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.

Judith A. Scott
Education Dept
University of California, Santa Cruz
1156 High St,
Santa Cruz, CA 95064
(831) 459-5066
FAX: (831) 459-4618
jascott@ucsc.edu

Merrit Hoover
104 Bond Court
Los Gatos, CA 95030
mhoover@ucsc.edu

Susan Leigh Flinspach
Education Department
University of California, Santa Cruz
Santa Cruz, California 95064
(831) 459-2239
flinspac@ucsc.edu

Jack Vevea
School of Social Sciences, Humanities and Arts
The University of California
5200 North Lake Road
Merced, CA 95343
jvevea@ucmerced.edu

The VINE Project is funded by IES Reading and Writing Education Research Grant Program #R305G060140 (U.S. Department of Education), FY2006-2009. This presentation is the sole responsibility of the authors and does not necessarily reflect the opinions of the U.S. Department of Education.

For more information on the VINE Project, see http://vineproject.ucsc.edu/.

A Multiple-Level Vocabulary Assessment Tool: Measuring Word Knowledge Based on Grade-Level Materials

Judith A. Scott
Merrit Hoover
Susan Leigh Flinspach
University of California, Santa Cruz

Jack L. Vevea
University of California, Merced

In the past three decades, the assessment of student learning has influenced both policy and research pertaining to literacy practices more than in the previous history of American education (Afflerbach, 2004; Wixson & Pearson, 1998). Large-scale assessments of reading, in particular, have become prominent tools for making important educational decisions, often trumping other sources of information (Afflerbach, 2004; IRA, 1999). In discussing the limitations of reading assessments, the RAND Reading Study Group report (2002) points out that, unless measures of assessment reflect and are consistent with underlying theories of reading comprehension, we as researchers "will be severely hampered in our capacities to engage in excellent research" (p. 129).

We know quite a bit about vocabulary acquisition, and yet this knowledge has not filtered into assessments of vocabulary knowledge, or assessments of reading comprehension, beyond the idea of adding context (Pearson, Hiebert, & Kamil, 2007). Although measures of vocabulary knowledge have been included in reading assessments since the 1920s, the traditional means of assessing vocabulary have recently come under fire for being "driven by tradition, convenience, psychometric standards and a quest for economy of effort" (Pearson, Hiebert & Kamil, 2007, p. 282). This hampers both research and decision making, for most of the measures do not reflect our current understanding of the nature of word learning (Scott, Lubliner, & Hiebert, 2006). Instead, the words on current large-scale vocabulary assessments are chosen arbitrarily, with no theoretically grounded principles, theories or frameworks to guide their selection for the tests (Pearson, Hiebert & Kamil, 2007). Words are chosen for their discriminatory power and their ability to operate in psychometerically appropriate ways to differentiate students, without consideration of the complexity and mulitfaceted nature of word learning (Pearson, Hiebert & Kamil, 2007).

Because issues of vocabulary learning are intertwined with conceptual knowledge, sociocultural realities, instructional opportunities, and the slippery nature of words, this is not an easy subject to tackle (Scott, Nagy & Flinspach, 2008). However, in their review of vocabulary assessment, Pearson, Hiebert and Kamil (2007) suggest that it is feasible to do a better job than what is currently done, for "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" (p. 295).

Kame'enui, *et al.* (2006) sought to develop a framework to evaluate K-3 assessment tools for vocabulary and the four other components of reading (phonemic awareness, phonics, fluency, and reading comprehension) specified in the Reading First Initiative within the No Child Left Behind

legislation. Confronted with variation and discrepancies in the information available about the tests of interest, however, they narrowed their efforts to formulating an overall evaluative judgment of the trustworthiness of—or sufficiency of evidence about—the assessment functions. They found that few instruments used to assess K-3 student reading performance, including vocabulary acquisition, met their criteria of trustworthiness. Although a step forward in thinking about good tests, their efforts did not address the need for assessments that reflect the underlying processes of learning to read and master vocabulary.

This paper discusses the development of two forms of a vocabulary test that build on what is known about vocabulary learning as an incremental and multifaceted process. They capture elements of the complexity of word learning and are tied to the curriculum being taught in fourth grade. The assessments have been administered to a diverse sample of students in California, and the results indicate that the tests are highly reliable.

SOME PRINCIPLES OF LEARNING VOCABULARY THAT SHOULD MATTER IN ASSESSMENTS

What it means to know a word is not simple or straightforward (Anderson & Nagy, 1991; Beck & Mckeown, 1991). Words have multiple meanings, are often abstract, are used in idioms, vary according to register, and differ according to context. Nagy and Scott (2000, p. 575) identify five aspects of word knowledge that indicate the complexity of word learning: incrementality, multidimensionality, polysemy, interrelatedness, and heterogeneity:

> (1) incrementality—knowing a word is a matter of degrees, not all-or-nothing; (2) multidimensionality—word knowledge consists of several qualitatively different types of knowledge; (3) polysemy—words often have multiple meanings; (4) interrelatedness—one's knowledge of any given word is not independent of one's knowledge of other words; and (5) heterogeneity—what it means to know a word differs substantially depending on the kind of word.

These components of word learning are not commonly measured, or even taken into account, in vocabulary assessments. Current assessments also tend to ignore a student's opportunities to learn the words in the classroom before taking the test. Given that these factors are likely to influence what students know, they should matter in the construction of vocabulary assessments. Three of these factors—incrementality, multidimensionality, and opportunity to learn—are addressed by the tests described in this paper.

