (PEER) as a framework for the overall design of this research. Snow (2015) described PEER as a methodology for bringing rigor and realism to educational research, allowing for partnerships with teachers to explore urgent problems of practice. Snow characterized this methodology as emphasizing "interconnections of research and practice rather than the gap between them." (p. 460) recognizing that there are intriguing and challenging problems of teaching practice and research that do not fit analytical models espoused by "basic science" (p. 460). PEER focuses on improving practice by studying how innovations are developed and implemented and evaluating outcomes within a context of problems that are meaningful to classroom instruction and learning. We enacted PEER methodology by working with an expert science teacher to address the problem of developing relevant vocabulary assessment for science that could capture ownership: the ability to use and comprehend selected science vocabulary in reading, writing, speaking, and listening.

Complexity of Word Knowledge

To understand the challenges of developing ownership, we looked to Nagy and Scott (2000), who provided a theoretical framework for understanding the complexity of word knowledge. Nagy and Scott (2000) pushed against the reductionist views of vocabulary learning that many students have experienced, in which vocabulary instruction consists only of lists of words to memorize for weekly vocabulary tests. They argued that knowledge of a word is multifaceted and complex and includes five aspects: incrementality, multidimensionality, polysemy (multiple meanings), interrelatedness, and heterogeneity (what counts as knowing a word varies depending on word type). We used two of those aspects as the theoretical frames. The first aspect is incrementality: knowing a word is a matter of degree, not an "all or nothing" venture. The second aspect is multidimensionality: there are different ways of knowing a word.

An example of incremental knowledge can be illustrated with the word "response," one of the words on the vocabulary assessment. Using the lens of incremental knowledge, students may know "response" has something to do with a behavior or action, but may not know that in science, "response" is connected to some kind of stimulus. Nagy and Scott (2000) cited several teams of researchers who found continual growth in degrees of knowing for particular words (Beck, Perfetti, & McKeown, 1982; McKeown, Beck, Omanson, & Perfetti, 1983; McKeown, Beck, Omanson, & Pople, 1985) as students refine knowledge of word meanings, building expert schema. Knowing a word proceeds by degrees: it is incremental.

Additionally, any word can be known in many forms, including spoken or written, by its meaning, when it is associated with a certain context, by its morphological form, and so on. Knowing a word in different forms is multidimensional knowledge, which can be illustrated with another word on the vocabulary assessment: "abiotic". The word "abiotic" might be recognized and decodable if a student knows the morpheme "bio", because the other morphemes ("a" and "tic") are easily decodable. But knowing how to decode and pronounce "abiotic" and even recognizing "bio" as having something to do with life does not mean a student can use the word correctly in writing or speech. A search of dictionary.com shows "abiotic" is defined as "the absence of life". There are many aspects to knowing a word, and this study aimed to explore both incremental and multidimensional word knowledge.

Literature

Mastering the Language, Making Science Accessible

Extensive research has examined disparities in students' self-concept, identity, and efficacy within science learning. Students perceive their possible success in science very differently depending on their race, ethnicity, and gender; and persistent, significant achievement gaps exist in science test scores between Black and Hispanic student groups and whites (Berends, Lucas, & Penaloza, 2008; Philips, Crouse, & Ralph, 1998). Comparisons of fourth- and eighth-grade cohorts in the Trends in International Mathematics and Science Study (TIMSS) data reveal a significant drop in enjoyment of science among girls as they moved into adolescence (Riegle-Crumb, Moore, & Ramos-Wada, 2010). In the United States, Keiffer, Lesaux, Rivera, and Francis (2009) reported discouraging results on the National Assessment of Educational Progress and other large-scale assessments for particular groups of students. They surmised that "all sophisticated academic tasks, [including] reasoning with scientific information, are mediated by language and literacy skills" (p. 1188).

Narrowing the Scope of Vocabulary Learning

Our lens on the language and literacy skills needed for reasoning with scientific information focused first on vocabulary and classroom discourse levels. The first task for this research was to work with a seventh-grade science teacher, Ann, to make a joint examination of the lengthy vocabulary lists that accompanied each nine-week instructional unit in her science curriculum. A growing body of research strongly suggests that instruction in the United States rarely results in long term learning of vocabulary (McKeown, Deane, Scott, Krovetz, & Lawless, 2017; Nagy & Townsend, 2012; Pearson, Hiebert, & Kamil, 2007), and one of the primary difficulties is that teachers teach too many words at once. The result is that words are not remembered long term, but are forgotten after completion of the associated project or assessment. Lesaux, Keiffer, Faller, and Kelley (2010) asserted that the complexity of word learning requires more intensive instruction with smaller groups of words.

Indeed, years of research on typical ways we teach and assess vocabulary, with long lists of words that are abandoned as each vocabulary test is passed, have found that such an approach does not work (McKeown, et al., 1985, Stahl & Fairbanks, 1986). The abundance of specialized vocabulary in science (Brown, et al., 2010; Gee, 2005; Honig, 2012) necessitates intensive focus, and some researchers have equated science vocabulary knowledge with science conceptual knowledge (Pearson, 2010; Pearson, Moje, & Greenleaf, 2010). However, current reviews (Jagger & Yore, 2012; Nixon, Saunders, & Fishback, 2012) indicate that very little research has explored language in science and the ways it supports meaning-making. Researchers across many years, from Dale (1965) to Pearson and colleagues (2007), have called for explorations into the difference in learning vocabulary from disciplinary texts versus stories, and Keiffer et al. (2009) called for research to operationalize academic language that is specific to disciplines, and examine how instruction in such language could impact students' performance in disciplinary instruction and assessment.

