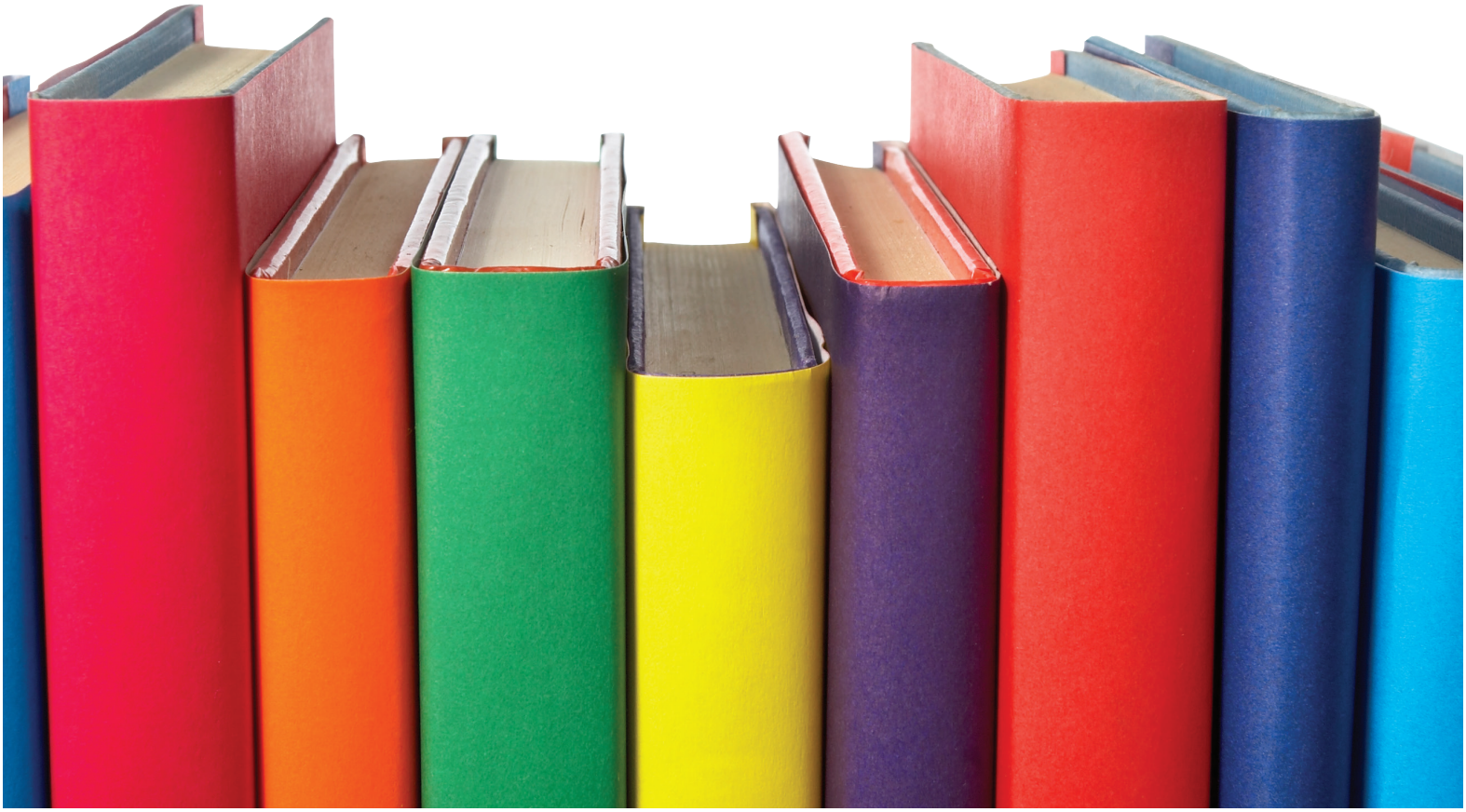
pals™

Phonological Awareness Literacy Screening

Marcia Invernizzi • **Joanne Meier** • **Connie Juel**
University of Virginia • Curry School of Education