Researchers have shown that mastering a new word is usually an incremental process; it occurs along a continuum (Beck & McKeown, 1991; Beck, Perfetti & McKeown, 1982; Dale, 1965; Nagy & Scott, 2000; Paribakht & Wesche, 1999; Shore & Durso, 1990; Stahl, 2003). Children's knowledge of a word's meaning is often incomplete initially, and it grows with repeated exposures to the word. Beck, McKeown, and colleagues found that up to 40 well-taught instructional encounters with a word do not necessarily ensure full mastery (Beck, Perfetti, & McKeown, 1982; McKeown, Beck, Omanson, & Perfetti, 1983; McKeown, Beck, Omanson, & Pople, 1985).

In vocabulary assessment, Dale (1965) translated the notion of "word learning on a continuum" into four levels of word knowledge ranging from "I never saw it [the word] before" to "I

know it." More recently, Paribakht and Wesche (1996) added a fifth gradation of knowledge, asking students to indicate whether or not they could use the word in a sentence. Stallman, Pearson, Nagy, Anderson & Garcia (1995) have also attempted to gauge incremental growth in understanding words through the manipulation of distractors in the test that assess gradations of knowledge about each word. These attempts to incorporate incremental learning into vocabulary assessments have all been limited, and much more needs to be done to "operationalize the construct of incrementality" (Pearson, Hiebert & Kamil, 2007, p. 290).

Another principle of word learning, multidimensionality, underscores that all words are not considered equal. Words differ on multiple dimensions such as part of speech. Syntactic awareness, or knowing the part of speech, is identified as an important, and often neglected, component of learning from context (McKeown, 1985; Werner & Kaplan, 1952) and dictionary use (Scott & Nagy, 1997). Understanding the role of syntax in word learning is implicit in cloze tasks, where a correct guess depends on understanding the part of speech needed to fill in the blank correctly. Cloze tasks are often included in traditional vocabulary assessments, but knowledge of words' parts of speech can be assessed in other ways as well.

A second feature of multidimensionality deals with words that have families of morphological variants, such as *relate, relates, related, relating, relative, relation*, and *relationship*. Hiebert (2005; 2006) has identified morphological family frequency as an important element of a principled vocabulary curriculum. Word frequency is usually attributed to individual words, but Hiebert suggests that family frequency, the larger set of words that share the same morphological root, may be a more important factor in word learning. For instance, although the word *consume* appears only 5 times per million in a large corpus of words (Zeno, Ivens, Millard & Duvvuri, 1995), it is part of a larger set of related words (*consumer, consumers, consumed, consumption, etc.*) that appear an additional 90 times per million words (Pearson, Hiebert, & Kamil, 2007). Words that are seen more often are more likely to be learned (Stahl, 2003), which indicates that morphological family frequency should be a consideration when selecting the words in vocabulary assessments.

A third factor affecting the words that students know is their opportunity to learn words in class. Consequently, the vocabulary items tested ought to relate to the materials and curriculum covered in a particular domain or grade level. Most states have developed content standards for subjects taught at each grade level. For instance, fourth graders in California study the state's history including the missions that were established by Spanish priests. Although the students may encounter words such as *indigenous* or *mission* on a social studies test, they are unlikely to find them on a vocabulary assessment. Rather, the words on most vocabulary assessments do not reflect students' opportunities to learn words; they are chosen arbitrarily and are unconnected to grade-level curricula or materials (Read & Chapelle, 2001).

Despite their importance for vocabulary knowledge, word-learning principles and students' opportunities to learn have been largely ignored in test construction. The VINE vocabulary assessments described in this paper represent an attempt to begin to remedy this oversight.

VINE VOCABULARY ASSESSMENTS

The impetus for developing a measure of vocabulary growth based on the words that students actually study and on principles of word learning emerged from our work in the VINE (Vocabulary

Innovations in Education) Project, a federally funded research project that explores the development of vocabulary knowledge in fourth grade through word consciousness. VINE teachers assist students in building an associative network surrounding words, with knowledge of the subtle distinctions between related words that occur in the same semantic field (Scott, 2004). By focusing on word learning as a generative process, in which students learn metacognitively about words as opposed to learning a specific set of words, VINE seeks to help students develop both the skill and the will to learn more words.

We thought that developing vocabulary assessments that follow the fourth grade curriculum and that pay attention to levels of word knowledge, parts of speech, and morphological family frequency was necessary and compelling. In the process, we wanted the instruments to have the psychometric properties that would make them trustworthy in the eyes of others.

Development of the VINE Assessments

To test words that our students were likely to encounter during fourth grade, we asked local fourth-grade teachers to identify novels that were typically read aloud in their classrooms or used in literature circles with many of their students. In addition, we asked them about the textbooks they used in math, social studies, and science. We formed a list of 21 trade books and four textbooks for academic subjects commonly used in this region of California (see Appendix A). These were the source texts for our corpus of words. As our research team read these books, they pulled out words that they thought fourth-grade students might find difficult. From over 30,000 words culled initially, we formed a corpus of more than 5,000 distinct words that fourth-graders in surrounding school districts were likely to read and/or hear in class.

From this corpus, we formed two vocabulary assessments, a fiction test (with words drawn from the trade books) and a non-fiction test (with words from the math, science, and social studies textbooks). Only words that appeared in two or more of the source books were included, as these were words students seemed more likely to encounter during the school year. About 20% of the words on the assessments appeared in both fictional and non-fictional sources. Our rationale for testing an equal number of words from trade books and from subject-specific textbooks was that approximately three hours of each school day in fourth grade are devoted to reading and writing and three hours to everything else. Consequently, the distribution of words reflected the time that fourth-grade teachers may devote to vocabulary development across all subjects in the school day.