Conceptualizing Word Learning

Starting very broadly, Dale (1965) conceptualized vocabulary learning as proceeding in stages. Learners may begin with no knowledge of a word at all, and proceed in incremental steps of partial knowledge before they finally gain full knowledge of a word. Along the way, they may be able to recognize a word but not know its meaning, and may be able to recognize a word and its meaning in context, but not outside of that context. After conducting numerous studies attempting to quantify the vocabularies of students at various ages, Dale recommended that future research focus on specialized vocabularies versus vocabulary counts, with an emphasis on levels of growth in knowledge, and distinguishing between “acceptable and inadequate definitions” to determine growth in understanding of concepts (p. 901). Later, Beck, McKeown, and Omanson (1987) argued that growth in word knowledge could instead be thought of as a continuum and that, when thinking about how word knowledge contributes to concept knowledge, there is more to consider. One piece of the puzzle is level of knowledge, whether or not a learner has seen the word before or has knowledge of it in a certain context. To demonstrate deeper understanding or ownership of a word a student would need to be able to give examples of correct use, extending to decontextualized and metaphorical uses.

Beck et al. (1987) were the first to theorize the sorting of words into three Tiers for instruction, and this sorting can be a preliminary mechanism for operationalizing words for learning. High-frequency words constitute Tier-1 and need very little instruction; highly specialized words constitute Tier-3 and need context-specific teaching. Inhabiting the middle ground are Tier-2 words: general utility words that appear across domains and that call for the greatest amount of focus in vocabulary instruction. The words selected for this assessment did not all fit the Tier-2 description of general-utility words that appear across domains; they were selected because they called for the greatest amount of focus in the science classroom.

Operationalizing Vocabulary Selection

There are numerous frameworks, word lists, and instructional recommendations for further operationalizing vocabulary instruction in schools. The most widely used of these are based on Marzano (2004), who reviewed U.S. standards documents from 13 content areas to produce a list of nearly 8,000 subject-specific vocabulary words organized into four grade ranges. Marzano’s list included 225 science words at the 6–8 grade level, raising questions about cognitive load. Whether related to a particular unit of disciplinary study or to general academic vocabulary, typical curricular word lists can be quite lengthy. Fisher and Frey (2014), Fisher (2007), Lesaux et al. (2010), and Nagy and Townsend (2012) all cautioned against choosing too many words for students to successfully integrate, and Snow, Lawrence, and White (2009) built their Word Generation program integrating only five “all-purpose” (p. 326) academic words per week of study. The question of cognitive load and how many words to focus on for retention and ownership that Pearson et al. (2007) raised remains open and unsolved.

In a Practice Report reviewing research on adolescent literacy for the U.S. Institute of Education Sciences, Kamil et al. (2008) cited strong evidence for specific vocabulary instruction in science and called for selection of focus vocabulary words based on “how important the words are for learning in a particular discipline” (p. 15). These authors also cited the Tier framework (Beck, et al., 1987) for vocabulary selection; however, they cautioned that decisions should be based not only on Tiers but also on importance of the word to content learning. We selected words for our assessment in partnership with the science teacher, and we considered not only the recommended focus on Tier-2 words (Beck et al., 1987) but also the suggestion of Kamil et al. (2008) that selected words be those that are important to learning in a particular discipline, in this case, high-utility words for learning in science. Words selected for this assessment include Tier-3 words such as “photosynthesis” and Tier-2 words such as “response” that are central to understanding important and recurring science concepts, but also occur in other contexts with different meanings.

Research on Vocabulary Instruction with Assessment

Fisher, Grant, and Frey (2009) asserted that vocabulary instruction should be intentional and transparent, and should present selected words in ways that are usable and personal. Assessment should follow these same caveats, as much as possible. Several large-scale studies have explored the impact of instructional interventions focused on carefully selected, broadly useful words, and have included assessment to evaluate student learning and thus the effectiveness of the interventions.

Snow, et al. (2009) focused on providing students in sixth, seventh, and eighth grades multifaceted ways to learn and use 120 selected "all-purpose" academic words when they designed Word Generation, a weekly vocabulary intervention. They found significant positive correlations between student participation in regular intervention activities and performance on posttest measures of vocabulary as well as on state assessments. However, their posttest was lengthy, with 48 items, making it potentially cumbersome for regular classroom use. There were also significant differences by gender, with boys outperforming girls, and vocabulary for Word Generation was selected from broadly identified general academic areas not localized to science.

In a study with sixth-grade students, McKeown, Crosson, Moore, and Beck (2018) used a progression of assessments to explore growth in knowledge of 99 general academic words taught across two academic years in their Robust Academic Vocabulary Encounters (RAVE) intervention. Assessments explored both proximal and distal outcomes, and the process was quite lengthy. A word knowledge assessment gathered data on all 99 words using fill-in-the-blank format, a lexical decision task utilized three sets of 20 words and nonwords, and a comprehension assessment required reading two passages of approximately 300 words each. Additionally, a morphological awareness task was conducted with a subset of the total participants in the RAVE study, and a standardized, large-scale distal reading comprehension measure administered pre/post. Results across this progression of assessments were mixed and modest, and were not disaggregated by gender. The authors did provide direction for future interventions, including some aspects that we included in our study: introducing words in typical contexts and "directly confronting various senses and uses" of a word (p. 610).

These lengthy interventions and assessments developed by Snow and colleagues (2009) and McKeown et al. (2018) focused on learning and assessing a much larger and broader set of words than we sampled. Interventions were highly structured, sometimes even scripted, and McKeown and colleagues in particular intended to explore larger questions about the nature of vocabulary learning within the broad body of academic knowledge. Our focus was on learning the language of a specific and highly challenging discipline: science. Our decision to work in partnership with a science teacher, whose expertise we considered vital to the selection and assessment design process, meant that our assessment was more concentrated and our outcomes were different from the broad interventions and analyses conducted by other researchers. We sought to describe the process of building science vocabulary knowledge in a classroom setting, and to describe methods of selection and assessment that could translate theory on the incremental and multidimensional nature of vocabulary learning into classroom practice. Our focus on science addressed an area where achievement gaps based on race, ethnicity, and gender are well documented.

