

# Vocabulary assessment: What we know and what we need to learn

**P. DAVID PEARSON, ELFRIEDA H. HIEBERT**

University of California, Berkeley, USA

**MICHAEL L. KAMIL**

Stanford University, California, USA

**A**fter a nearly 15-year absence from center stage, vocabulary has returned to a prominent place in discussions of reading, and it is alive and well in reading instruction and reading research. We have no doubt that the renaissance is due, at least in part, to the salutary findings about vocabulary in the report of the National Reading Panel (NRP; National Institute of Child Health and Human Development [NICHD], 2000) and, even more important, the use of the NRP findings to shape policy and practice via the Reading First component of No Child Left Behind (2002). We regard these developments as positive, for we think there is good reason to teach vocabulary more aggressively and even better reason to study its relation to comprehension more carefully. However, if we are going to teach it more effectively and if we are going to better understand how it is implicated in reading comprehension, we must first address the vexing question of how we assess vocabulary knowledge and, even more challenging, vocabulary growth. In this essay, we argue that vocabulary assessment is grossly undernourished, both in its theoretical and practical aspects—that it has been driven by tradition, convenience, psychometric standards, and a quest for economy of effort rather than a clear conceptualization of its nature and relation to other aspects of reading expertise, most notably comprehension. We hope that our essay will serve as one small step in providing the nourishment it needs.

There is no doubt that vocabulary is closely tied to comprehension (Davis, 1942; Just & Carpenter, 1987; Whipple, 1925)—in study after study, vocabulary knowledge predicts comprehension performance consistently with positive correla-

tions typically between .6 and .8. But a correlation is not an explanation of a conceptual relation between factors. Anderson and Freebody (1985) understood this complexity well when they put forward three hypotheses to explain the ubiquitous finding of a high correlation between comprehension and vocabulary. The instrumentalist hypothesis argues that learning the words *causes* comprehension. The verbal aptitude hypothesis suggests that general verbal ability is the root cause of both vocabulary and comprehension performance. The knowledge hypothesis argues that both vocabulary and comprehension result from increases in knowledge.

More to the point, it is one thing to demonstrate a correlation and quite another to demonstrate a causal relation between vocabulary instruction or learning and comprehension. In that regard, it is worth noting the conclusions of the subgroup for vocabulary of the NRP (NICHD, 2000), which document a consistent and robust relation between learning vocabulary in specific texts and performance on experimenter-designed comprehension measures derived from those same texts. By contrast, the group found only two studies showing that vocabulary instruction transferred beyond text-specific increases in vocabulary to far transfer measures, such as norm-referenced comprehension reading tests. A question of interest raised by the NRP report is whether its conclusions are generalizable or are the artifact of some special characteristic of the ways in which the outcomes were measured in the studies they examined.

Even though experimentally documented effects of vocabulary instruction on measures of general reading comprehension are weak, at least as indexed by effects on standardized measures, vocabulary instruction has returned to a place of prominence in the reading curriculum; vocabulary serves a core role in commercial reading programs and in other curricular areas such as science, history, or foreign language. Its ubiquity and gravity are captured by the complaint, at least of science educators, that 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 (Armstrong & Collier, 1990).

There are at least three plausible explanations for the weak empirical link between vocabulary instruction and some transfer measures of reading comprehension. The first position is that there is no actual link between the two: that a vocabulary myth has clouded our reasoning and our pedagogy for centuries and that learning words does not cause comprehension. The second is that vocabulary

instruction does not promote far transfer—that is, it is conceptually incapable of moving beyond the texts to which it is tied. Hence it shows up in local but not global indicators of text understanding. The third position, and the one we take up in this essay, is that our measures of vocabulary are inadequate to the challenge of documenting the relationship between word learning and global measures of comprehension. That is, it might be that our instruction is improving vocabulary learning, which might lead to improvements in general comprehension, but the instruments we use to measure vocabulary are so insensitive that they prevent us from documenting the relationship. In particular, the fact that standardized assessments do not often include types of text that are found in textbooks is an example of this potential masking of effects. The National Assessment of Educational Progress (NAEP) 2009 framework has addressed this issue by dividing what have traditionally been labeled expository texts into more explicit and descriptive subcategories (National Assessment Governing Board [NAGB], 2005). Exposition has been separated from, for example, literary nonfiction in recognition of the fact that these different genres have, at the very least, different vocabulary loads.

We don't want to dismiss the first two positions out of hand, but we want to press the measurement question so that it can be ruled in or out as the most plausible explanation for the paucity of documented transfer effects. We will never know until and unless we have developed and tested vocabulary measures that are as conceptually rich as the phenomenon (vocabulary knowledge) they are intended to measure.

We begin by defining vocabulary and offering a short historical account of vocabulary assessment. Then we examine the literature—research, common practices, and theoretical analyses—on vocabulary assessment to answer three questions:

1. What do vocabulary assessments (both past and current) measure?
2. What could vocabulary assessments measure?
3. What research will we have to conduct over the next decade in order to develop and validate measures that will serve us in our quest to improve both vocabulary research and, ultimately, vocabulary instruction?

## How is vocabulary defined?

Any analysis of the domain of vocabulary assessment should first consider what it means to know a word. The first definition of *vocabulary* in

the *Random House Webster's Unabridged Dictionary* (Flexner, 2003) is "the stock of words used by or known to a particular people or group of persons." A subsequent definition is "the words of a language." In turn, *word* is defined as "a unit of language, consisting of one or more spoken sounds or their written representation, that functions as a principal carrier of meaning."

These dictionary definitions provide little specificity and hence little guidance to researchers who are studying vocabulary acquisition and understanding. Faced with the immediate task of reviewing the instructional research, the NRP (NICHD, 2000) was forced to establish parameters for the types of vocabulary that were taught and learned in research studies. The NRP categorized various types of vocabulary as a function of the cognitive operations involved and the context in which vocabulary is measured. The panel asked two questions: (a) Is the use of vocabulary productive or receptive? (b) Is the mode of communication written or oral? Thus, one quickly ends up with the familiar quartet of vocabulary types: listening, speaking, reading, and writing. In general, receptive vocabulary is larger than productive vocabulary; we can understand more words through listening and reading than we use in speech or writing. This conclusion should not be surprising given the general psycholinguistic principle that comprehension normally precedes production and the recognition that additional cueing systems (various textual and contextual aids) are available to individuals during language reception, but not during production.