The individual words for each test were selected from the corpus with attention to multidimensionality regarding part of speech, membership in morphologically related families, and frequency. The words on each test reflect the parts of speech across the entire corpus—35% are verbs, 39% are nouns, 5% are adverbs, and 21% are adjectives. Most test words have numerous morphological variants, but a few words (e.g., *lunar*) with no variants were also included. The morphological family frequency of the selected words ranges from low (*sizzle*: family U-value = .29) to relatively high (*century:* family U-value = 61.34). U-value is a standardized measure of frequency per million words adjusted for variation in distribution.

To capture the incremental aspect of word learning, each selected word was developed as a testlet, a five-question measure of gradations of knowledge about the word. While other vocabulary assessments tend to focus on a student's ability to define a word, the VINE tests were designed to encompass a broader base of knowledge. The first question about each word asks test-takers to

make a recognition judgment: "*Have you ever seen or heard of this word before?*" Students who have neither heard of, nor seen, the test word before are directed to indicate that on the first question and then skip to the next test word. If they have seen or heard the word before, test-takers go on to the second question, which asks them about their confidence in their knowledge of the word: "*I've heard it, but I'm not sure what it means*" versus "*I think I know what it means.*" The third question requires students to identify the word's semantic field, and for the fourth question, test-takers choose the correct definition of the word. The last question asks students to identify the word's part or parts of speech. They indicate one or more parts of speech for each word. The test booklet was formatted so that one word and its 5-question testlet appear per page. See Figure 1.

According to Pearson, Hiebert and Kamil (2007), distractors used in a vocabulary assessment "play an important role in operationalizing the underlying theory of vocabulary knowledge" (p. 293). Our distractors on the third and fourth questions of each testlet include at least two parts of speech so that the student's answers to these questions do not influence the response on the last question regarding part of speech. The distractors were also selected so that the correct semantic field matches the correct definition, with the majority of the distractors falling outside "close matches" if students knew almost anything about the word. Thus, the students had to choose whether the word *flick* might have something to do with: a) dolphins; b) looking; c) fruit; or d) movement. Although dolphins might flick their tails or eyes might flick back and forth, choosing either of these would be a stretch. The

Figure 1. Sample Test Item

PYRAMID

Circle either" yes" or "no." If you circle "no," then go on to the next word. If you circle "yes," then please answer the rest of the questions below.

Have you ever seen or heard this word before?
NO—skip to next word
YES—please answer all the questions below

Circle the letter of one answer.

How well do you know this word?
a. I've heard it, but I'm not sure what it means.
b. I think I know what it means.

I think the word may have something to do with:
a. Farming
b. Noise
c. Movement
d. Math

I think this word means:
a. Making a sound like metal hitting metal
b. Moving down something
c. A triangle shape with a square base
d. A type of tractor

Circle all that apply.

I think this word may be:
a. A noun
b. A verb
c. An adjective
d. An adverb

distractors for the definition of *flick* were: a) the seed of a jungle fruit; b) the way dolphins swim; c) looking carefully and d) moving something quickly.

Piloting and Administering the VINE Vocabulary Assessments

Both VINE vocabulary assessments, the one from fictional sources and the one from non-fictional sources, were piloted in a fourth-grade classroom that was not in the VINE Project. Based on the pilot tests, we reduced the number of words on each test from 50 to 36, decided that 15 minutes was the best length for the timed testing period, and agreed that students should take the two tests on different days.

Members of the research team gave the VINE fiction and non-fiction vocabulary assessments to intact classrooms in both the fall and spring. On a testing day, students completed one vocabulary assessment and a separate VINE writing prompt. Each administrator followed a script that introduced the vocabulary assessment and ran through the 5-question testlet for two sample words. During the timed testing period, administrators and teachers circulated among the students to ensure that they were filling in the assessments correctly. Students were asked to put their pencils down at the end of 15 minutes exactly, and administrators collected the test booklets.

In following the standardized scripts, administrators reviewed several key aspects of the assessments. They informed students that the tests were timed and, should the students be completely unfamiliar with a word, that they should respond to the first question only for that word and then skip to the next word in the test. They demonstrated how students who recognized a word should respond to the first two questions according to their own individual degree of familiarity with the word. Then they showed students that the semantic field and definitional questions each had one correct answer and that the question on the part of speech could have one or more correct responses. They reviewed the parts of speech through a short example sentence, labeling a noun, verb, adjective, and adverb in the sentence. After going through the two sample words and answering student questions, they left the example sentence with the labeled parts of speech visible to the class throughout the testing period.

VINE Student Participants

The VINE vocabulary tests were administered in the fall and again in the spring of the 2006-07 academic year to VINE student participants. The 380 VINE students were from 11 fourth-grade classrooms and two split classrooms with both fourth and fifth graders. The classrooms were located in metropolitan, town, and rural schools in seven districts in one region of California. VINE selected these classrooms because the enrollment of English learners was between 10 and 60%, and because the teachers had experience with writing workshop, had some flexibility in their teaching, and committed to the VINE Project for at least a year.

In 2006-07, about half of the VINE students, 46 percent, spoke a language other than English at home, and 32 percent were designated English learners. In the town and rural schools, English learners usually came from Spanish-speaking households. The language situation was more complicated in the metropolitan schools, though, with students from households representing 26 home languages. After English, Spanish was the most common home language for VINE students, spoken by slightly more than 31%.