Brown et al. (2010) also focused on science language learning and achievement gaps. In a mixed-methods study with ethnically and linguistically diverse students, they found that when students were allowed to use everyday language to communicate their understanding of science concepts, they were more likely to develop rich and correct understandings of those concepts. Their assessment included extensive discourse analysis of participants' use of focus words in post-intervention interviews, and pre/post data on 18 multiple-choice and 10 open-response items, with a 7-point rubric to score the open-response items. These researchers surmised that teaching science concepts in naturalistic ways that connected to real experiences, and analyzing conceptual learning without requiring the unnecessary formalism of science language, supported students in accessing complex science concepts. Their findings reinforce the naturalistic instructional methods the teacher used and the way we assessed the productive writing of participants in this study.

Receptive, Productive, Practical Assessment

Outside of the research assessments already described, current, broadly used, large-scale vocabulary assessments for middle-grade students such as the Gates MacGinitie Reading Tests (e.g. MacGinitie, MacGinitie, Maria, & Dreyer, 2000), tap into receptive, passive understanding of word meanings through multiple-choice items with no measure of productive knowledge: the ability to use words correctly (Nagy & Townsend, 2012). What is needed are practical classroom assessments that can provide a picture of students' ownership of words through both receptive and productive modes. To be useful for teachers, such assessments would abbreviate the lengthy analysis utilized in Word Generation (Snow et al., 2009) and RAVE (McKeown et al., 2018) while also avoiding the extensive discourse analysis used by Brown and colleagues (2010) but would still focus on specific disciplinary language necessary for access to content.

Scott, Hoover, Flinspach, and Vevea (2008) developed multiple-level vocabulary assessments that drew heavily on incrementality (word learning is a matter of degree) and multidimensionality (there are many aspects of word knowledge, including part of speech and semantic and morphological family) (Nagy & Scott, 2000). They selected words from topics students study, identified through conversations with fourth-grade teachers who identified 21 novels and four textbooks used by fourth-grade students in their region of the United States. From these textbooks and novels, Scott and colleagues identified 30,000 words they judged would be difficult for fourth-grade students, eventually culling the list to 5,000. For each word, they developed a "testlet": four related questions for each word (Thissen, 1989). They developed fiction and nonfiction forms, each incorporating only words from their list that appeared in two or more of the reviewed texts. Words chosen represented a range of grammatical forms (noun, adjective, etc.) and initially each assessment tested 50 words. This number was reduced to 36 to keep assessment time to 15 minutes. The vocabulary testlet format, with multiple related questions for each word, was piloted with 380 fourth- and fifth-grade students, including 46% who were English language learners or bilingual students coming from 26 different home languages. Analysis found the testlet format to be psychometrically sound and reliable, and to have high discrimination scores, indicating ability to distinguish between students with differing abilities.

If teachers hope to teach science vocabulary in ways that will help students access important conceptual knowledge while also honoring decades of research on vocabulary learning, they need to revisit and repurpose their approach. Focus words should be those that are optimally useful for science, the number of focus words should be limited, and assessment should reveal the development of the related conceptual knowledge. In order to concentrate our approach in these ways, we started with the model developed by Scott et al.

Methodology

Context

The seventh-grade science teacher we worked with for this study, Ann, was selected because she possessed many markers of expertise that gave weight to her instructional decision-making. These markers included leadership experiences both at her school and within her school district in both science and language arts, and multiple participant experiences in select national on-site science study and research opportunities. Ann viewed her content holistically and made instructional decisions accordingly. She thought about standards and outcomes: what students would be asked to know and do in her science classroom, as well as real-life experiences her students might have that she could connect to science content and how to explain concepts in terms her students would understand. These markers, combined with extensive teaching experience (more than 20 years) and background in both science and English/Language Arts instruction made Ann's practices worthy of study (Dyson & Genishi, 2005; Yin, 1994).

To study how practices of vocabulary assessment could be innovated, the first researcher asked Ann to choose 10 words per semester she believed were essential learning for her students during the year. These became focus words in this study. We accepted Ann's appraisals of the curriculum, basing our trust on the particular and extensive markers for expertise she possessed in content knowledge (science) pedagogical knowledge (teaching) and pedagogical content knowledge (the intersection of content and teaching) (Shulman, 1987). Although she continued to teach her entire curriculum, including all 30+ words per 9-week unit that were emphasized in her curriculum, Ann incorporated special emphasis on these focus words into her instruction in naturalistic ways, by connecting understanding of concepts to real experiences and then attaching language to those concepts (Pearson, 2010).

Instructional methods. Ann used multidimensional instructional methods that included reading, writing, listening, and speaking. A few of her techniques are described here, all of them undergirded by her belief that "Students have to be able to transfer their knowledge of words to using it in some other way—writing, speaking, etc.—in order to demonstrate that they can use it flexibly. Then we have some evidence that they've got it!" In a nod to traditional vocabulary learning, she incorporated regular definition activities, but in a new ways. Three days a week, Ann spent 5 minutes reviewing words selected randomly from the current science topics, including the focus words. She provided the definition, students wrote down the corresponding word, and then graphed their personal mastery progress in their science notebooks. This was a low stakes activity with no grade attached, and Ann did not require perfect spelling or grammar in students' responses for this activity. She compiled an overall performance graph for each class, but students monitored their own progress individually (Hayden & Eades-Baird, 2016).

The words used for these quick review activities were connected to real-life experiences during classroom instruction. For example, Ann connected the word "sublimation" to shrinking snow forts students built after the latest snowstorm, and connected the word "divergent," used in science to discuss boundaries between tectonic plates, with the math term "divide" (Hayden & Eades-Baird, 2016). Words were also used productively, in science journal writing that included informal summaries of daily class learning, narrative writing with topic and focus words, formal write-ups after inquiry activities, and visual arts activities such as postcard or poster-making to demonstrate ownership of science concepts.

In another nod to typical instructional practice, Ann used “Ticket Out” activities to review learning at the end of class time, but in ways that demonstrating incremental levels of knowledge. The most basic Ticket Out activities asked students simply to provide an oral definition for a word or use a word in a complete sentence, but the more advanced activities asked students to describe new science concepts learned that day or to illustrate and label a new science concept. Because of her language arts background, Ann made extensive and exceptional use of opportunities to connect literacy and language to science concepts, and she provided multiple ways that students could demonstrate their concept knowledge productively, not just receptively.