1–3 Technical Reference



For questions about PALS 1–3, please contact:
Phonological Awareness Literacy Screening (PALS)
1-888-UVA-PALS (1-888-882-7257) or (434) 982-2780
Fax: (434) 982-2793
e-mail address: *pals@virginia.edu* • website: *https://pals.virginia.edu*

Graphic Design: Branner Graphic Design

Printed in the United States of America

pals™

Phonological Awareness Literacy Screening

1–3 Technical Reference

This document is a supplement to the
PALS 1–3 materials binder.

Marcia Invernizzi • Joanne Meier • Connie Juel
Virginia State Department of Education
University of Virginia • Curry School of Education

Acknowledgments

Development of the Phonological Awareness Literacy Screening for Grades 1–3 has been supported by the Virginia Department of Education through Virginia's Early Intervention Reading Initiative. Without the support provided by the Department, the test development activity required for this assessment would not be possible.

The PALS Office would like to thank Dr. Francis Huang at the University of Missouri for his contribution to the technical adequacy of PALS 1–3, and to Dr. Heather Warley for her editorial assistance. Thanks go also to division representatives, principals, and teachers throughout Virginia who have participated in the pilots. Thanks to their participation, the PALS office is able to ensure that classroom teachers have a literacy screening tool with good evidence of reliability and validity.

Section I

Section II

Section III

Section IV

Section V

Section VI

Section VII

Section VIII

Section I

Phonological Awareness Literacy Screening: Reading and Spelling Inventories for Grades 1–8 (PALS Plus)

In this section we
- provide an overview of the purpose and use of PALS Plus;
- show how PALS Plus supports Virginia's Standards of Learning (SOL);
- describe briefly the PALS Plus instrument.

More detailed information about the instrument is available from our website (pals.virginia.edu).

Purposes, Uses, and Limitations

The Phonological Awareness Literacy Screening for Grades 1–8 (PALS Plus) can identify students at risk of reading difficulties and delays. It can also assess what students know about words and what they need to learn next to become better readers.

With over 15 years of classroom testing and expert review, PALS has been shown to have good evidence of reliability and validity as an assessment of students' reading and writing skills. (See Technical Adequacy, pp. 29–50). However, like any other assessment tool, PALS Plus should be used as one among several potential sources of evidence about any given reader's overall competence. Instructional decisions are best based on multiple sources of evidence: reading assessment data from other kinds of tests; reading group placement; lists of books read; and, most important, teacher judgment.

Overview

Consisting of three screening instruments, the Phonological Awareness Literacy Screening (PALS-PreK, PALS-K, and PALS Plus for grades 1–8), measures young children's knowledge of important literacy fundamentals:
- oral passage reading
- word recognition in isolation
- spelling and morphology
- alphabet knowledge and letter sounds
- phonological awareness
- concept of word

The major purpose of PALS Plus is to identify students who are performing below minimal competencies in these areas and may be in need of additional reading instruction beyond what is provided to typically developing readers. Note that meeting the Entry Level Summed Score benchmark does not imply that the student is on grade level, but only that the student met the level of minimum competency necessary to benefit from typical classroom literacy instruction. A secondary and logical extension of this purpose is to provide teachers with explicit information about what their students know of these literacy fundamentals so that they can more effectively tailor their teaching to their students' needs.