The assessment of vocabulary as it pertains to reading comprehension has almost exclusively emphasized the receptive dimension of vocabulary. For the most part, at least on large-scale tests, reading is the medium, but a prominent set of vocabulary assessments use the listening mode of the receptive dimension. The Peabody Picture Vocabulary Test (PPVT-III) (1997), a widely used standardized measure of vocabulary development, typifies the latter group of tests. Rarely is the productive aspect of vocabulary examined, especially as it relates to comprehension; for example, when students are taught new words in relation to new texts or topics in subject matter classes, do those words spontaneously emerge in their speaking and writing? The results of one recent analysis document substantial transfer of newly learned vocabulary to writing (an unobtrusive measure—simply looking for the spontaneous occurrence of such words) for students who had participated in an intervention where complex science vocabulary was emphasized in reading, speak-

ing, and listening (Bravo & Tilson, 2006). Despite what we know, much needs to be learned about these complex relationships between the various modes of vocabulary learning and assessment.

## What do vocabulary assessments measure?

### *A brief history*

The assessment of students' knowledge of word meanings, what we generally call vocabulary assessment, is as old as reading assessment itself. Vocabulary assessment dates back to at least the development of the early tests of intelligence by Binet and Thurstone (see Johnston, 1984; Pearson & Hamm, 2005) that preceded formal measures of reading comprehension. The earliest measures of reading vocabulary consisted of asking students to define or explain words that were selected because they were likely to be found in the texts they would encounter in schools; an early item might have asked a student explain individually to an interviewer what a "fork" is used for. With the movement toward mass testing prompted by the need to test recruits for World War I (Resnick & Resnick, 1977) came the need for more efficient, easily administered, and easily scorable assessments; hence the move to standardized, multiple-choice versions of items the students read and responded to. Prototypic items are illustrated in the first row of Table 1.

That sort of item dominated formal vocabulary assessment until the 1970s (Read, 2000), when changes in thinking about language and reading, which emerged from the new fields of psycholinguistics and cognitive science, motivated more contextualized vocabulary assessments such as those found in the second row of Table 1.

The press for contextualization increased systematically, at least in the most ambitious context for vocabulary assessment, English as a Second Language (ESL; see Nation, 2001; Read, 2000), resulting in a progression of items as illustrated in the final three rows of Table 1.

As one can see from the progression of items in Table 1, the field has witnessed the increasing contextualization of vocabulary assessment during the previous quarter century. One would expect greater contextualization to increase the sensitivity of vocabulary assessment to comprehension growth precisely because increasingly contextualized formats require text comprehension as a part of the process of re-

**TABLE 1**  
**SAMPLE ITEMS OF DIFFERENT ERAS**

Time period	Sample item(s)
1915–1920: Decontextualized vocabulary assessment	<p>Pick the word that fits in the blank:</p> <p>A _____ is used to eat with.</p> <p>saw spoon pin car</p> <p>Pick the best meaning for the italicized word:</p> <p><i>glad</i> clever mild happy frank</p>
1970s: Early efforts to contextualize vocabulary	<p>Pick the best meaning for the italicized word:</p> <p>The farmer <i>discovered</i> a tunnel under the barn.</p> <p>built found searched handled</p>
1980s: Steps toward contextualization	<p>In a (1) <i>democratic</i> society, we presume that individuals are innocent until and unless proven guilty. (2) <i>Establishing</i> guilt is (3) <i>daunting</i>. The major question is whether the prosecution can overcome the presumption of (4) <i>reasonable</i> doubt about whether the suspect committed the alleged crime.</p> <p>For each item, select the choice closest in meaning to the italicized word corresponding to the number:</p> <p>2. <i>establishing</i> a. attributing b. monitoring c. creating d. absolving</p> <p>3. <i>daunting</i> a. exciting b. challenging c. intentional d. delightful</p>
1995: Embedded vocabulary assessment	<p>Among a set of comprehension items, you might find the following:</p> <p>In line 2, it says, “Because he was responsible for early morning chores on the farm, John was often tardy for school.”</p> <p>The word <i>tardy</i> is closest in meaning to</p> <p>a. early b. loud c. ready d. late</p>
Late 1990s: Computerized format	<p>Baseball has been a favorite American pastime for over 120 years. Each year, fans flock to diamonds all over the country to pursue this passionate hobby.</p> <p>Look at the word <i>hobby</i> in the passage. Click on the word in the text that has the same meaning.</p>

sponding to the vocabulary items. That, however, is a claim that deserves an empirical rather than a rational test to determine its validity.

This is not to say that, because of this history, all current assessments assess vocabulary in a contextualized manner. In fact, many of the major assessments still use fairly isolated approaches. To illustrate

the nature of current vocabulary assessments, we have analyzed items on four prominent vocabulary assessments that are among those identified by a national panel as fitting the criteria for use in Reading First (Kame'enui, 2002). We chose two individually administered assessments—the PPVT–III and the Woodcock Reading Mastery Test (WRMT–R)

(Woodcock, 1998)—and two that are group administered—Iowa Test of Basic Skills (ITBS; 2005), and the Stanford Achievement Test (SAT–10; 2004). Items characteristic of those included in these assessments are presented in Table 2, except for the PPVT–III. It was difficult to portray the PPVT visually because when taking it, a student sees only pictures. The task is to identify the picture that matches the word spoken by the test administrator. If the target word was *surfing*, the picture set might include someone surfing, someone playing water polo, someone swimming, and someone driving a speedboat.

### *Toward a theory of vocabulary assessment*

Words may seem like simple entities, but they are not. Their surface simplicity belies a deeper complexity. For example, they connect with experience

and knowledge, and their meanings vary depending on the linguistic contexts in which they can be found, including in a variety of literal and figurative contexts. Complexity of word knowledge is evident in Nagy and Scott's (2000) identification of five aspects of word knowledge used in reading:

(a) *Incrementality*: knowing a word is not an all-or-nothing matter; to the contrary, each time we encounter a word and each time we use it, our knowledge becomes a little deeper and a little more precise—eventually leading to nuanced understanding and flexible use.

(b) *multidimensionality*: word knowledge consists of qualitatively different types of knowledge such as understanding nuances of meaning between words such as *glimpse* and *glance* or typical collocations of words (e.g., a *storm front* not a *storm back*).

(c) *polysemy*: many words have multiple meanings, and the more common the word, the more meanings it is likely to

**TABLE 2**  
**PARALLEL ITEMS OF VOCABULARY TASKS ON THREE NORM-REFERENCED TESTS**

Test	Prototypical item(s)
ITBS	To <i>sink</i> in the water play rest wash go down
SAT	Item type 1: To <i>cut</i> is to— slice bark run save  Item type 2: Put the money in the <i>safe</i> . In which sentences does the word <i>safe</i> mean the same thing as in the sentence above? The puppy is <i>safe</i> from harm. I am <i>safe</i> at home. It is <i>safe</i> to go out now. Michael opened the <i>safe</i> .  Item type 3: Ron only has one hat, but he has <i>several</i> coats. <i>Several</i> means— funny some hungry large
WRMT	Subtest 1: Antonyms (read this word out loud and then tell me a word that means the opposite). near (far) dark (light)  Subtest 2: Synonyms (read this word out loud and then tell me another word that means the same thing). cash (money) small (little)  Subtest 3: Analogies (listen carefully and finish what I say [text is visible but experimenter reads the text]). dark—light night—(day) rain—shine wet—(dry)

ITBS = Iowa Test of Basic Skills; SAT = Stanford Achievement Test; WRMT = Woodcock Reading Mastery Test

have; a common word like *run* may have 20 meanings, but a rare word like *geothermal* has but one.