Assessment Development

. Our goal was to develop a classroom instrument that assessed science vocabulary in usable, naturalistic ways that allowed students to use everyday language to communicate their understanding (Brown et al., 2010) and that also drew on “definitional and contextual information” (Blachowicz, Fisher, Ogle, & Watts-Taffe, 2006, p. 528) as well as the multifaceted nature of vocabulary knowledge. Mindful of cognitive load and the realities of time in classroom teaching, we focused our assessment on 10 words per semester. We followed recommendations of Lesaux et al. (2010), Beck et al. (1985), Stahl and Fairbanks, (1986) and others to assign fewer focus words and use them extensively in classroom discourse.

We modified the final item on the Scott et al. (2008) testlet, which asked for part of speech of the focus word. We included this item initially, but dropped it when Ann reported it did not provide useful information for her science instruction. The decision to drop the grammatical form item from the original testlet format (Scott et al., 2008) is in keeping with the Practice Embedded Educational Research (PEER) (Snow, 2015) caveat to listen to what teachers say, “so that through a process of … design more easily usable versions of evidence-based programs will emerge.” (p. 462). We also agreed with the critique provided by Nagy and Townsend (2012) that vocabulary assessments lacking a productive element failed to assess depth of students’ knowledge. In response, we substituted a sentence-writing item for the grammatical form item to address this gap.

We modified our assessment in another important way. While Scott et al.’s (2008) vocabulary selection contrasted with typical assessments, which contain arbitrarily chosen words (Pearson, et al., 2007) we localized selection even more, to seventh-grade science content. With Ann, we collected lists of words provided by her school district for each instructional unit and used consideration of science standards and local/state assessments in conjunction with Tier frameworks (Beck, et al., 2002) to select words. Thus, we addressed concerns raised by Kamil et al. (2008) who urged that words be considered for usefulness as well as frequency and utility.

Participants

Students in Ann’s seventh-grade science classes (n=126) provided the data for this study. Ann taught in a lower-middle class, moderately educated Midwestern U.S. city (census.gov). The school served 754 students in grades 6–8, where 81.7% of students were white, .3% American Indian, 3.2% were Black/African American, 1.3% Asian, 6.6% Hispanic/Latinx, .1% Pacific Islander, and 6.8% were two or more races.

Twenty-six percent of students qualified for free/reduced lunch. The sample was nearly evenly distributed across gender lines, with 64 males and 62 females.

Data Sources

Pre/post assessment data was collected during the spring semester of seventh grade science, when unit topics included living organisms, behavioral inquiry, and environmental science. The assessment was designed to examine students' growth in ownership of the 10 focus vocabulary words (Table 1) Ann selected for special emphasis in her instruction. It contained both binary items (Questions 2 and 3) and ordinal scales (Questions 1 and 4) for multipoint items that resulted in nonlinear data (see Table 2). Two of the four questions on the assessment were re-evaluated for inclusion in the analysis. While Question 1 provided a baseline for students' perceived knowledge levels, especially at pretest, this information did not contribute materially to our research questions and was dropped from the analysis. Question 4 asked students to write a sentence using the focus term. Sentences were scored using a three-point system, where "0" indicated no sentence or the sentence used the focus vocabulary incorrectly, "1" indicated partial understanding of the term demonstrated by its use in the sentence, and "2" indicated complete knowledge of the term for seventh-grade level.

Table 1

Vocabulary Terms, Seventh-Grade Science, Second Semester

Habitat
Photosynthesis
Stimulus
Adaptation
Species
Response
Instinct
Ecosystem
Virus
Abiotic

Sentence scoring was completed by a literacy researcher and former classroom teacher, and a PhD candidate with degrees in biology and education and 13 years of experience teaching middle-school biology and chemistry. For example, a sentence for "response" needed to include both cause and effect to earn a full score of 2, and needed to demonstrate knowledge that a response is involuntary, different from a reaction. A sentence such as "When I felt a bug on my arm I reacted from the stimulus to swat it off my arm" would be scored at 1 since it demonstrated some understanding of a "response" as connected to a stimulus, but characterizes it as a planned reaction rather than involuntary. A sentence such as, "When someone claps right in your face you blink" would earn a score of 2 since it demonstrated more complete seventh-grade knowledge of "response" as involuntary and including both a cause (stimulus) and effect.

While this finely grained rubric was used for initial scoring of the written sentences, responses to Question 4 were recoded for our analysis. Scores of 0 remained at 0 for the analysis, and scores of 1 (partial knowledge) and 2 (complete knowledge) were collapsed and entered as 1 into the analysis. This was done to meet the parameters for this study, which focused on analysis and modeling of a pre/post classroom assessment, developed through a teacher–researcher partnership, that assessed both incrementality and multidimensionality of word learning in science and included a productive item as well as receptive items. Future analysis will focus specifically on the differences between a sentence scored at 1 (partial knowledge) and a sentence scored at 2 (complete knowledge). For this research, we focused our analytical lens on the ability to produce meaning, even partial meaning, with a written sentence, and on the capacity of the assessment to capture three aspects of vocabulary ownership: Awareness, represented by Question 2, Receptive Ownership, represented by Question 3, and Productive Ownership, represented by Question 4 (Table 2).

Table 2

Pre/Post Assessment Example for Vocabulary Term "Habitat" and Scores for Each Response

Item	Scoring
Circle the letter of one answer.	
How well do you know this word?	
I've never heard this word before	0
b. I've heard this word, but I don't know what it means	1
c. I think I know what this word means, or what it is related to	2
d. I know this word and can use it correctly	3
If you chose answer "a" from question #1, please go on to the next word.	
If you chose "b," "c," or "d" from question #1, then continue.	
I think the word may have something to do with:	
a person	0
b. organizing	0
c. environment	1
d. food	0
I think this word means:	
the way the sun warms the Earth	0
b. a natural home of an animal, plant, organism	1
c. gathering food	0
d. placing things in order	0
4. Write a sentence using this word:	0, 1, or 2

Data Screening

Once our data set was established, it was treated for missing values using multiple-imputation. Next, the data set was tested and treated for univariate and multivariate normality. Under both pre/post conditions, the word "abiotic" was identified as problematic across all model configurations and ultimately dropped. It was the most dichotomous word in the set, with all but two students earning a score of 0 on the written sentences at pretest (no knowledge demonstrated), and 116 students earning a 2 (more complete knowledge demonstrated) at posttest. This unique word, which seems to be both Tier-3 and important to understanding content, caught the full attention of students, or was taught determinedly. In any event, no other word produced a pre/post split this extreme.