The VINE classrooms were also diverse in other ways. Based on state classifications of primary ethnicity, 41% of the students were Latino, 28% were White, just over 9% were Filipino, slightly under 9% were Asian, and about 4% were African American. About 0.5% were American Indian or Alaskan Native, another 0.5% were Pacific Islanders, and the primary ethnicity of the remaining 8% was not specified. Most of the classrooms included students with disabilities, usually specific learning disabilities or speech or language impairment. Five VINE classrooms were located in a district that allocates all of its Title I monies to specific schools. Three of those classrooms were in all-Title I schools, and two were not. In the remaining eight VINE classrooms, student Title I enrollment varied from 0 to 100%. By numerous criteria, VINE classrooms in 2006-07 were diverse.

VINE test administrators assumed that all VINE students who were attending school on an assessment day would take the vocabulary test. Occasionally a teacher requested that a student be excluded because of a severe disability or because of newcomer status (where a student speaks no English), but almost all VINE students receiving special resource or language services took the VINE tests. When alternative programs (band or English language development) pulled more than four students out of the classroom during the assessment period, VINE test administrators returned at a later date to make up the test with the missing students.

The two VINE vocabulary assessments were developed to reflect the incremental and multidimensional nature of word learning and to take into account students' opportunities to learn vocabulary in the classroom. At the beginning and end of the 2006-07 academic year, they were administered in a systematic way to most of the 380 VINE student participants in 13 diverse classrooms. The next section discusses the psychometric properties of the tests based on the student responses.

APPLYING GRADED ITEM RESPONSE THEORY TO THE VINE VOCABULARY TESTS

Traditional vocabulary scales typically rely on classical test theory: they sum each item and report the summed score as an index of vocabulary knowledge. This approach treats each vocabulary word equivalently, which results in a considerable loss of information. These traditional measures fail to account for the range in an item's ability to make fine distinctions between persons of similar proficiency levels. Vocabulary items differ with respect to this characteristic, known as the item's discrimination. Using an item response theory (IRT) approach results in a test that is more sensitive to the range of difficulties in vocabulary words and that differentially weights items that are more discriminating between test-takers of different proficiency levels. In IRT, the probability of successfully completing an item is dependent on specific item parameters, including discrimination and difficulty of the item, and individual proficiency.

An IRT analysis assumes independence of items (conditional on proficiency), which is problematic for tests like the VINE vocabulary assessments that are composed of sets of items that are linked by their reference to a common word. In order to circumvent this issue, we grouped the five items based on each vocabulary word into a composite item. This practice is based on Thissen's (1989) idea of a "testlet" in which related items are linked to create one larger, graded item. We then

applied Samejima's (1996) graded IRT model, allowing for a range of scores on each composite item or testlet, rather than traditional binary scoring methods. Thus, our vocabulary test is made up of a series of these testlets with each vocabulary word contributing a single graded response.

We calculate scores for a testlet by summing the correct responses on the five items related to each vocabulary word. For example, the word *flutter* is scored as follows:

- Initial score = 0
- Have you ever seen or heard this word before? If yes, add 1; if no, add 0.
- How well do you know this word? a) I've heard it, but I'm not sure what it means; b) I think I know what it means. If a, add 0; if b, add 1.
- I think the word may have something to do with: a) blankets; b) walking; c) houses; d) butterflies. If d, add 1; otherwise, add 0.
- I think this word means: a) walking quickly in a straight line; b) the edge of a blanket; c) flying like a butterfly; d) a large and fancy house. If c, add 1; otherwise, add 0.
- I think this word may be: a) a noun; b) a verb; c) an adjective; d) an adverb. If b, add 1; otherwise, add 0.

In this example, the possible scores range from zero to five. On other items, multiple parts of speech can be counted as correct, allowing for a score range of up to seven.

Scaling. The testlet items were scaled using public domain software called Augment v2, which implements the same scaling algorithm used in Multilog (Thissen, Chen, & Bock, 2003). We selected testlets for which every possible score was adequately represented in our sample (no categories of response were omitted). We then calibrated the testlet items and selected for those with high discrimination. Three testlets (*admirable, brazenly,* and *ancient*) were dropped from the fiction portion of the test and four (*landscape, crop, rough,* and *location*) were dropped from the non-fiction portion due to their low discrimination. This resulted in tests composed of 33 testlets on the fiction portion and 32 on the non-fiction portion, with all testlets selected for high discrimination and a range of difficulty levels.

Each testlet varies with regard to its discrimination, but most have unusually high discrimination scores (greater than 1.0). Discrimination scores indicate how well an item distinguishes between test-takers of slightly differing abilities. The estimated discrimation values for each testlet are reported in Tables 1 and 2 (for non-fiction and fiction items respectively).

Testlets have a range of possible scores (from zero to seven), so we also calculated threshold parameters associated with each discrete score for a particular testlet. Threshold parameters are related to the probability that a respondent with a particular proficiency level will respond in the next higher category, and they are reported in the same metric as raw proficiency. So, for example, the first threshold parameter for the "*migrate*" testlet, -1.14, implies that a respondent with proficiency equal to –1.14 has a 50% chance of moving from the a score of 0 to a score of 1 on that particular testlet. The full range of threshold parameters for each testlet is provided in Tables 1 and 2 for non-fiction and fiction items respectively.

Marginal reliability. In IRT, the quality of measurement depends on proficiency. In classical test theory construction, the reliability of a particular test is the same regardless of proficiency, as long as respondents are in the proficiency range targeted by the test. To obtain an analogue of classical reliability, then, we averaged reliability across the range of possible proficiencies, weighting

Table 1. Non-fiction testlets, discrimination parameter & threshold parameters.