The PALS Plus Technical Reference includes a description of the background and rationale underlying the assessment, the process through which tasks and items were developed and field tested, and the technical adequacy of the instrument (validity and reliability). In preparing the PALS Plus Technical

Table 1 PALS Plus and Virginia's SOL for Grades 1–8

PALS Plus Level	PALS Plus Task	Virginia SOL	Objective
Entry	Spelling	1.6h	Read and spell common, high-frequency sight words
		2.4a	Use knowledge of consonants, consonant blends, and consonant digraphs to decode and spell words
		2.4b	Use knowledge of short, long and r-controlled vowel patterns to decode and spell words
		3.10j, 4.8g, 5.8j, 6.8h, 7.8h, 8.8g	Use correct spelling of frequently used words
		3.4a.b., 4.4b, 5.4c	Use knowledge of roots, affixes, synonyms, antonyms, and homophones
		6.4a.b., 7.4a.b., 8.4c	Identify word origins and derivations. Use roots, cognates, affixes, synonyms, and antonyms.
	Word Recognition	1.6e	Blend beginning, middle, and ending sounds to recognize and read words
		2.4	Use phonetic strategies when reading and spelling
		3.3	Apply word-analysis skills when reading
		3.4f, 4.4e, 5.4g	Use vocabulary from other content areas when reading
		6.4f, 7.4f, 8.4f	Extend general and specialized vocabulary through reading
Level A	Oral Reading Accuracy	1.6	Apply phonetic principles to read
		1.7	Use meaning clues and language structure to expand vocabulary when reading
		1.7d	Reread and self-correct
		2.4, 2.5	Use phonetic strategies, meaning clues, and language structure when reading
		2.7	Read fiction and nonfiction, using a variety of strategies independently
		3.4	Use strategies to read a variety of fiction and nonfiction materials
		4.6l, 5.6m	Read nonfiction texts with accuracy
		6.6, 7.6, 8.6	Read a variety of nonfiction texts
	Fluency & Rate	1.8., 2.7c	Read familiar stories, poems, and passages with fluency and expression
		3.4e	Read fiction and nonfiction fluently and accurately
		4.6l, 5.6m	Read nonfiction texts with fluency
		7.2a	Use verbal communication skills, such as word choice, pitch, feeling, tone, and voice
		8.2b.c	Deliver oral presentations: choose appropriate tone, use appropriate verbal presentation skills
	Comprehension	1.9, 2.8	Read and demonstrate comprehension of fiction and nonfiction
		3.7	Demonstrate comprehension of information from a variety of print resources
		4.6c.d.f.i.j.k, 5.6.b.d.f.k.l, 7.6a.b.d.e.f.g.h.i.l, 8.6a.b.e.f.g.h.l	Read, comprehend, and analyze a variety of nonfiction texts

Table 1 (Continued)

PALS Plus Level	PALS Plus Task	Virginia SOL	Objective
Level B	Alphabet Recognition	K.7a	Identify and name the uppercase and lowercase letters of the alphabet
		1.5c	Identify letters, words, and sentences
	Letter Sounds	K.7b	Match consonant and short vowel sounds to appropriate letters
	Concept of Word	1.5b	Match spoken words with print
Level C	Sound-to-Letter	1.4	Orally identify and manipulate phonemes in syllables and multisyllabic words
	Blending	1.6e	Blend beginning, middle, and ending sounds to recognize words

Reference, we have followed current professional standards for educational tests.[1] Explicit instructions for the administration and scoring of PALS instruments are included in separate *Administration and Scoring Guides* for each instrument.

Background

PALS-K and PALS 1–3 were originally designated as the state-provided screening tools for the Virginia Early Intervention Reading Initiative (EIRI), and were specifically designed for use in kindergarten- through third-grade classrooms. The purpose of the EIRI is to reduce the number of children with reading problems by detecting those problems early and by accelerating the learning of research-identified emergent and early literacy skills among kindergarten, first-, second-, and third-grade students.

Although the EIRI is a voluntary initiative, the vast majority of Virginia schools opted to participate in an effort to reduce the incidence of reading problems in the primary grades. Over the years, many schools came to rely on the PALS Internet database system and began asking for PALS to extend through the upper elementary and middle grades. As a result, in 2012, the Virginia Department of Education provided funds to develop PALS Plus for use in grades 1–8.