Analysis

We first analyzed pre/post scores on the vocabulary assessment with paired sample t-tests. Then, confirmatory factor analysis (CFA) identified the measurement model that explained the covariance among the observed vocabulary items using Mplus software and the weighted least square with mean and variance adjusted (WLSMV) estimation method, widely recommended for analyses of categorical (e.g., binary) data (Brown, 2015). Four factor models were specified, including: a) a unidimensional model, which assumes that all covariance in science vocabulary acquisition can be explained by one common factor, b) a correlated three-factor model, c) a second-order factor model, and d) a bifactor model. The bifactor model included individual aspects of ownership: awareness, (Question 2, which asked students to select a one-word/partial-knowledge definition); receptive ownership (Question 3, which asked students to select a sentence-length definition); and productive ownership (Question 4, which required sentence-writing). General vocabulary acquisition was represented in the bifactor model through overall analysis of responses to all items. This is referred to as the general factor below.

A bifactor measurement model specifies that for a given set of item responses, correlations among items can be accounted for by: (a) a general factor representing shared variance among all the items and (b) a set of group factors where variance over and above the general factor is shared among subsets of items presumed to be highly similar in content (Rodriguez, Reise, & Haviland, 2016). Based on the theoretical framework used in this study, we wanted to test incrementality (word learning is a matter of degree) of science vocabulary learning and multidimensionality (there are many aspects of word knowledge). Questions 2, 3, and 4 focused especially on the theory of incrementality, while Ann's decision-making in selection of vocabulary words and her instructional methods acknowledged the multidimensionality theory.

Model fit was evaluated using multiple indices, including chi-square ($\chi 2$), Comparative Fit Index (CFI), Tucker Lewis Index (TLI) , RMSEA, and standardized root-mean-square residual (SRMR). RMSEA values below .08, CFI and TLI values equal to or greater than .95, and SRMR values equal to or less than .05 are preferred for excellent model fit (Hu & Bentler, 1999). For nested models, the $\chi 2$ difference test was also used to evaluate model fit. A significant $\chi 2$ difference test indicates that the more constrained model (with more degrees of freedom) provides significantly worse fit to the data than the less constrained model (with fewer degrees of freedom). Following the measurement model specification, structural equation modeling (SEM) tested the extent to which latent constructs of awareness, receptive ownership, and productive ownership from the CFA predicted vocabulary knowledge.

Results

As this study included data from one science teacher's classroom, we interpreted results with care. Our focus was not on broad generalization, but instead on exploring the usefulness of the assessment for capturing growth in vocabulary knowledge across varied dimensions, using both receptive and productive items. We also explored how to best model the results.

Table 3

Descriptive Statistics and Reliabilities for Vocabulary Knowledge

	Mean Score Pretest	Standard Deviation Pretest	Mean Score Posttest	Standard Deviation Posttest	Alpha Reliability Posttest
	n=126		n=126		n=126
Awareness (Q2)	0.54	0.21	0.83	0.18	0.61
Receptive Ownership (Q3)	0.59	0.22	0.88	0.18	0.75
Productive Ownership (Q4)	0.55	0.22	0.85	0.19	0.72

Usefulness of Pre/Post Assessment

Paired sample t-tests for pre/post results were significant at .000, with a large effect size (2.144) at posttest. Since improvement is expected with any instructional intervention, this finding is not especially compelling from a research perspective. Of greater interest was that no significant differences were found between performance of males and females in our virtually equally divided sample. These equivalent outcomes are of interest, because previous research in science vocabulary instruction has found less robust performance for females than males on similar assessments (Snow et al., 2009) or has not disaggregated results to explore gender differences (McKeown, et al., 2018) even though TIMSS data revealed significant gender differences in uptake and enjoyment of science in middle-grade years (Riegle-Crumb, et al., 2010). This vocabulary assessment appeared to be sensitive, perhaps even equitable, to capturing significant growth in receptive and productive vocabulary knowledge. Ann's disciplined selection and discursive instruction of science vocabulary, combined with the vocabulary assessment evaluated here, could provide more equitable science instruction that improves accessibility of science content for all students.

Many replications with much broader samples will be needed to determine if these instructional and assessment methods are indeed statistically powerful, but the results of this limited study indicate that such further studies are warranted in order to uncover equitable methods of science instruction. This assessment was highly useful as a measurement of growth in vocabulary-receptive and -productive ownership in Ann's seventh-grade science classroom, where instructional methods included careful selection of vocabulary words based on importance to content learning and Tier, and highly discursive instruction incorporating disciplinary language into every aspect of science inquiry.

Modeling Results

This study also sought to explore modeling of a science vocabulary assessment designed to capture varied dimensions of vocabulary learning and knowledge. Our analytic exploration focused on a best-fit model, finding this to be a bifactor model where the main explanatory factor was overall vocabulary general factor, with additional uncorrelated second-order subskill factors of awareness, receptive ownership, and productive ownership. Descriptive statistics for subscales, including means, standard deviation, and Cronbach's alpha reliability estimates are reported in Table 3. Reliability estimates for each aspect of vocabulary ownership were acceptable: awareness (.61) receptive ownership (.75) and productive ownership (.71). A comparison of the three models is shown in Table 4. Model statistics are shown in Table 5.