	Discrimination Parameter	Score 1 Threshold	Score 2 Threshold	Score 3 Threshold	Score 4 Threshold	Score 5 Threshold	Score 6 Threshold
migrate	1.07	-1.14	-0.88	-0.72	-0.34	0.68	
pyramid	1.40	-1.08	-0.95	-0.77	-0.44	0.56	
decorative	1.40	-0.33	-0.25	-0.07	0.26	1.71	
petroleum	0.88	1.13	1.27	1.51	1.88	2.82	
settle	1.23	-1.19	-0.88	-0.43	0.16	1.45	
bison	1.25	0.23	0.32	0.54	0.67	1.08	
inference	0.70	0.46	0.80	1.35	2.30	5.25	
conserve	1.09	0.19	0.52	1.21	1.94	2.75	
shrub	1.40	-0.01	0.20	0.53	0.79	1.32	
irrigate	0.64	1.35	2.12	3.31	4.34	6.32	
lunar	1.38	0.42	0.62	0.81	1.34	3.38	
absorb	1.56	-0.67	-0.50	-0.23	0.20	1.07	
survival	1.23	-1.36	-1.25	-0.97	-0.42	2.20	
equivalent	1.40	-0.04	0.03	0.17	0.47	1.23	3.48
mild	1.68	-0.28	-0.08	0.20	0.51	1.65	
rarely	1.86	-0.23	-0.06	0.21	0.76	2.21	
northward	1.05	-0.24	-0.14	0.01	0.63	3.36	
sizzle	1.19	-0.60	-0.46	-0.26	0.14	1.12	3.05
earn	1.25	-1.10	-1.02	-0.88	-0.55	1.02	
hollow	1.38	-0.55	-0.44	-0.17	0.38	0.99	3.04
classify	1.57	0.50	0.71	1.06	1.55	2.56	
producer	1.60	-0.30	-0.21	-0.07	0.24	0.99	
express	1.05	-0.62	-0.51	-0.03	1.08	1.86	4.01
moist	1.78	-0.22	-0.15	0.08	0.48	1.76	
solution	1.75	-0.22	-0.12	0.11	0.43	1.43	
reservoir	1.44	1.51	1.73	1.98	2.29	2.55	
bellow	0.70	0.12	0.35	0.86	1.86	3.20	6.07
observe	1.67	-0.33	-0.21	-0.08	0.21	1.05	
circular	1.54	0.08	0.14	0.26	0.56	1.99	
century	1.34	-0.24	-0.17	-0.03	0.22	1.32	
seedling	1.29	-0.15	-0.08	0.09	0.37	1.02	
analyze	1.57	0.52	0.62	0.72	1.04	1.96	

by the likelihood of obtaining a proficiency in each range. In IRT, this concept is called *marginal reliability*. Marginal reliabilities on the VINE vocabulary tests were quite high: .90 for fiction and .92 for non-fiction.

Table 2. Fiction testlets, discrimination parameter & threshold parameters.

	Discrimination Parameter	Score 1 Threshold	Score 2 Threshold	Score 3 Threshold	Score 4 Threshold	Score 5 Threshold	Score 6 Threshold
flutter	1.08	-1.34	-1.21	-0.84	-0.47	0.32	3.20
heroine	0.57	1.58	2.43	3.31	4.29	5.42	
hiss	0.90	-3.19	-2.97	-2.45	-1.82	-0.55	2.49
grim	1.05	-0.43	-0.21	0.14	0.92	2.41	
flick	1.05	-2.17	-2.09	-1.81	-1.26	-0.37	3.71
perch	1.08	-0.68	-0.53	-0.30	0.05	0.82	3.64
pilot	1.02	-2.16	-2.08	-1.75	-1.45	-0.56	3.01
glance	1.49	-0.92	-0.84	-0.62	-0.37	0.35	3.11
bundle	1.30	-0.71	-0.55	-0.34	0.14	0.93	2.83
fade	1.12	-1.00	-0.82	-0.40	0.32	1.59	3.79
dazzling	1.61	-0.39	-0.14	0.26	0.70	1.60	
aisle	1.36	0.53	0.76	0.95	1.07	1.52	
grove	1.22	-0.42	-0.04	0.52	0.99	1.69	
glisten	1.27	0.34	0.52	0.77	1.39	2.89	
chirp	0.94	-1.79	-1.67	-1.47	-1.21	-0.17	2.81
hillock	0.56	2.37	2.81	3.51	4.34	6.22	
brim	1.41	0.34	0.56	1.01	1.38	1.97	
clammy	1.37	-0.33	0.21	1.09	1.79	2.74	
canopy	1.42	0.14	0.31	0.66	1.10	1.61	
frock	0.79	1.83	2.17	2.96	3.74	4.61	
hem	0.77	0.53	0.74	1.37	2.07	3.07	
glum	1.33	0.41	0.57	0.69	1.19	2.21	
haze	1.01	0.65	0.90	1.34	2.10	3.03	
hesitate	1.80	0.11	0.20	0.34	0.68	1.47	
frantic	1.70	0.50	0.65	0.82	1.20	2.34	
musician	1.19	-0.84	-0.80	-0.69	-0.49	0.85	
despair	1.25	0.16	0.31	0.51	1.09	2.23	
burst	1.34	-0.16	-0.03	0.33	1.11	2.98	
complain	1.04	-0.73	-0.63	-0.46	-0.06	1.56	
chatter	0.95	-0.39	-0.32	-0.23	-0.04	0.86	3.90
mission	0.91	-0.45	-0.37	-0.16	0.12	1.24	
announcement	0.83	-0.32	-0.25	-0.17	0.08	3.02	
glider	1.38	0.45	0.57	0.78	1.02	1.68	

Expected a posteriori scores. After calculating summed scores for each testlet and selecting optimal testlets using Samejima's (1996) graded IRT model, we developed expected a posteriori (EAP) scores that estimate participants' proficiency based on their individual test responses. In IRT, an individual's score is often estimated by calculating the average proficiency associated with a particular response pattern. These expected a posteriori (EAP) scores do not directly

correspond to the summed score that would be calculated in classical test theory. Instead, IRT accounts for the fact that the probability of attaining a particular set of item responses on a test is uniquely associated with the proficiency level of the examinee. Therefore different response patterns could have the same summed score, yet be associated with different proficiency level estimates.