Virginia's Standards of Learning (SOL) and PALS

The Virginia SOL for English in kindergarten and first grade were designed to enable students to become independent readers by the end of first grade.[2] Virginia's Early Intervention Reading Initiative provides further assistance for school divisions striving to meet that goal. The English Standards of Learning include many of the literacy skills assessed through PALS Plus. Phonemic awareness, alphabet knowledge, identification of letter sounds, concept of word, word recognition, oral reading in context, oral reading fluency, and reading comprehension are all listed in the Virginia SOL for English.

Table 1 illustrates the relationship between PALS Plus and the Virginia SOL for English in grades 1–8. These are fundamental components of the learning-to-read process. PALS Plus extends this relationship through grade 8. PALS Plus provides a straightforward means of identifying students who are relatively behind in their acquisition of fundamental literacy skills. Results from the PALS Plus screening also afford a direct means of matching reading instruction to specific literacy needs.

Section II

Description of PALS Plus

In this section we briefly describe the parts of PALS for Grades 1–8 (PALS Plus). Table 2 outlines the conceptual framework for the instrument.

Among the most effective strategies for preventing reading problems is first to identify early and accurately children who are experiencing difficulties in acquiring fundamental skills, and second to ensure that these children attain critical beginning literacy skills through additional instruction. This approach can be viewed as simultaneously proactive and preventative. Nevertheless, there will be students in the upper grades who still require ongoing intervention. PALS Plus for grades 1–8 is designed to identify such students and to provide the diagnostic information needed to effectively instruct them.

A substantial research base has suggested key variables that help identify children most likely to experience subsequent difficulties with reading achievement.[3] This research indicates that measures of phonological awareness, alphabet knowledge, letter-sound knowledge, and other elements of early literacy (e.g., phonetic spelling, word recognition) serve as robust predictors of children's later literacy achievement.

PALS Plus uses a three-tiered approach in which the first tier (or Entry Level) contains a routing appraisal that estimates a child's general level of skill in reading and spelling. The Entry Level tier also indicates the first required passage to be read in Level A. Level A assesses the accuracy, fluency, rate, and comprehension of a child's oral reading in context. Students can be identified as needing additional literacy intervention if they do not meet the Entry Level benchmark, or, for students in Grades 4–8, if they do not meet the criteria for accuracy, rate, and comprehension on the designated passage on Level A.

Level B assesses emergent and beginning reading essentials in alphabetic knowledge and concept of

Table 2 Conceptual Framework for PALS Plus

Level	Domain	Tasks
Entry Level	Orthographic Knowledge	Word Recognition
		Spelling
Level A	Oral Reading in Context	Oral Reading Accuracy
		Oral Reading Fluency
		Oral Reading Rate
		Oral Reading Comprehension
Level B	Alphabetics	Alphabet Recognition
		Letter Sounds
		Concept of Word
Level C	Phonemic Awareness	Blending
		Sound-to-Letter

word, and is taken only by students who do not have a measurable reading at the Preprimer or higher level. If Level B benchmarks are not met, children are routed to Level C for a more in-depth evaluation of phonemic awareness skills including blending and segmenting speech sounds.

Students demonstrate their skills in each domain to their classroom teacher, who administers PALS in the classroom (after reading the PALS Plus Administration and Scoring Guide). The performance-based tasks do not have a time limit; they are administered one-on-one, except for the Spelling task, which can be administered in small groups or in the class as a whole. Each task contains a criterion score or benchmark for a minimal level of competency. The benchmarks change from fall to spring.

Students in grades 1–3 who do not meet the Entry Level benchmark should receive a minimum of 2-1/2 hours of additional instruction each week for the equivalent of a school year, as per Virginia's Early Intervention Reading Initiative. Although not part of Virginia's Early Intervention Reading Initiative, students in grades 4–8 should also receive additional literacy instruction if they: (a) do not meet the Entry Level benchmark and/or (b) do not have an instructional reading level that is one level below grade-level in the fall, or on-grade level in the spring.