Table 4

Assessment of Model Fit Under WLSMV at Posttest, 36 items

Model	# Estimated Parameters	Chi-Square Value	Chi-Square Scale Factor	Chi-Square DF	Chi-Square p-value	CFI	RMSEA Estimate	RMSEA Lower CI	RMSEA Higher CI	RMSEA p-value
Model 1, Word-Factor Graded Response	-	-	-	-	-	-	-	-	-	-
Model 2, One-Factor Graded Response	110	1072.432	1.000	594	0.0000	.813	.078	.070	.085	.000
Model 3, Higher-Order Graded Response	119	880.415	1.000	585	0.0000	.885	.061	.053	.070	.015
Model 4, Bifactor Graded Response	146	744.884	1.000	558	0.0000	.927	.050	.040	.059	.493

Note: The latent variable covariance matrix (PSI) is not positive definite. The model covariance matrix is not positive definite.

Table 5

Standardized estimates of fit using WSLMV pretest and posttest

	χ2	df	χ2/df	CFI	TLI	RMSEA	WRMR	Good Fit
Post Test								
	1438.908	325	4.42	0.877	0.867	0.117	2.56	-
Second-Order Model	563.156	323	1.74	0.74	0.71	0.077	1.54	-
	433.507	323	1.34	0.88	0.87	0.052	1.14	Yes
Bi-Factor Model	381.346	300	1.27	0.91	0.87	0.046	0.99	Yes
Pre Test								
	1141.244	325	3.51	0.843	0.830	0.141	2.48	-
Second-Order Model	587.519	323	1.81	0.93	0.92	0.114	1.66	-
	436.225	323	1.35	0.97	0.96	0.075	1.26	-
Bi-Factor Model	440.191	300	1.46	0.97	0.97	0.06	1.21	Yes

The data were first fitted to a unidimensional (vocabulary knowledge) model, with fit less than satisfactory (Ullman, 2006). In Table 4, Models 1 and 2 represent inadequate fit while Models 3 and 4 represent good fit. Since the CFI and TLI should ideally be 0.95 or greater and the root mean square error of approximation (RMSEA) 0.06 or less, the bifactor model represents the best way to model the data collected with this vocabulary assessment (Table 4, Table 5). Table 6 displays estimated item loadings for the one-factor model for post- and pretest. The average loading for pretest is 0.57, ranging from 0.165 to 1. The average loading for posttest is 0.65, ranging from 0.31 to 1. Both average item loading and minimum value of item loading increased in posttest results in the bifactor model.

Table 6

Posttest Results: Demonstration of Correlated Factor Model Using a Subset of Assessment Items

Item Loadings

Item	Awareness	Receptive Ownership	Productive Ownership
Habitat	1		
Photosynthesis	0.535		
Stimulus	0.811		
Adaptation	0.541		
Species	0.924		
Response	0.58		
Instinct	0.874		
Ecosystem	0.655		
Virus Habitat	0.305		
		1	
Photosynthesis		0.744	
Stimulus		0.88	
Adaptation		0.687	
Species		0.664	
Response		0.79	
Instinct		0.872	
Ecosystem		0.865	
Virus Habitat		0.721	
			1
Photosynthesis			0.736
Stimulus			0.812
Adaptation			0.76
Species			0.766
Response			0.615
Instinct			0.64
Ecosystem			0.767
Virus			0.656

These findings are of significance theoretically, as they imply that knowledge of science vocabulary is a specialized skill and should be assessed at various levels: incrementally, and including receptive and productive aspects. Pearson, et al. (2007) described the need for research with explicit attention to these types of distinctions in aspects of vocabulary knowledge. This study provides a limited first step.

Discussion

Assessing Science Language

The findings of this study reinforce the notion that scientific language is complex and multifaceted and should be assessed in ways that give students varied opportunities to express their ownership of important and carefully selected vocabulary. Based on the large effect size found here, we can posit that a good assessment of vocabulary will include the types of items we included on this assessment: multiple-choice items to narrow down definitions as in Questions 2 and 3 (Scott et al., 2008), and a productive sentence-writing element (Nagy & Townsend, 2012). This assessment met our goals of including definitional and contextual information, and of addressing incremental and multidimensional aspects of word knowledge through both receptive and productive items. Questions 2 and 3 provided students with scaffolded opportunities to activate schema for a word through progression from choosing a one-word definition to choosing a sentence-length definition. This sequence captures the incremental nature of vocabulary knowledge: that word learning progresses by degrees and is not necessarily an "all or nothing" proposition. It also addresses multidimensionality by activating semantic and syntactic networks as a way to access word meaning. Questions 2 and 3 are receptive in nature, and Question 4 provides an assessment of productive knowledge: ability to use a word correctly in written language. These are essential aspects of ownership: the ability to understand a word in reading and use it correctly in writing.

Multidimensionality, incrementality, definitional and contextual information, and receptive and productive knowledge are all essential aspects of vocabulary ownership. These multifaceted ways of knowing a word are not captured by vocabulary assessments that ask only for receptive knowledge. Likewise, assessments that do not tap into semantic networks and syntactic support for a word, or do not provide ways for students to demonstrate their incremental knowledge are insufficient. If we accept Pearson's (2010) view of vocabulary knowledge as equivalent to conceptual knowledge for science: "a label for an idea you carry around in your head" (n.p.), then assessing word knowledge in ways that do not provide opportunities for students to demonstrate multifaceted understandings will fail to capture the rich contextual knowledge that can be developed when students truly "own" words in all their richness of meaning.

Two additional features differentiated this assessment of vocabulary learning from others (Scott et al., 2008; Snow et al., 2009; McKeown et al., 2018). One is the inclusion of sentence writing as a productive outcome. The other is that the words were selected purposefully for science, utilizing powerful instructional tools: Ann's expert disciplinary knowledge and the Tier framework. The Tier framework alone is not sufficient for vocabulary selection; it must be combined with extensive disciplinary knowledge that considers content, standards, and state and local assessments. This combination provides further support for acknowledging the differences between content area literacy and the more nuanced inquiry involved in disciplinary literacy (Shanahan & Shanahan, 2008).

Thinking About and Using Science Language

Ann's instruction provided opportunities for students to hear and use focus vocabulary during classroom discourse. She intentionally integrated focus words into discursive activities that accompanied inquiry-based science instruction, weaving the use of analytical tools such as etymology and morphology into daily discussions (Hayden, Eades-Baird, & Singh, 2019). This instruction did not develop word knowledge in isolation, but instead focused on developing ownership of the words over time, through practice and use within the daily life of the science classroom and science inquiry.