(d) *interrelatedness*: learning or knowing a word often entails derivation or association with the meanings of related words, either in a linguistic context (dogs bark or buffaloes roam) or in one's semantic memory store (dogs are members of the canine category and related to cats because they share the attribute that they can be domesticated).

(e) *heterogeneity*: a word's meaning differs depending on its function and structure (e.g., frequency in written English, and syntactic roles). Contrast, for example, the sentences, "I spilled the cocoa, get a broom," with, "I spilled the cocoa, get a mop." Over time, by experiencing a word like *spill* in different contexts, we learn more about the range of its application.

Nagy and Scott (2000) also identified the ability to reflect on and manipulate vocabulary, or metalinguistic knowledge, as an important feature of word knowledge. Although such categories illustrate the complexity of vocabulary, few studies of vocabulary attend to these variables in any systematic fashion, especially when it comes to choosing the words for instructional interventions or for assessments (Scott, Lubliner, & Hiebert, 2006). At the present time, these distinctions are unlikely to be highly productive as filters for reviewing assessments that are commonly used in large-scale assessment. These variables do, however, suggest important new directions for exploration in vocabulary research. They currently exist, in Nagy and Scott's work, as features of a theory of vocabulary knowledge. However, one would hope to see them eventually as a part of a theory of vocabulary instruction and assessment.

In an analysis of vocabulary assessments for ESL learners, Read (2000) identified three continua for designing and evaluating vocabulary assessments; we believe all three are useful: (a) discrete—embedded, (b) selective—comprehensive, and (c) contextualized—decontextualized. (Read actually uses the terms *context-dependent* and *context-independent* to anchor the two ends of the continuum, but we have substituted our own terminology here.) They represent dimensions that are not only conceptually interesting but also derived from careful analyses of existing tests. We discuss each continuum in turn.

### *Discrete—embedded*

This distinction addresses whether vocabulary is regarded as a separate construct with its own separate set of test items and its own score report, which is the discrete end of the continuum, or whether vocabulary is an embedded construct that contributes to, but is not regarded as separate from, the larger construct of text comprehension. All four of the assessments repre-

sented in Table 2 treat vocabulary as a discrete construct separate from comprehension. The PPVT-III is an entire test devoted to oral receptive vocabulary. The other three assessments each have a separate subtest or set of subtests devoted to vocabulary or, in the case of the WRMT-R, word comprehension. As is typical of norm-referenced reading tests, these subtests allow vocabulary to be reported both as a separate score or as a part of a combined reading score that is some aggregate of vocabulary plus some other reading subscores, most notably comprehension.

By contrast, the NAEP has traditionally taken an embedded stance to vocabulary assessment, being content to ensure that contextualized vocabulary items are a part of one or more of the stances assessed in creating aggregate comprehension scores for text genres. A typical item is immersed in a larger set of comprehension questions and queries the meaning of specific words as used in context, such as the following:

The word *misanthrope* on page 12 means

- (a) an ill-intentioned person
- (b) an ill person
- (c) a person who reacts well to misery
- (d) a person who mistrusts anthropology

In the 2009 NAEP Framework (NAGB, 2005; Salinger, Kamil, Kapinus, & Afflerbach, 2005), the goal is to report vocabulary separately, assuming that the construct, as measured, stands up to the psychometric validation of its statistical independence, and as a part of the overall comprehension score.

More often than not, if the option for a separate score is available, there will be free-standing vocabulary section in the test battery, and it will have its own unique item format and separate time allotment. Conversely, when vocabulary items are included as a part of the overall comprehension score (i.e., embedded), they are most likely to be physically embedded within and distributed among the set of comprehension test items. Note, however, that one could report a separate vocabulary subtest score even if the vocabulary items were physically interspersed among comprehension items. That is exactly the approach that will be taken in the new NAEP assessment.

Inherently there is neither vice nor virtue in a separate vocabulary score; empirically, however, the case for reporting a separate score is strong. As far back as 1942, when Frederick Davis reported the first factor analysis of comprehension test items, he was able to extract a factor for word, along with factors for gist and reasoning. Further, again and again,

analyses of the infrastructure of comprehension assessments implicate something independent about vocabulary knowledge (see Pearson & Hamm, 2005, for a summary of these studies). Hence, the decision by NAGB to report a separate score seems appropriate. As with other questions of vocabulary assessment, the wisdom of the new NAEP approach awaits empirical validation.

### *Selective—comprehensive*

This distinction refers to the relationship between the sample of items in a test and the hypothetical population of vocabulary items that the sample represents. Thus, if one assesses students' grasp of the allegedly new vocabulary in a story from an anthology or a chapter in a science text, the sample is inherently selective; one wants to know if the students learned the words in that particular sample. In general, the smaller the set of words about which we wish to make a claim, the more selective the assessment. We could, however, want to make a claim about students' mastery over a larger corpus of words, such as all of the words in the *American Heritage Dictionary* or the 2,000 most frequently occurring words in English, or all the words in Level 8 of a basal anthology, or all of the words in a science textbook. At the comprehensive end of the continuum, larger hypothetical corpora of words prevail.

This distinction is not just an idle mental exercise; it has enormous implications for the generalizations that can be made from assessments. Consider the items that assess vocabulary in Table 2. Because the items on real tests are copyrighted and cannot be shared publicly, we tried to convey the nature of the tests by creating items that paralleled what we saw on the actual assessments. The process of trying to identify parallel vocabulary to exemplify typical tests was both frustrating and instructive. What immediately struck us was that there were no guidelines, no theories, and no frameworks to guide our choices. We could not infer how or why particular words were chosen for these tests.