Two forms of PALS Plus are now in use. Forms A and B are used in alternate years. Form C is the optional mid-year form for grades 1–3 and will be available soon for grades 4–8. A description of how the criterion scores or benchmarks were established may be found later in this manual. The following section contains a detailed description of how PALS Plus items and tasks were developed and field-tested.

Section III

Item Development and Field-Testing

In this section we outline the various tasks included in PALS Plus:

- **Entry Level Tasks: Orthographic Knowledge**
- **Level A: Oral Reading in Context**
- **Level B: Alphabetics**
- **Level C: Phonemic Awareness**

The tasks presented in PALS Plus are a representative sample of tasks found in other measures of literacy achievement. Items were selected because of their standing in literacy research and because of their correlation to the Commonwealth of Virginia's Standards of Learning (SOL) in first through eighth grades.

Many of the tasks and items in PALS Plus are similar to other tasks in commonly used informal reading and spelling inventories. These tasks have been used for a number of years with hundreds of thousands of first through third grade children in Virginia, Wisconsin, Colorado and many other states across the country. Previous research on PALS 1–3 tasks[4] provides support for similar tasks on the PALS Plus expansion, which were piloted with thousands of students in grades 4–8 between 2012 and 2014.

Entry Level Tasks: Orthographic Knowledge

Orthographic knowledge refers to knowledge about the form of written words. Because written words are made of letters that represent speech sounds, and letter patterns that represent speech sounds and meanings, orthographic knowledge is impossible to achieve without knowing the alphabet and letter sounds, or without being able to attend to the speech sounds those letters represent. Thus, orthographic knowledge subsumes two of the most powerful predictors of early literacy achievement: (1) phonemic awareness, and (2) the alphabet. If a student demonstrates orthographic knowledge, he or she necessarily has cracked the alphabetic code. The two most cost-effective, time-efficient, and instructionally useful measures of orthographic knowledge are word recognition and spelling.

Word Recognition

The capacity to obtain meaning from print depends strongly on accurate, automatic recognition of core reading vocabulary at each grade level. As a result, PALS Plus provides ten benchmark word lists to gauge students' progress throughout the year: preprimer (pre-1), primer (1.1), end-of-first (1.2), end-of-second (2.2), end-of-third (3.2), fourth (4.2), fifth (5.2), sixth (6.2), seventh (7), and eighth (8) grades. The words on each list represent a random sample from a data- base of words created from a variety of sources.

Originally, word lists for grade one, two, and three were generated from a database of words created from basal readers most frequently used in the Commonwealth of Virginia. These included the Harcourt Brace Signature series and the Scott Foresman series from 1997 and 1999. Then, words from the first-, second-, and third-grade lists from the EDL Core Vocabularies in Reading, Mathematics, Science, and Social Studies (1997) were added to the database. The EDL Core Vocabularies provides a reading core vocabulary by grade, comprised of words derived from a survey of nine basal reading series. Words from the 100 Most Frequent Words in Books for Beginning Readers[5] were added to the primary and first-grade word pools.

The PALS 1–3 database was expanded to include word pools for grades one through six using words from grade-level lists in spelling and vocabulary books. These include words from *Teaching Spelling*,[6] *A Reason for Spelling*,[7] *A Combined Word List*,[8] *A*

Basic Vocabulary of Elementary School Children,[9] and S*pelling and Vocabulary*.[10] Our database now includes all of these words plus the words from graded word lists from informal reading inventories and other well-known published assessments that include grade-level lists. Words were added to the database from the *Qualitative Reading Inventory* (QRI-II),[11] the *Stieglitz Informal Reading Inventory*,[12] the *Bader Reading and Language Inventory*,[13] the *Decoding Skills Test*,[14] the *Ekwall/Shanker Reading Inventory*,[15] the *Book Buddies Early Literacy Screening* (BBELS),[16] and the *Howard Street Tutoring Manual*.[17]