We were interested to discover that while Question 4 asked students to "Write a sentence using this word" what students usually did was write a description of the concept instead, and often did not actually use the focus word in their sentence. During scoring of this item we, like Brown and colleagues (2010), accepted students' use of everyday language to communicate their understanding of the science concept attached to the word (Pearson, 2010). We did not restrict our notion of correctness to the use of the formal label. This was true in Ann's instructional methods as well. Like Ann, we were less interested in grammatical or syntactic forms and more interested in students' developing ownership of a word, and we determined this to mean that they could understand the concept connected to the focus word well enough to describe their rich and correct understandings in their own informal language.

Ann accepted students' everyday language as well during her instruction, and she shared examples of the incremental growth in concept knowledge she observed in their sentence writing (Table 7). Sentences written in science journals showed progression from incorrect knowledge to more complete knowledge with increments of partial knowledge along the way, much as Dale described in 1965. Some sentences even extended to decontextualized and metaphorical use, an indicator of ownership described by Beck and colleagues (1987) when they first theorized the concept of Tiers as a way to select focus words for instruction. Thus, Ann's instruction appeared to develop students' word knowledge in ways that prove again the wisdom of early thinking about vocabulary learning, while the vocabulary assessment provided an outlet for students to demonstrate their ownership through use of words in ways that revealed their understanding of the concept behind the word as well. Pearson's (2010) notions of the inextricable links between science conceptual knowledge and the vocabulary that labels it were supported by our findings.

Table 7

Incremental Growth in Sentence Writing

Focus Word	Moving From Incorrect or Less Complete Understanding to More Complete Understanding
Host	I'm a host. All people and animals are hosts.
	A dog is a host for fleas.
	He was the host of the virus.
	I am a host. Viruses use me for energy
	A virus or bacteria needs a host in order to grow and develop.
Photosynthesis	Plants use photosynthesis for sunlight.
	Plans use photosynthesis to make food
	Hey, now that I think about it, when machines are solar powered it is pretty much photosynthesis.
Response	I clapped my hands and everyone clapped as the response.
	My response to that clap was I blinked.
	My response to the loud, sudden sound was to jump a little.
	If someone pretended to slap you (stimulus) then you might flinch (response).

Limitations and Future Directions

These findings are based on a sample of students from one science class, so generalization to broader samples is not possible. Many replications of this study with samples of students from across many schools and locations will be needed before any type of generalization is possible. However, it was not the goal of this research to produce broadly generalizable results. Instead, we sought to develop an assessment of vocabulary learning that tapped varied dimensions of vocabulary knowledge identified through many years of research (incrementality, multidimensionality) and captured this knowledge using both receptive and productive items. We chose a science classroom as the setting for developing this assessment because mastery of science language has been identified as highly challenging for many students, and lack of mastery of language in science is particularly detrimental to learning science concepts and content. The results of this limited study of such a vocabulary assessment are highly encouraging, and we hope to replicate them with broader, larger student samples.

This study did not control for student factors such as general word learning or science ability. Correlations are expected between high general abilities and specific science learning, but we were interested in developing a method of assessment that could draw on the disciplinary expertise of the classroom teacher and assess vocabulary ownership of general populations of students such as are found in classrooms everywhere.

The productive item on our assessment was brief, and future research should explore additional ways to assess productive ownership with longer writing samples within science. Additionally, we scored sentences as categorical data: either partial knowledge or no knowledge of the term. Partial knowledge does not equate to complete understanding of a word, and future research will explore the growth from partial to complete knowledge demonstrated in the sentence-writing item. Further intervention studies designed with larger groups of teachers and diverse groups of students are needed, as is analysis of the work of research partnerships that include both disciplinary and literacy experts working to open access for all students to the highly specialized language that accompanies disciplinary study in middle grades and beyond.

References

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

Beck, I., Perfetti, C., & McKeown, M. (1982). Effects of long-term vocabulary instruction on lexical access and reading comprehension. *Journal of Educational Psychology,* 74(4), 506–521.

Berends, M., Lucas, S. R., & Penaloza, R. V. (2008). How changes in families and schools are related to trends in black-white test scores. *Sociology of Education,* 81(4), 314–344.

Blachowicz, C. L. Z., Fisher, P. J. L., Ogle, D., & Watts-Taffe, S. (2006). Vocabulary: Questions from the classroom. *Reading Research Quarterly*, 41(4), 524–539.

Brown, B. A., Ryoo, K., & Rodriguez, J. (2010). Pathway towards fluency: Using "disaggregate instruction" to promote science literacy. *International Journal of Science Education, 32*(11), 1465–1493.

Brown, T. A. (2015). *Confirmatory factor analysis for applied research.* New York, NY: Guilford Publications.

Dale, E. (1965). Vocabulary measurement: Techniques and major findings. *Elementary English, 42*(8), 895–901, 948.

Dyson, A. H., & Genishi, C. (2005). *On the Case: Approaches to Language and Literacy Research*. New York, NY: Teachers College Press.

Fisher, D. (2007). Creating a schoolwide vocabulary initiative in an urban high school. *Journal of Education for Students Placed at Risk, 12*, 337–351.

Fisher, D. & Frey, N. (2014). Content area vocabulary learning. *The Reading Teacher, 67*(8), 594–99.

Fisher, D., Grant, M., & Frey, N. (2009). Science literacy is › strategies. *The Clearing House: A Journal of Educational Strategies, Issues and Ideas, 82*(4), 183–186.

Gee, J. P. (2005). Language in the science classroom: Academic social languages as the heart of school-based literacy. In R. Yerrick & W. M. Roth (Eds.), *Establishing Scientific Classroom Discourse Communities: Multiple Voices of Teaching and*

Learning Research (pp. 19–37). Mahwah, NJ: Lawrence Erlbaum Associates.