We ended up choosing our parallel words by matching the word frequency and decodability of the target words in the actual items. However, this information was not provided in the technical manuals of these assessments. Such a lack of clarity on the source of vocabulary in large-scale assessments is typical of current assessments. Most of our current vocabulary assessments have no theoretically defined population of words at all or, if they do, we have not been able to infer it from the materials they provide for test users. The core of the development process is

psychometric, not theoretical. Test developers obtain a bunch of words, often by asking professionals in the field to tell them what words students at a particular grade level should know; then, they administer all the words to a sample of students at known levels of development or expertise (usually indexed by grade level). The words are sorted by their difficulty, expressed often as the percentage of students in a particular population who answered the question correctly. Ultimately, scores for individuals on such a test derive their meaning from comparisons with the population of students, not words, at large, which is why we call them norm-referenced tests. Under such circumstances, all we know is that a given student performed better, or worse, than the average student on the set of words that happened to be on the test. We know nothing about what the scores say about students' knowledge of any identifiable domain or corpus of words. Whether we want to possess information about domain mastery is, of course, a matter of policy and of educational values. Do we care about the terms in which we describe the vocabulary growth of individuals or groups? Will it suffice to know how a student or a group performed in relation to some other individuals or groups?

In analyzing assessment tasks in archival vocabulary studies, Scott et al. (2006) reported that most researchers had devised assessments that tested knowledge of the specific words that had been taught in an instructional intervention. The only common construct underlying word selection across a majority of studies was students' prior knowledge. That is, it was assumed that the words taught, or at least the majority of them, were unknown to the target students. This assumption was validated in one of three ways: (a) by using a pretest that tested each word directly, (b) by selecting words with a low *p* value (percent correct) from a source such as *The Living Word Vocabulary* (Dale & O'Rourke, 1981), or (c) by asking teachers or researchers to select words not likely to be known by the target population. The criterion of being likely known by the target age group provides little indication of what larger vocabulary students can access as a result of an intervention. For example, if students have learned *consume*, how likely is it that they will also learn something about members of its morphological family, such as *consumer* or *consumable*, or about words likely to be in a mental semantic network with *consume*, such as *eat* or *devour*?

Later in this essay, we review current proposals for theoretically grounded means of selecting words for instruction and assessment. Although none of these frameworks has yet been used for the design of

an assessment, such frameworks suggest that it may be possible to move our assessments to the more comprehensive end of this continuum. Only then will we be able to make claims such as, “The average student in a given school exhibits basic mastery over X% of the words in a given corpus (e.g., the words encountered in a given curriculum in a given grade level).” One could even imagine a computerized assessment system in which all 200 students enrolled in 10th-grade biology took a different sample of 25 vocabulary items from the same corpus for purposes of estimating each student’s mastery over that corpus. Given a corpus of, for example, 150 vocabulary items, there are an indefinitely large number of random samples of 25-word tests that could be generated by computer. One could also imagine a similar approach with smaller corpora within a course (e.g., all of the content vocabulary within a chapter, unit, or project). At the comprehensive end of this continuum, we can begin to think of domain-referenced assessment in the ways in which proponents such as Hively (1974) or Bock, Thissen, and Zimowski (1997) conceptualized the approach.

### *Contextualized—decontextualized*

This continuum refers to the degree that textual context is required to determine the meaning of a word. Any word can readily and easily be assessed in a decontextualized format. But simply assessing a word in a contextualized format does not necessarily mean that context is required to determine its meaning. In order to meet the standard of assessing students’ ability to use context to identify word meaning, context must actually be used in completing the item. Table 3 details several examples to illustrate the continuum.

Item 1 falls firmly on the decontextualized side of the continuum. Even though context is provided for item 2, it is not needed if someone knows the meaning of *consume* as eat or drink. Because the context provides literally no clues about the meaning of *consume*, the item provides no information about a reader’s ability to use context to infer the meaning of a word. In item 3, because all four meanings denote one or another meaning of *consume*, context is essential for zeroing in of the meaning as used in the sentence. Item 4 is even trickier than item 3. Unlike item 3, which requires the selection of the most common meaning of *consume*, item 4 requires a student to reject the default (most common) meaning in favor of a more arcane sense of *consume*. Note also that a very fine semantic distinction is required in item 4 to select *spent wastefully* over *used up*. As a general

**TABLE 3**  
**DEGREES OF CONTEXTUAL RELIANCE**

1. *consumed*
  - a. ate or drank
  - b. prepared
  - c. bought
  - d. enjoyed
2. The people *consumed* their dinner.
  - a. ate or drank
  - b. prepared
  - c. bought
  - d. enjoyed
3. The people *consumed* their dinner.
  - a. ate or drank
  - b. used up
  - c. spent wastefully
  - d. destroyed
4. The citizens *consumed* their supply of gravel through wanton development.
  - a. ate or drank
  - b. used up
  - c. spent wastefully
  - d. destroyed

rule, it is nearly impossible to assess vocabulary in context without reliance on polysemous words and distractor sets that reflect at least two of the meanings of each assessed word. In all fairness, we must admit that there are formats that do not require polysemous words or extremely rare words to assess contextual usage. For example, if one selects really rare, arcane words, the meanings of which can be derived from the textual context, then a straightforward format can be used. Also, one can argue that “picking a word” that fits a blank space absolutely requires the systematic analysis of context. Even so, we like the polysemous format because of its emphasis on close reading of the surrounding context to make a selection from among a set of real meanings of real words.

## What could vocabulary assessments measure?

To begin to answer the question of what vocabulary assessments could measure, we decided to look closely at research that had already been completed or was currently underway. Thus, we looked at existing reviews of research, current investigations, and current developments, particularly the new NAEP vocabulary assessment (Salinger et al., 2005). Our logic was that by looking broadly at extensive reviews of vocabulary and narrowly at cutting edge work, we would get a clear picture of the possible and the feasible.

### *Insights on vocabulary assessments from reviews of research*

The RAND Reading Study Group (RRSG; 2002) was convened to examine what was known about comprehension, with the goal of formulating a plan for research and development. The resulting document includes an analysis of vocabulary research as well as questions in need of intensive research. RRSG acknowledged the strong link between vocabulary knowledge and reading comprehension and speculated that it is an especially important factor in understanding the reading problems experienced by second-language learners. However, RRSG cautioned that the relationship between vocabulary knowledge and comprehension is extremely complex, because of the relationships among vocabulary knowledge, conceptual and cultural knowledge, and instructional opportunities. Surely, as we look to the future, we will want to know more about whether there are any special considerations for assessing vocabulary for students learning English as a second language.

What we know about the nature of instruction that influences vocabulary learning can aid in the design of assessments. The NRP (NICHD, 2000) reviewed 50 experimental and quasiexperimental studies published in English in refereed journals. One provocative finding from the NRP report is that students acquire vocabulary best when it is used in meaningful, authentic contexts; indeed, they are less

able to remember words that are presented in isolated formats, such as lists. As was apparent in the analysis of the current assessments on the decontextualized—contextualized continuum, many current vocabulary assessments present words in a decontextualized context. Contrasting the power of isolated versus contextualized vocabulary assessments to predict both passage specific and general comprehension should be a priority.