Words were eliminated from the first through sixth grade-level word pools if they appeared on more than one grade-level list within the database. The remaining words were those that all sources agreed to be unique to that grade level. The validity of each word's grade-level placement was cross-checked for consistency within frequency bands in several word frequency reference sources such as *The American Heritage Word Frequency Book*.[18] Words on the preprimer and primer word lists appear in at least three of the word pool sources. Words on the first through sixth grade word lists appear in at least two of the word pool sources, and are unique to that specific grade level.

Different forms of the PALS 1–3 word lists were piloted between 2000 and 2005 with over 7,500 students in 246 first-, 194 second-, and 80 third-grade classrooms from over 55 different school divisions across all eight regions of Virginia. Student scores generated from these field tests were used to assess the reliability and validity of the word lists.

Each individual word on each list was analyzed using the following criteria:

- teacher feedback,
- amount of variance,
- item-to-total correlations, and
- Cronbach's alpha.

Words and/or word lists were considered for removal if they had alphas lower than .80, low item-to-total correlations, little to no variance, or if they received negative feedback from more than two teachers in the pilot sample. Words with low item-to-total correlations, little to no variance in response patterns, and/or negative feedback from teachers were substituted with words that had higher item-to-total correlations, moderate variance, and positive teacher feedback. In a few isolated cases, plural endings were changed to singular. Currently, three sets of graded word lists (Preprimer through the sixth grade level), with good evidence of reliability and validity are used in rotation across PALS screening windows.

The original word database used to establish the word lists for PALS 1–3 was expanded for the development of the seventh and eighth grade words lists. These words were collected from student reading anthologies (basal readers and seventh and eighth grade literature anthologies), spelling and vocabulary lists, and informal reading inventories to create an extensive list of unique words at each grade level. Additionally, all the major corpora were consulted for words within frequency bands associated with the higher grade-levels. These included: *The Corpus of Contemporary American English: 425 Million Words,*[19] *1990-Present; The Educator's Word Frequency Guide;*[20] *The Living Word Vocabulary;*[21] and *Words Worth Teaching*.[22]

However, word difficulty is much more complex than frequency of occurrence alone. Words differ by number of phonemes, by number of syllables and morphemes, by parts of speech and syntactic categories. They also differ in frequency of occurrence in written versus spoken language, referential concreteness, imageability, and other dimensions.[23] These features are particularly critical for longer words. Therefore, the word lists were also analyzed linguistically using the MRC Psycholinguistic Database[24] and balanced for all of these attributes.

From this expanded database, the word lists for each grade level were created with balanced SFI (Standard Frequency Index) totals and linguistic attributes. The word lists were piloted across four assessment windows to establish reliability and validity. The number of students participating in each pilot ranged

from 4,150 (in fall 2013) to 8,860 (in fall 2012). Each word list contained approximately eight additional items to pilot so that poorly performing items could be excluded. Each round of analyses included the calculation of difficulty indices, item-to-total correlations, calculation of alpha coefficients, and discrimination indices. Items were ranked according to how discriminating they were between good and poor performers. The good and poor performers were defined using two separate measures: the overall score using the word lists and an external measure, the Virginia Standards of Learning (SOLs). Words were considered for removal if they had discrimination indices lower than .30, low item-to-total correlations (i.e., point biserial correlations), little to no variance, or if they received negative feedback from more than two teachers in the pilot sample.

Item bias was also reviewed through the use of differential item functioning (DIF) based on gender and race/ethnicity groups using a Mantel-Haenszel procedure to flag misbehaving items[25]. Flagged items were then reviewed and assessed for further validation. To help with summarizing results, the Educational Testing Service (ETS) DIF classifications were used (i.e., A = negligible, B = slight to moderate, C = moderate to large, - and + indicating against the focal or reference group, respectively).