Hand, B. M., Alvermann, D. E., Gee, J., Guzzetti, B. J., Norris, S. P., Phillips, L. M., Prain, V., & Yore, L. D. (2003). Message from the "Island Group": What is literacy in science literacy? *Journal of Research in Science Teaching, 40*(7), 607–615.

Hayden, H. E. & Eades-Baird, M. (2016). What Can Students Do With the Words They Know? An ELA Teacher Takes on Science. *Literacy Research: Theory, Method, and Practice*, 65, 182-199.

Hayden H. E., Eades-Baird, M. E., & Singh, A. (2019). Learning the words AND knowing the concepts: An in-depth study of one expert teacher's use of language as a cultural tool to support inquiry. *Literacy*. doi: 10.1111/lit.12192

Honig, S. (2012). Teaching science, teaching language. *Illinois Reading Council Journal, 40*(3), 32–39.

Jagger, S. L., & Yore, L. D. (2012). Mind the gap: Looking for evidence-based practice of "Science Literacy for All" in science teaching journals. *Journal of Science Teacher Education, 23*, 559–577.

Kamil, M. L., & Hiebert, E. H. (2005). The teaching and learning of vocabulary: Perspectives and persistent issues. In E. H. Hiebert & M. Kamil (Eds.), *The Teaching and Learning of Vocabulary* (pp. 1–23). Mahwah, NJ: Lawrence Erlbaum Associates.

Kamil, M. L., Borman, G. D., Dole, J., Kral, C. C., Salinger, T., & Torgesen, J. (2008). *Improving Adolescent Literacy: Effective Classroom and Intervention Practices*. (NCEE #2008- 4027). Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education. Retrieved from http://ies.ed.gov/ncee/wwc.

Keiffer, M. J., Lesaux, N. K., Rivera, M., & Francis, D. J. (2009). Accommodations for English language learners taking large scale assessments: A meta-analysis on effectiveness and validity. *Review of Educational Research, 79*(3), 1168–1201.

Lee, O., Quinn, H., & Valdes, G. (2013). Science and language for English language learners in relation to next generation science standards and with implications for common core state standards for English language arts and mathematics. *Educational Researcher, 42*(4), 223–233.

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.

MacGinitie, W. H., MacGinitie, R. K., Maria, K., & Dreyer, L. G. (2000). *Gates-MacGinitie Reading Tests Manual for Scoring and Interpretation*. Itasca, IL: Riverside Publishing.

Marzano, R. J. (2004). *Building Background Knowledge for Academic Achievement: Research on What Works in Schools*. Alexandria, VA: Association for Supervision and Curriculum Development.

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

McKeown, M. G., Beck, I. L., Omanson, R. C., & Pople, M. T. (1985). Some effects of the nature and frequency of vocabulary instruction on the knowledge and use of words. *Reading Research Quarterly, 20*(5), 522–535. doi:10.2307/747940

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.

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

National Governors Association Center for Best Practices, Council of Chief State School Officers (2010). *Common Core State Standards for Science*. Washington DC.

National Research Council (2011). *A Framework for K–12 Science Education: Practices, Crosscutting Concepts, and Core Ideas.* Washington, DC: National Academies Press.

Nixon, S. B., Saunders, G. L., & Fishback, J. E. (2012). Implementing an instructional framework and content literacy strategies into middle and high school science classes. *Literacy Research and Instruction, 51*(4), 344–365.

Pearson, P. D. (2010). *Embedding Literacy Learning and Instruction Within Inquiry-Based Science: Impact and Synergies*. Retrieved from http://www.nlnw.nsw.edu.au/videos10/7830_2010_pearson/7830/links/pearson.pdf

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

Pearson, P. D., Moje, E., & Greenleaf, C. (2010). *Literacy and science: Each in the service of the other.* Science, 328, 459–463.

Phillips, M., Crouse, J., & Ralph, J. (1998). Does the black-white test score gap widen after children enter school? In C. Jencks & M. Philips (Eds.), *The Black-White Test Score Gap* (pp. 229–272). Washington, DC: Brookings Institution Press.

Riegle-Crumb, C., Moore, C., & Ramos-Wada, A. (2010). Who wants to have a career in science or math? Exploring adolescents' future aspirations by gender and race/ethnicity. *Science Education, 95*(3), 458–476.

Rodriguez, A., Reise, S. P., & Haviland, M. G. (2016). Evaluating bifactor models: Calculating and interpreting statistical indices. *Psychological Methods, 21*(2), 137.

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

Shanahan, T., & Shanahan, C. (2008). Teaching disciplinary literacy to adolescents:

Rethinking content-area literacy. *Harvard Educational Review, 78*(1), 40–59.

Shulman, L. S. (1987). Knowledge and teaching: Foundations of the new reform. *Harvard Educational Research, 57*(1), 1–22.

Snow, C. E. (2015). Rigor and realism: Doing educational science in the real world. *Educational Researcher, 44*(9), 460–466. doi:10.3102/0013189X15619166

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

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.

Ullman, J. B. (2006). Structural equation modeling: Reviewing the basics and moving forward. *Journal of Personality Assessment, 8*7(1), 35–50. doi: 10.1207/s15327752jpa8701_03

Yin, R. (1994). *Case Study Research: Design and Methods, 2nd ed.* Thousand Oaks, CA: Sage Publications.

About the Authors

Emily Hayden is an assistant professor of literacy education in the School of Education at Iowa State University. Having taught for 17 years in K–12 classrooms, she is most interested in research that prepares and sustains better teaching in literacy and better preparation of students to be actively literate citizens. This interest includes research that explores disciplinary literacy practices in science and research that describes how teachers develop adaptive expertise and enact pedagogical content knowledge.

Anupma Singh graduated in May 2018 with a Ph.D. in higher education from Iowa State University. She is currently working in the Office of Institutional Analysis at the University of Wyoming. Her research interests include quantitative research methods and access and success of diverse student populations.

Michelle R. Eades-Baird is an assistant professor of science education within the Master of Arts in Teaching Program at Empire State College. Prior to entering academia, she was a secondary science teacher for 17 years. Her research interests include science disciplinary literacy and academic language, pedagogical content knowledge, and science inquiry.