In order to avoid having different EAPs for participants who have the same summed score, we calculated "summed-score EAPs" by weighting the EAP estimate for each possible response pattern by the likelihood of that response pattern conditioned on proficiency. The raw summed score is translated into these summed-score EAPs, which are expressed in a metric that provides the test-taker and examiner with a convenient way to translate the raw summed score into a useful approximation of the EAP. The loss of information is minimal when we compare these simpler scores to pattern-scored EAPs. Both fiction and non-fiction summed score EAPs were centered at 50, with $s = 10$, for fourth-grade participants at the beginning of the academic year.

Tables 3 and 4 provide the summed-score-to-EAP translation tables for the non-fiction and fiction tests, respectively. These tables facilitate the translation of an individual's summed score into the final metric of the test. We provide the tables to illustrate the simplicity of using the tests in

Table 3. Non-fiction summed score to EAP translation table. For brevity, not every possible summed score is provided, but the complete range of observed summed scores on the non-fiction portion of the test and their EAP equivalents are covered.

Summed Score	EAP Equivalent
0	24
10	33
20	39
30	43
40	46
50	49
60	52
70	55
80	57
90	60
100	63
110	66
120	69
130	73
140	78
150	84
160	93

Table 4. Fiction summed score to EAP translation table. For brevity, not every possible summed score is provided, but the complete range of observed summed scores on the fiction portion of the test and their EAP equivalents are covered.

Summed Score	EAP Equivalent
0	20
10	29
20	34
30	39
40	43
50	47
60	50
70	53
80	56
90	59
100	62
110	65
120	68
130	72
140	76
150	81
160	87
170	95

practice. For example, a respondent who attained a summed score of 80 on the fiction test would earn an EAP proficiency estimate of 56. The user of the test, then, can simply apply the translation using the tables and need not be concerned with details related to the IRT scaling.

DISCUSSION

Any multiple-choice test that tries to capture student word knowledge is likely to fall short of that goal. Assessing word knowledge may seem simple, just as acquiring word knowledge may seem simple at first blush. But, words are intertwined with experiences and background knowledge, can be used literally or figuratively, occur in idioms and with multiple meanings, and have nuanced meanings in different contexts. Assessment of word knowledge is not an easy task.

The VINE vocabulary tests build on a theory of word learning as a complex and incremental process to assess the depth of an examinee's knowledge of each vocabulary word. The tests summarize performance across words by weighting based on the degree to which incremental knowledge of individual words discriminates between respondents who have different abilities. They have excellent reliability.

The design of the VINE vocabulary tests also ensures that they have content validity, and we expect to establish their construct validity as well. The tests are based on an underlying theory of vocabulary acquisition, using two principles of word learning—incrementality and multidimensionality (Nagy & Scott, 2000). The test questions reflect the principle that mastering a new word usually happens incrementally along a continuum of growth. The tests also take into account that words are multidimensional in two ways. First, the final question about each word asks about its syntactic roles or parts of speech. In fact, one component of word selection was that the ratio of words representing the parts of speech on each test mirrors the distribution of parts of speech across the entire corpus of source words. In addition, words belong to morphological families that vary greatly in overall frequency, and that frequency was a criterion for selecting the words on the tests. Thus, VINE vocabulary tests were constructed in accordance with a theory of how children actually learn words.

Other design features of the VINE vocabulary assessments also contribute to their validity as a set of measures of fourth-grade vocabulary development in this region of California. Students' opportunity to learn the words in their fourth-grade classrooms was fundamental to the process of word selection for the tests. The source materials were all trade books or textbooks used by local fourth-grade teachers, and the words on the tests occur in more than one source book. This honors the incremental principle that multiple exposures are often needed to learn new vocabulary. Moreover, the two tests, one with words from fictional sources and the other with words from non-fiction textbooks, take account of vocabulary learning outside of language arts; they represent the heterogeneity of word learning throughout the student's day. In these ways, the VINE vocabulary tests move towards the development of better, more valid vocabulary assessments.

In their discussion of vocabulary assessment, Pearson, Hiebert & Kamil (2007) call for research that takes several new directions. They seek research that pays attention to distinctions among various types of vocabulary, including the vocabulary of different text genres or possibly different subjects. They also recommend that researchers focus on how words are selected for vocabulary tests. They

are interested in vocabulary assessments that estimate mastery of particular domains of knowledge, in opposition to the conventional, norm-referenced tests that set a benchmark of vocabulary growth at the average performance of other students. Lastly, they call for research that examines more thoroughly the vocabulary development of non-native speakers through assessment.

The VINE vocabulary tests represent a step in the directions proposed by Pearson, Hiebert, and Kamil (2007). The tests are based on the curriculum covered in local fourth-grade classrooms rather than on words found to be good discriminators of vocabulary knowledge across a national sample. The fictional and non-fictional sources for the VINE tests reflect subject-matter differences in the vocabulary that students encounter. In addition, the words on the test were chosen with frequency and syntactic categories in mind, matching the distribution of syntactic categories identified in the source materials. These design principles could be used by other researchers as they work on new vocabulary assessments.