Another critical finding in the NRP is that students often do not fully understand the task when asked to show evidence of vocabulary knowledge. If tasks are restructured so that students understand what is expected, students often do better. Restructuring seems to be particularly effective for low-achieving or at-risk students. Again, this conclusion has important implications for assessment, given the general difficulty of assessing skills and knowledge among low-achieving or at-risk students.

Two of the characteristics of vocabulary learning from the Nagy and Scott's (2000) list of important characteristics of vocabulary acquisition have implications for assessment research: incrementality and heterogeneity. If a new word meaning is acquired incrementally rather than in an all-or-nothing fashion, it would seem useful to gauge students' developing depth of understanding of important words. There have been a few attempts to begin such an endeavor. For example, Stallman, Pearson, Nagy, Anderson, and García (1995) found a way to discriminate among levels of depth by manipulating the set of distractors from which a student was asked to select a correct response. Students encountered the same test item several times, as illustrated in Table 4. As one moves from one level to the next, the discrimination task becomes more refined and, presumably, more difficult. However, this represents only a beginning; much remains to be done to operationalize the construct of incrementality.

Heterogeneity in Nagy and Scott's (2000) view suggests that the more contexts in which a word is encountered, the greater the likelihood that its meaning will be acquired, or more precisely, the greater the likelihood that a precise, nuanced, and even sophisticated meaning will be acquired. To assess the influence of heterogeneity, we could assess word meaning across situations in which a new or rare word appeared in varying frequencies; say once, twice, and five times. In addition, of course, the quality of the context matters too; it may be that when a word is encountered in a highly supportive context (where the semantic relatedness of the surrounding words is high), students perform differently than in a less supportive context.

**TABLE 4**  
**ASSESSING DEPTH OF VOCABULARY KNOWLEDGE**

1. A *gendarme* is a kind of
  - a. toy
  - b. person
  - c. potato
  - d. recipe
2. A *gendarme* is a kind of
  - a. public official
  - b. farmer
  - c. accountant
  - d. lawyer
3. A *gendarme* is a kind of
  - a. soldier
  - b. sentry
  - c. law enforcement officer
  - d. fire prevention official
4. One would most likely encounter a *gendarme* in
  - a. New York
  - b. Nice, France
  - c. London, England
  - d. New Orleans

### *Insights on assessment from perspectives on selecting words for instruction*

Our previous discussion of the selective—comprehensive dimension of vocabulary selection emphasized the point that vocabulary on current assessments is not selected on the basis of any evident criteria. For all intents and purposes, any word in the English language could be found on a typical vocabulary test, provided that it discriminates across students. The question of interest is how *could* word choices be made in a more principled way. Three prominent perspectives on word selection offer underlying theoretical, or at least conceptually interesting, frameworks that could be translated into principles for selecting words to appear on a vocabulary test. To be clear, the scholars whose work we review have developed these frameworks as tools to select words for instruction. We are the ones who are extrapolating their potential as tools to select words for assessment; nonetheless, it may be a useful extrapolation.

The most prominent perspective on word selection at the present time is that of Beck, McKeown, and Kucan (2002). Beck and her colleagues view vocabulary as falling into three tiers. The first tier is comprised of high-frequency words (e.g., *come, go, happy, some*) that do not need to be taught, except perhaps to English learners, and the third tier is comprised of rare words that are specific to particular content domains (e.g., *chlorophyll, photosynthesis, xylum*). They believe that vocabulary instruction should focus on second-tier words. Words in that second tier characterize the vocabulary of mature language users when they read and write. They are best thought of as less common labels for relatively common concepts: *stunning* in place of *pretty*, *pranced* instead of *walked*, *astonished* but not *surprised*. As such, they constitute the language of sophisticated academic discourse, at least as it is represented in narrative fiction. In research and programs guided by the tier model, Beck and her colleagues have identified words from texts, mainly narrative, and either provided teachers with candidate words for Tier 2 instruction or taught them how to select Tier 2 words for their own lessons. The rules for selecting Tier 2 words are not precisely expressed in the Beck et al. research. This presents a problem for the development of vocabulary assessments. However, one could imagine a principle stipulating that only Tier 2 words (or perhaps Tier 2 words in a given frequency band—say English nouns, verbs, and adjectives that rank between 2,000 and the 5,000 on a frequency count), are candidates

for assessment at a given grade level and that the correct foil in a multiple-choice item is always the most common synonym (e.g., *pretty*) for any given Tier 2 target (e.g., *stunning*). The validity of such a rule would have to be established through research on the ultimate utility of such a definition of Tier 2 words.

There are other approaches to the selection of words. Biemiller (2005; Biemiller & Boote, 2006; Biemiller & Slonim, 2001) has identified a group of words judged to be worth teaching during the primary grades. These are words that are known by 40 to 80% of students at the end of grade 2. Such words might be thought of as a set of “Goldilocks” words—not too easy and not too hard (Stahl & Nagy, 2005). There is a deeper rationale behind Biemiller’s work. He and his colleagues assume that, other things being equal, students are likely to acquire these words in roughly the order of their “knownness” by a large sample of students at the end of second grade, with the least commonly known words learned last. Equipped with such a hypothetical list, if we select and sequence words for instruction in descending order of how well they are known among end-of-year second-grade students, we can make it possible, at any given point in the school year, for students to be on track to learn the next set of words they are likely to need in their everyday reading. In this way, we could eliminate, or at least minimize, the vocabulary gap between various groups of students who, by virtue of differences in experience and instruction, differ markedly in their vocabulary knowledge. Biemiller found just such corpus of words in Dale and O’Rourke’s (1981) *The Living Word Vocabulary* grade levels 2, 4, and 6. Level 2 words were considered easy and not recommended for teaching. Through testing approximately 2,870 *The Living Word Vocabulary* level 4 and level 6 root word meanings and rating another 1,760 meanings, Biemiller has identified some 1,860 root-word meanings that are appropriate for instruction during the primary grades. These could easily become the corpus of words from which samples could be drawn for assessments of various sorts, including standardized assessments.

Hiebert’s (2005, 2006) framework employs three elements as part of a principled vocabulary curriculum. The first principle—the richness of a word’s semantic associations—builds on and extends the work of Beck et al. (2002). As new labels for already known concepts (Graves, 2000), the Tier 2 words are part of semantic networks with words that are similar in meaning. In the principled curriculum, the richness of a word’s semantic network is established by reference to an analysis of semantic associations.

Hiebert has used Marzano and Marzano's (1988) semantic clusters to establish the richness of the semantic network of which a word is part. Marzano and Marzano classified 7,230 words from elementary school texts into three levels: superclusters (61), clusters (430), and miniclusters (1,500). Some superclusters have numerous clusters and these, in turn, have numerous miniclusters. For example, the word *hue* can be described as having a sparse set of semantic associations in that it is part of a miniclustert with only two additional words (*color*, *tint*) and is part of the supercluster devoted to words about *color*, consisting of only 29 words. By contrast, *plunge* is part of the *descending motion* cluster with 19 words that, in turn, is part of the *types of motion* supercluster with 321 words.