Words from both forms were combined and re-sorted into separate forms based on word difficulty to create a balanced assessment (parallel forms). A small portion of words were retained as common, linking items and appear in both forms. Tables 1 and 2 in the Appendix present the item analyses for extended word lists.

Spelling

Application of letter-sound knowledge in invented spelling tasks is an excellent predictor of word recognition in young children[26] and among the best predictors of word analysis, word synthesis[27], and even reading comprehension[28]. Research on how children learn to read and spell words in an alphabetic orthography has consistently revealed that orthographic features are internalized in a systematic, developmental progression. Invented spellings provide a diagnostic window into students' understanding of alphabetic orthography and can help teachers determine when to teach which phonics or spelling features of English orthography.[29]

According to this body of research, the acquisition of basic orthographic features within one-syllable words occurs in the following progression: beginning consonants; ending consonants; consonant digraphs; medial short vowels in simple three-letter words; consonant blends; pre-consonantal nasals; silent-e marker for long vowels; other long vowel patterns; r- and l-influenced vowel patterns; ambiguous vowel- diphthongs and digraphs; syllable structures; affixes; and morphemes of Greek or Latin derivation. Although students vary in their rate of skill acquisition, the order of acquisition is more or less the same, though some of the early features may be learned simultaneously.[30]

Words for the PALS Plus spelling inventories were selected from a pool of words used in previous research in the Virginia Spelling Studies.[31] Specific words were chosen by frequency of occurrence for each grade level and informed by developmental spelling theory in regards to grade-level expectations. That is, we selected specific words to recreate the progression of phonics/spelling features acquired by typically achieving students in the course of their schooling. Grade-level expectations for word features are outlined in Table 3. Features that are often acquired simultaneously are shaded. Examples of each feature are shown in the second column.

We selected four words for each feature category, each one within a similar frequency band. Forms were field-tested with over 6,800 kindergarten through third-grade students in 55 different school divisions and across all eight regions of Virginia. For grades 4–8, alternate forms were field tested with nearly 9,000 fourth through eighth grade students in 279 different schools across all eight geographical regions of Virginia. All of the pilot tests assessed stu-

dent performance on specific orthographic features, the number of words spelled correctly, and a combination of the two. For each word, students received a point for the presence of a specific orthographic feature, whether the entire word was spelled correctly or not. Another point was awarded if the entire word was spelled correctly. In this way, students were credited with the application of phonics principles regardless of whole-word spellings.

In every pilot, we analyzed each individual word on each spelling list using the same criteria we used for analyzing the internal consistency of the words in the graded word lists (teacher feedback, amount of variance, item-to-total correlations, item difficulty, discrimination indices, and Cronbach's alpha). Words were removed if they had low item- to-total correlations, poor discrimination capabilities, or little to no variance.

Item bias was examined through differential item functioning analyses broken down by student gender and race/ethnicity groups. Flagged items were reviewed for potential bias using a Mantel-Haenszel procedure to flag misbehaving items (Holland & Thayer, 1986) To help with summarizing the results, the Educational Testing Service (ETS) DIF classifications were used (i.e., A= negligible, B= slight to moderate, C= moderate to large, -and + indicating against the focal or reference group respectively). A small proportion of words (< 3% for gender and < 1% for race/ethnicity) were classified as having moderate to large evidence of DIF though differential items were counterbalanced between focal (i.e. female, white) and reference (i.e. male, nonwhite) groups and were further examined for potential bias.

Tables 3 and 4 in the Appendix show the percent of items by DIF classification for gender and race/ethnicity, using the ETS classifications. As with the word list pilots previously described, the number of pilot participants for the phonics/spelling words for grades 4–8 ranged from 4,150 (in fall 2013) to 8,860 (in fall 2012).

We also developed sentences for teachers to use when calling out words for their students to spell. The sentences were piloted along with the spelling word lists, and teacher feedback informed changes.