The population used to assess the viability of the VINE tests was quite diverse. The classrooms represented diversity in terms of language abilities, home language, disabilities, primary ethnicity, and economic status. Thirty-two percent of VINE students were English learners. When testing students, particularly those from homes where academic English is rarely used, it seems imperative that we have vocabulary measures that capture what we are asking students to learn. The VINE tests were found to be highly reliable with this large and varied population of fourth graders.

Although a strength of the VINE vocabulary assessments is their reflection of the local curriculum, it is also a limitation. The vocabulary covered in a fourth-grade curriculum is location-specific. Fourth-grade texts and trade books in California may overlap to some extent with those in Nebraska, but there will also be differences. To reflect regional differences in curriculum on a test, the vocabulary words need to be regionally specific. The capacity of computers to store and analyze large corpora of words can facilitate this effort, with states, districts or teachers selecting words for a test from a broad range of source materials. Eventually, districts or teachers may be able to align vocabulary tests with the curriculum actually taught in local schools.

One criticism of the VINE vocabulary assessments is that they test words out of context. The desire to question students' knowledge of parts of speech, which we did on the VINE tests, came into conflict with placing the tested words into a sentence context. We plan to continue our assessment inquiry by examining the contribution of the parts-of-speech question to the discriminating power of the testlets. It may be possible to develop a good testlet without asking about parts of speech, thus enabling the tested word to appear in a sentence.

We hope to extend our use of this approach so that vocabulary tests targeted at different grade levels can be linked using a common proficiency scale. The IRT methods that we employed make such linking feasible, and they also provide a way to assess whether the test functions in similar ways for different sub-populations (e.g., native speakers versus English language learners). This approach to vocabulary assessment provides tests that are grounded in current theories of word learning and in students' opportunities to encounter the vocabulary in their classrooms.

Overall, we hope that others will build on the innovations in the VINE vocabulary assessments. This study shows that it is possible to create assessments of vocabulary learning that capture gradations of vocabulary knowledge and that incorporate what we know about the complexity of word knowledge into a psychometrically sound and reliable set of tests. We hope the VINE

vocabulary tests will encourage test publishers and researchers to look beyond current assessment offerings to find or develop alternatives that are more theoretically sound.

We gratefully acknowledge funding by the IES Reading and Writing Education Research Grant Program #R305G060140 (U.S. Department of Education) for this Research.

REFERENCES

Afflerbach, P. (2004). High Stakes Testing and Reading Assessment: National Reading Conference Policy Brief. *Journal of literacy research, 37*(2),151 -162.

Anderson, R. C., & Nagy, W. E. (1991). Word meanings. In R. Barr, M. Kamil, P. Mosenthal, & P. D. Pearson (Eds.), *Handbook of reading research: Vol. 2.*, (pp. 690-724). New York: Longman.

Beck, I. & McKeown, M., (1991). Conditions of vocabulary acquisition. In R. Barr, M. Kamil, P. Mosenthal, P. D. Pearson, P. (Eds.), *Handbook of Reading Research Volume II* (pp. 789-814). Mahwah, NJ: Erlbaum.

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.

Dale, E. (1965). Vocabulary measurement: Techniques and major findings. *Elementary English, 42*, 82–88.

Graves, M. (1987). The roles of instruction in fostering vocabulary development. In M. McKeown & M. Curtis (Eds.), *The nature of vocabulary acquisition* (pp. 165–184). Hillsdale, NJ: Erlbaum.

Hiebert, E. H. (2005). In pursuit of an effective, efficient vocabulary curriculum for elementary students. In E. H. Hiebert & M. 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.

International Reading Association (1999). *High-stakes assessments in reading: A position statement of the International Reading Association.* Newark, DE: International Reading Association.

Kame'enui, E., Fuchs, L., Francis, D., Good, R., O'Connor, R., Simmons, D., Tindal, G., & Torgesen, J. (2006). The adequacy of tools for assessing reading competence: a framework and review. *Educational Researcher, 35*(4), 3-11.

McKeown, M. G. (1985). The acquisition of word meaning from context by children of high and low ability. *Reading Research Quarterly, 20*, 482-496.

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*, 3-18.

McKeown, M., Beck, I., Omanson, R., & Pople, M. (1985). Some effects of the nature and frequency of vocabulary instruction on the knowledge and use of words. *Reading Research Quarterly, 20*, 522-535.

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: Volume III* (pp. 269-284). Mahwah, NJ: Erlbaum.

[NICHD] National Institute of Child Health and Human Development. (2000). *Report of the National Reading Panel: Teaching children to read.* (NIH Publication No. 00-4754). Washington, DC: U.S. Government Printing Office.

Paribakht, T. S. & Wesche, M. (1996). Enhancing Vocabulary Acquisition Through Reading: A Hierarchy of Text-Related Exercise Types. *Canadian Modern Language Review, 52*, 155-78.

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.

Paribakht, T. S., & Wesche, M. (1999). Reading and incidental L2 vocabulary acquisition: An introspective study of lexical inferencing. *Studies in Second Language Acquisition, 21*(2), 195–224.

Rand Reading Study Group. (2002). *Reading for understanding: Toward a research and development program in reading comprehension.* Prepared for the Office of Educational Research and Improvement. Washington, DC: US Department of Education.

Read, J. & Chapelle, C. (2001). A framework for second language vocabulary assessment, *Language Testing 18*(1), 1-32.

Samejima, F. (1996). The graded response model. In W. J. van der Linden & Hambleton, R. K. (Eds.), *Handbook of modern item response theory* (pp. 85-100). New York: Springer.

Scott, J. A. (2004). Scaffolding vocabulary learning: Ideas for equity in urban settings. In D. Lapp, C. Block, E. Cooper, J. Flood, N. Roser, and J. Tinajero (Eds.), *Teaching all the children: Strategies for developing literacy in an urban setting* (pp. 275-293). NY: Guilford.