To give a general indication of the opportunities that students have had with the semantic concept underlying a word, Hiebert (2006) has used the Marzano and Marzano (1988) categories to identify words as members of one of three different groups: (a) rich semantic connections (superclusters with 200 or more members), (b) moderate semantic connections (superclusters with 100–199 members), and (c) sparse semantic connections (superclusters with 21–99 members).

“Knownness” is the second principle of Hiebert's curriculum, and it builds directly on the work of Biemiller (2005; Biemiller & Boote, 2006) and Dale and O'Rourke (1981) described in a previous section of this essay. Knownness is operationally defined as those words that students at particular grade levels respond to correctly on the vocabulary assessments developed by Dale and O'Rourke and Biemiller and Boote.

The third principle, family frequency, combines the insights on the centrality of word frequency among second-language scholars such as Nation (1990, 2001) and the discoveries about the importance of the morphological families (Carlisle & Katz, 2006; Nagy, Anderson, Schommer, Scott, & Stallman, 1989). If one assumes that students are capable of recognizing common roots across instances of occurrence, then the notion of frequency must be modified dramatically from counts of the frequency of individual words. For example, although the word *consume* can be expected to appear 5 times per million words (Zeno, Ivens, Millard, & Duvvuri, 1995), members of its morphological family appear an additional 90 times per million words (*consumed*, 7; *consumer*, 37; *consumers*, 28; *consumers'*, 1; *consumes*, 1; *consuming*, 2; *consumption*, 14).

Hiebert (2006) has described how the words on four prominent vocabulary tests distributed themselves according to the three elements in her model—semantic connectedness, knownness, and frequency of morphological families. The analysis of words on vocabulary assessments used the third-grade forms of the same vocabulary tests for which items are illustrated in Table 2—the PPVT, WRMT, ITBS, and SAT. Table 5 presents the results of this analysis. It is relevant to our assessment concerns because it illustrates how the words on vocabulary assessments could be viewed in terms of Read's (2000) selective—comprehensive continuum.

Specifically, Table 5 portrays the distribution of one of Hiebert's (2005, 2006) elements, morphological families, across the four assessments. The data in Table 5 show that the types of words on three of the four assessments, the WRMT–R, ITBS, and the

**TABLE 5**  
**TARGET WORDS AND THEIR MORPHOLOGICAL FAMILIES IN PARTICULAR WORD ZONES AS PERCENTAGES OF TOTAL WORDS ON AN ASSESSMENT**

	PPVT		WRMT		ITBS		SAT	
	Target word	Target word + morphological family	Target word	Target word + morphological family	Target word	Target word + morphological family	Target word	Target word + morphological family
Zones 0–2 (High: 100+)	6	14	39	48	36	57	49	7
Zones 3–4 (Moderate (10–99.99))	27	36	41	37	46	33	43	24
Zones 5–6 Rare (>1 –9.99)	67	50	20	15	18	10	8	4

Reprinted with permission from Hiebert (2006).

SAT-10, were similar, with only a small percentage of the words in the rare category (8–20%). By contrast, the majority of the words (67%) on the PPVT-III fell into the rare zones. The pattern on this feature was similar to that for the other two features (semantic associations and knownness).

The PPVT-III is the outlier on these assessments. But why? One possibility is that the PPVT-III, as an individually administered, administrator-paced test, is designed to span a wide range of levels of vocabulary knowledge in a single text—thus it will necessarily require a large number of “obscure” words in order to be sensitive to individual differences at the high end of the vocabulary knowledge scale. But if range is responsible for differentiating the PPVT-III from the other assessments, then one would also expect the WRMT-R (another test with a wide range of items spanning various grades), to behave like the PPVT-III rather than like the other group reading tests. It doesn't. Another possibility is that the PPVT-III taps oral, not written, receptive vocabulary knowledge. Hence there is no need to worry about the decodability of words, making it possible to assess children's mastery over the conceptual knowledge of even orthographically rare words. In the final analysis, however, we admit that we are not sure what makes the PPVT-III behave so differently from other widely used, wide-range assessments.

### *Insights from new assessments*

Nowhere is a theory of contextualized vocabulary assessment more prominent than in the recent NAEP framework (NAGB, 2005; Salinger et al., 2005). In developing that framework, the Framework Committee took a new stance on the role of reading vocabulary assessment. In previous frameworks and assessments, vocabulary items were included, but only to ensure breadth of coverage of important aspects of the reading framework, and vocabulary items were folded into an overall comprehension score. In the previous model, a word was selected for two reasons: (a) because it was deemed important and (b) in order to assess whether a reader was able infer its meaning from context. In the new framework, the committee signaled an important shift: “*vocabulary items will function both as a measure of passage comprehension and as a test of readers' specific knowledge of the word's meaning as intended by the passage author*” (NAGB, p. iv, emphasis added). Thus, vocabulary takes on a more important role, with a hope that it will prove to be a sufficiently robust construct that it could be reported as a separate

score in addition to serving as a part of the overall comprehension score.

In addition, the theory behind the role of vocabulary as a part of comprehension is quite different. In the new framework, the emphasis is “the meanings of the words that writers use to convey new information or meaning, not to measure readers' ability to learn new terms or words” (NAGB, 2005, p. 35). This principle is operationalized in a set of criteria for choosing words according to the following: (a) words that characterize the vocabulary of mature language users and written rather than oral language; (b) words that label generally familiar and broadly understood concepts, even though the words themselves may not be familiar to younger learners; (c) words that are necessary for understanding at least a local part of the context and are linked to central ideas such that lack of understanding may disrupt comprehension; and (d) words that are found in grade-level reading material (NAGB; Salinger et al., 2005). In short, the specified words are of the type that Beck et al. (2002) have called Tier 2 words—uncommon labels for relatively common concepts. As noted earlier, these words constitute the language of sophisticated academic discourse, particularly in literary text. In fact, in science and mathematics, much of academic discourse is new labels for new concepts—what Beck and her colleagues call Tier 3 words. The NAEP framework has emphasized information texts and recognizes the different vocabulary loads in information and literary text. We have limited knowledge of the generality of the “Tier” concept because the research of Beck and her colleagues has been focused on literary texts.

In order to achieve complete operationalization of this approach to vocabulary assessment, the committee has established a set of rules for generating items and distractors. A set of distractors may include (a) a word that has a more common meaning of a target word, but that must be ignored in favor of the meaning in context; (b) a word that presents correct information or content from the text that is *not* what is meant by the target word; (c) a word that has an alternative interpretation of the context in which the target word occurs; or (d) other words that look or sound similar to the target word (NAGB, 2005; Salinger et al., 2005). Distractors play an important role in operationalizing the underlying theory of vocabulary knowledge as key component of comprehension, especially in the requirement that students must reject an alternative, and presumably sometimes more common, sense of the word (e.g., ignore stunning as *bewildering* in favor of stunning as *splendid* or *beautiful*).

This development within NAEP would have a significant influence even if it had no interesting theoretical grounding simply because it is NAEP and therefore influential in shaping other assessments. Given the fact that the new assessment venture is both theoretically interesting and provocative (i.e., it takes a stand on which aspect of vocabulary acquisition is worth assessing), it is likely to be exceptionally influential in shaping a broader set of vocabulary assessment practices.

## What could be the research agenda for the next decade?

The questions we have raised in this essay, where we have tried to draw inferences about vocabulary assessment issues from current efforts to understand or improve vocabulary instruction and assessment, would constitute an ambitious research agenda. However, we would certainly endorse such ambitious efforts. But we feel the need to raise additional assessment issues in closing, albeit without unpacking any of them in depth, just to make sure they get into the queue for future efforts.

1. A first priority should be to devote explicit research attention to the distinctions among various aspects of vocabulary that we have discussed in this essay, rather than simply using a global definition of vocabulary or some general concept of word meaning. One of the major issues is the type of vocabulary that is being taught and tested. For example, often reading vocabulary is intended to be assessed, although the instrument used might measure expressive vocabulary, or vice versa. Similarly, the term *vocabulary* is used almost interchangeably as we move between writing, listening, speaking and reading without making either conceptual or operational distinctions. We contend that these relatively simple changes would yield great dividends in our knowledge of the relationships between vocabulary knowledge, vocabulary instruction, and literacy. A simple example would be the targeted instruction of reading vocabulary based on a receptive vocabulary measure rather than an expressive vocabulary measure, which might be more important for speaking.

2. In order to conduct the research described in the preceding paragraph, much effort needs to be exerted in the development of assessments that are clear about the components and types of vocabulary. Researchers need to focus on the components and formats of vocabulary assessment, particularly with regard to the selection of words, sampling proce-

dures, and so forth as we have as noted in this essay. That research is needed to determine whether any single assessments can represent the various aspects of vocabulary we have identified (and, perhaps, some we have not) or whether we need individual and targeted assessments for each of the types of vocabulary. Without that information, progress in vocabulary research will be limited.

3. It is clear that informational text typically carries a heavier vocabulary load than does literary text. Currently, that difference is a hidden variable in many studies. Research is needed to untangle the relationship between text genre and vocabulary variables such as how words are chosen for instruction and the vocabulary load of the text. Regardless of what the answers will be, they will have profound implications for vocabulary instruction, and transfer. Because vocabulary is dealt with currently in a holistic fashion, one dividend might be to differentiate methods of instruction for vocabulary by text genre. Learning technical vocabulary from a biology text is clearly different from learning vocabulary in a story, where most of the word can be related to personal experiences.

4. The three preceding points all converge on the issue of transfer of vocabulary knowledge to other components of reading. The research alluded to here would almost certainly offer insights on the difficulties we have raised in this essay about issues of transfer and the specific effects of vocabulary instruction on comprehension. More important is the explicit attention to the issues of transfer, both near and far, for the tasks under investigation. In addition, the strength of transfer over time should be a part of this effort, particularly given the relatively short duration of many vocabulary instruction interventions in the literature.

5. The NAEP venture bears close watching, to see whether it is capable of generating a new paradigm for conceptualizing and measuring vocabulary. In particular, we hope that someone undertakes some value-added studies to determine what the new paradigm adds above and beyond more prosaic and conventional approaches to vocabulary assessment. The first administration using this new paradigm will not occur until 2009, giving us some time to address some of these questions.

6. There is still a set of unanswered issues that were raised in the RRSg (2002) report about the conditions and effects of vocabulary and vocabulary instruction that would, if answered, provide quantum leaps in our knowledge base. Among the issues raised in the RRSg report is the relationship of vocabulary instruction to literacy for non-native speak-

ers of English. At least a part of any research agenda should include an emphasis on the RRSB issues.

7. Finally, we need a serious attempt to implement computerized assessments of vocabulary domains, along the lines of those suggested in the section of this essay detailing the selective—comprehensive continuum. In a better world, we would not be limited to conventional norm-referenced assessments of vocabulary acquisition, where our only benchmark for gauging vocabulary growth is the average performance of other students. We could opt instead for estimates of mastery over particular domains of interest (e.g., all of the words in a given curriculum or a given frequency band) or estimates of control over other characteristics that might prove to be effective indexes of vocabulary learning (e.g., all words with a common morpheme, such as *spec*). Given the capacity of computers to store and analyze large corpora and our recent advances in computer adaptive assessment, the time appears right for such an exploration.

As we said at the outset, it is our hope that this essay will help to fuel the recent enthusiasm in the field for vocabulary research, in particular research on vocabulary assessment. Only when we are sure about the validity and sensitivity of our assessments will we be able to determine the relations among various modes of vocabulary development and the relations between vocabulary knowledge and other aspects of reading development. This agenda, we believe, is a wise investment for the field.

**P. DAVID PEARSON** serves as Dean of the Graduate School of Education at the University of California, Berkeley and as a faculty member in the Language, Literacy, and Culture program. His current research focuses on issues of reading instruction and reading assessment policies and practices. Before coming to UC Berkeley, Pearson served on the reading education faculties at Minnesota, Illinois, and Michigan State. He can be contacted at the Graduate School of Education 1501 Tolman Hall #1670, University of California, Berkeley, Berkeley, CA 94720-1670, USA, or by e-mail at ppearson@berkeley.edu.

**ELFRIEDA H. HIEBERT** is an adjunct professor at the University of California, Berkeley. Her research interests focus on the effects of texts on the fluency, vocabulary, and comprehension of elementary-level students, especially English-language learners. She can be contacted at 106 Phelan Court, Santa Cruz, CA 95060, USA, or by e-mail at hiebert@berkeley.edu.

**MICHAEL L. KAMIL** is professor of psychological studies in education at the Stanford University School of Education. His research interests are in the intersection of literacy and technology and second language learners. He chaired the Vocabulary, Teacher Professional Development, and Technology subgroups of the National Reading Panel. He can be contacted at 123 Cubberley Hall, 485 Lasuen Mall, Stanford, CA 94305, USA, or by e-mail at mkamil@stanford.edu.