Scott, J., Lubliner, S. & Hiebert, E. H. (2006). Constructs underlying word selection and assessments tasks 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.

Scott, J, & Nagy, W. (1997). Understanding the definitions of unfamiliar verbs. *Reading Research Quarterly, 32,* 184-200.

Scott, J. A., Nagy, W. E., & Flinspach, S. L. (in press). More than merely words: Redefining vocabulary learning in a culturally and linguistically diverse society. In A. Farstrup & J. Samuels (Eds.), *What research has to say about vocabulary instruction.* Newark, DE: International Reading Association.

Shore, W. & Durso, F. (1990). Partial knowledge in vocabulary acquisition: general constraints and specific details. *Journal of Educational Psychology, 82(2),* 313-318.

Stahl, S. (2003). How words are learned incrementally over multiple exposures. *American Educator, 27*(1). 18-19.

Stallman, A., Pearson, P. D., Nagy, W. E., Anderson, R. C., & Garcia, G. E. (1995). *Alternative approaches to vocabulary assessment (Tech Rep. No. 607).* Urbana, IL: Center for the Study of Reading, University of Illinois.

Thissen, D., Chen, C. & Bock, D. (2003). Multilog (Version 4.0) [Computer software]. Lincolnwood, IL: Scientific Software International.

Thissen, D., Steinberg, L. & Mooney, J. (1989). Trace lines for testlets: a use of multiple-categorical-response models. *Journal of Educational Measurement, 26,* 247-260.

Wainer, H., Vevea, J. L., Camacho, F., Reeve, B., Rosa, K., Nelson, L., Swygert, K., & Thissen, D. (2001). Augmented Scores —"Borrowing Strength" to Compute Scores Based on Small Numbers of Items. In Thissen, D. & Wainer, H. (Eds.), *Test Scoring.* New Jersey: Lawrence Erlbaum Associates, Inc.

Werner, H., & Kaplan, E. (1952). *The acquisition of word meanings: A developmental study.* Monograph of the Society for research in Child Development (Vol. 15, Serial 51(1)).

Wixson, K. & Pearson, P. D. (1998). Policy and Assessment Strategies to Support Literacy Instruction for a New Century. *Peabody Journal of Education, 73*(3/4). 202-227.

Zeno, S. M., Ivens, S. H., Millard, R. T., & Duvvuri, R. (1995) *The educator's word frequency guide* New York, NY: Touchstone Applied Science Associates, National Institute of Child Health and Human Development.

APPENDIX A

Original Source Materials for the VINE Word Corpus

Tradebooks:

Applegate, K. A. (1998). *Animorphs.* New York, NY: Scholastic.

Coerr, E. (2005). *Sadako and the 1000 Paper Cranes.* New York: Puffin Books.

Curtis, C. (2004). *Bud, Not Buddy.* Laurel Leaf Publishers.

Dahl, R. & Blake, Q. (1988) *James and the Giant Peach.* New York: Puffin Books.

DiCamillo, K. (2001). *Because of Winn Dixie.* Massachusetts: Candlewick Press.

DiCamillo, K. (2006). *The Tale of Despereaux: Being the Story of a Mouse, a Princess, Some Soup and a Spool of Thread.* Massachusetts: Candlewick Press.

Fleischman, S. (2000). *Bandit's Moon.* Yearling Publishers.

Fleischman, S. (1988). *Great Horn Spoon.* Little, Brown, & Young Readers.

Fletcher, R. (1997). *Twilight Comes Twice.* New York: Houghton Mifflin.

Hannigan, K. (2004). *Ida B: ... and Her Plans to Maximize Fun, Avoid Disaster, and (Possibly) Save the World.* New York: Harper Collins.

Hesse, K. (1997). *Out of the Dust.* New York: Scholastic.

Lowry, L. (1998) *Number the Stars.* New York: Bantam Doubleday Books for Young Readers.

March, C. (2003) *The Mystery on the California Trail.* Georgia: Gallopade International.
Mihelic, M. (2002). *Come My Gentle Ariel.* Didakta.
Mochiuki, B. (1997). *Baseball Saved Us.* New York: Lee & Low Books.
Munoz, P. (1999). *Riding Freedom.* New York: Scholastic.
Namioka, L & DeKiefte, K. (1995) *Yang the Youngest and His Terrible Ear* (Invitations to Literacy). Houghton Mifflin.
O'Dell, S. (1999). *Island of the Blue Dolphins.* Yearling Publishers.
Park, B. (2000). *Skinnybones.* New York: Random House Books for Young Readers.
Paulsen, G. (2006). *Hatchet.* Aladdin Publishers.
Ryan, P. (2000). *Esperanza Rising.* New York: Scholastic.
Soto, G. (1993) *Too Many Tamales.* New York: Putnam & Gosset Group.

Text Books:

Badders, W, Bethel, L, Fu, V., Peck, D., Sumners, C., Valentino, C. (2004) *Houghton Mifflin Science Discovery Works: Level 4 (Hardcover).* New York: Houghton Mifflin.
Banks, J., Beyer, B., Contreras, G., Craven, J., Ladson-Billings, G., McFarland, M., Parker, W. (2000). *California: Adventures in Time and Place.* New York. McGraw Hill School Division
Marht, A., Burton, G., Johnson, H., Luckie, L., McLeod, J., Newman, V., Schher, J. (2002). *Harcourt Math (California Edition).* Florida, Harcourt.
Slavick Frank, M, Jones, R, Kockover, G, Lang, M, McLeod, J, Valenta, C., VanDeman, B, (2000). *Harcourt Science.* Florida, Harcourt Inc.