## REFERENCES

- ANDERSON, R.C., & FREEBODY, P. (1985). Vocabulary knowledge. In H. Singer & R.B. Ruddell (Eds.), *Theoretical models and processes of reading* (3rd ed., pp. 343–371). Newark, DE: International Reading Association.
- ARMSTRONG, J.E., & COLLIER, G.E. (1990). *Science in biology: An introduction*. Prospect Heights, IL: Waveland.
- BECK, I.L., MCKEOWN, M.G., & KUCAN, L. (2002). *Bringing words to life: Robust vocabulary instruction*. New York: Guilford.
- BIEMILLER, A. (2005). Size and sequence in vocabulary development: Implications for choosing words for primary grade vocabulary instruction. In E. Hiebert & M. Kamil (Eds.), *Teaching and learning vocabulary: Bringing research to practice* (pp. 223–242). Mahwah, NJ: Erlbaum.
- BIEMILLER, A., & BOOTE, C. (2006). An effective method for building vocabulary in primary grades. *Journal of Educational Psychology*, 98, 44–62.
- 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, 498–520.
- BOCK, R.D., THISSEN, D., & ZIMOWSKI, M.F. (1997). IRT estimation of domain scores. *Journal of Educational Measurement*, 34, 197–211.
- BRAVO, M.A., & TILSON, J.L. (2006, April). *Assessment magazines: Gauging students' depth of reading comprehension and science understanding*. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA.
- CARLISLE, J.F., & KATZ, L.A. (2006). Effects of word and morpheme familiarity on reading of derived words. *Reading and Writing*, 19, 669–693.
- DALE, E., & O'ROURKE, J. (1981). *The living word vocabulary: A national vocabulary inventory*. Chicago: World Book/Childcraft International.
- DAVIS, F.B. (1942). Two new measures of reading ability. *Journal of Educational Psychology*, 33, 365–372.
- FLEXNER, S.B. (Ed.). (2003). *Random House Webster's unabridged dictionary* (2nd ed.). New York: Random House.
- GRAVES, M.F. (2000). A vocabulary program to complement and bolster a middle-grade comprehension program. In B.M. Taylor, M.F. Graves, & P. van den Broek (Eds.), *Reading for meaning: Fostering comprehension in the middle grades* (pp. 116–135). New York: Teachers College Press.
- HIEBERT, E.H. (2005). In pursuit of an effective, efficient vocabulary curriculum for elementary students. In E.H. Hiebert & M.L. Kamil (Eds.), *Teaching and learning vocabulary: Bringing research to practice* (pp. 243–263). Mahwah, NJ: Erlbaum.
- HIEBERT, E.H. (2006, April). *A principled vocabulary curriculum*. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA.
- HIVELY, W. (1974). Introduction to domain-reference testing. *Educational Technology*, 14, 5–10.
- JOHNSTON, P.H. (1984). Assessment in reading. In P.D. Pearson, R. Barr, M.L. Kamil, & P. Mosenthal (Eds.), *Handbook of reading research* (pp. 147–184). New York: Longman.
- JUST, M.A., & CARPENTER, P.A. (1987). *The psychology of reading and language comprehension*. Boston: Allyn & Bacon.
- KAME'ENUI, E.J. (2002). *An analysis of reading assessment instruments for K–3*. Eugene, OR: Institute for the Development of Educational Achievement, University of Oregon.
- MARZANO, R.J., & MARZANO, J.S. (1988). *A cluster approach to elementary vocabulary instruction*. Newark, DE: International Reading Association.
- NAGY, W., ANDERSON, R.C., SCHOMMER, M., SCOTT, J., & STALLMAN, A. (1989). Morphological families in the internal lexicon. *Reading Research Quarterly*, 24, 262–282.
- NAGY, W.E., & SCOTT, J.A. (2000). Vocabulary processes. In M.L. Kamil, P. Mosenthal, P.D. Pearson, & R. Barr (Eds.), *Handbook of reading research* (Vol. 3, pp. 269–284). Mahwah, NJ: Erlbaum.
- NATION, I.S.P. (1990). *Teaching and learning vocabulary*. Boston: Heinle & Heinle.
- NATION, I.S.P. (2001). *Learning vocabulary in another language*. Cambridge, England: Cambridge University Press.
- NATIONAL ASSESSMENT GOVERNING BOARD. (2005). *Reading framework for the 2009 National Assessment of Educational Progress*. Washington, DC: American Institutes for Research.

NATIONAL INSTITUTE OF CHILD HEALTH AND HUMAN DEVELOPMENT. (2000). *Report of the National Reading Panel. Teaching children to read: An evidence-based assessment of the scientific research literature on reading and its implications for reading instruction* (NIH Publication No. 00-4769). Washington, DC: U.S. Government Printing Office.

NO CHILD LEFT BEHIND ACT OF 2001, Pub. L. No. 107-110, 115 Stat. 1425 (2002).

PEARSON, P.D., & HAMM, D.N. (2005). The history of reading comprehension assessment. In S.G. Paris & S.A. Stahl (Eds.), *Children's reading comprehension and assessment* (pp. 13-70). Mahwah, NJ: Erlbaum.

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

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

RESNICK, D.P., & RESNICK, L. (1977). The nature of literacy: An historical exploration. *Harvard Educational Review*, 47, 370-385.

SALINGER, T., KAMIL, M.L., KAPINUS, B., & AFFLERBACH, P. (2005). Development of a new framework for the NAEP reading as-

essment. In C.M. Fairbanks, J. Worthy, B. Maloch, J.V. Hoffman, & D.L. Schallert (Eds.), *54th yearbook of the National Reading Conference* (pp. 334-349). Oak Creek, WI: National Reading Conference.

SCOTT, J.A., LUBLINER, S., & HIEBERT, E.H. (2006). Constructs underlying word selection and assessment tasks [plato4] in the archival research on vocabulary instruction. In J.V. Hoffman, D.L. Schallert, C.M. Fairbanks, J. Worthy, & B. Maloch (Eds.), *55th yearbook of the National Reading Conference* (pp. 264-275). Oak Creek, WI: National Reading Conference.

STAHL, S.A., & NAGY, W.E. (2005). *Teaching word meanings*. Mahwah, NJ: Erlbaum.

STALLMAN, A.C., PEARSON, P.D., NAGY, W.E., ANDERSON, R.C., & GARCÍA, G.E. (1995). *Alternative approaches to vocabulary assessment* (Tech. Rep. No. 607). Urbana-Champaign, IL: Center for the Study of Reading, University of Illinois.

WHIPPLE, G. (Ed.). (1925). *The 24th yearbook of the National Society for the Study of Education: Report of the National Committee on Reading*. Bloomington, IL: Public School Publishing.

ZENO, S., IVENS, S., MILLARD, R., & DUVVURI, R. (1995). *The educator's word frequency guide*. Brewster, NY: Touchstone Applied Science Associates.