Table 3 Grade-level Expectations for Phonics/Spelling Features

Phonics/Spelling Feature	Example	End of K	End of 1st	End of 2nd	End of 3rd	End of 4th	End of 5th	End of 6th	End of 7th	End of 8th
Beginning Consonants	**m**op	✓	✓	✓	✓	✓	✓	✓	✓	✓
Ending Consonants	wi**g**	✓	✓	✓	✓	✓	✓	✓	✓	✓
Consonant Digraphs (ch, th, sh)	**ch**in		✓	✓	✓	✓	✓	✓	✓	✓
Medial Short Vowels (CVC)	n**e**t		✓	✓	✓	✓	✓	✓	✓	✓
Consonant Blends	**tr**ap		✓	✓	✓	✓	✓	✓	✓	✓
Pre-consonantal Nasals	bu**mp**		✓	✓	✓	✓	✓	✓	✓	✓
Silent-e, Long Vowel Markers (CVCe)	sl**ide**			✓	✓	✓	✓	✓	✓	✓
Other Long Vowel Patterns	p**ai**nt, cl**ea**n				✓	✓	✓	✓	✓	✓
R- and L-influenced Vowels	st**ar**t, h**ur**t				✓	✓	✓	✓	✓	✓
Ambiguous Vowels	c**au**ght, sp**oi**l				✓	✓	✓	✓	✓	✓
Syllable Juncture	dro**pping**					✓	✓	✓	✓	✓
Affixes	**coo**perate						✓	✓	✓	✓
Derivations	ser**e**nity								✓	✓

Level A: Oral Reading in Context

Listening to students read aloud from graded passages provides direct information for estimating reading levels, diagnosing strengths and weaknesses, and evaluating progress.[32] This process allows teachers to determine a student's instructional reading level: the level at which he or she can profit from instruction. The reading selections for the primer through sixth grade levels were modeled on non-fiction basal reading passages published before 1990 and were written by local children's authors.

Passage Selection

For primer through sixth-grade level passages, we used non-fiction basal passages published prior to 1990 as models for several reasons. First, several reviews and meta-analyses of basal reading series have noted a relaxing of vocabulary control after 1990[33]. Some researchers have suggested that the lack of vocabulary control in basal readers published after 1990 obscured the classification of text difficulty by grade level. By imitating basal passages prior to 1990, we sought to achieve the vocabulary control pivotal to the historical construct of grade-level text. For each grade-level passage we targeted the end-of-year difficulty of the older basal readers: 1.2 for end of first, 2.2 for end of second, 3.2 for end of third, etc. Second, we wanted nonfiction passages to avoid the cultural bias inherent in narratives. Finally, we wanted to match topics represented in Virginia SOL for Science for each grade level.

We contracted with local children's literature experts to write nonfiction passages on topics represented in the Science SOLs for each grade level 1 through 6. These writers referred to grade-level word lists and grade-level science concepts, and incorporated them into the passages. They were also given general guidelines as to length, syntactic complexity, and word frequency.

The preprimer passages ("little books") were written by members of the PALS staff and illustrated before being piloted in the public schools. In writing the preprimer pas- sages, we used our collective experience derived from 26 years of teaching emergent and beginning readers using leveled texts. We also relied on published descriptions of books for beginning readers.[34] We paid attention to issues relating to the quantity and repetition of words, sentences, oral and written language patterns, and vocabulary. We were attentive to the layout and print features of the book. We attended to children's familiarity with objects and actions as well as story structures and related elements such as predictability, dialogue, and plot. We counted the number of syllables within words, the number of words within sentences, the number of sentences on a page, and the number of all of these per book. In addition, we were mindful of the number of decodable and high-frequency words, and we counted the number of phonic and morphemic elements such as consonant blends and digraphs, past-tense markers, and inflections. Before we piloted passages, several teachers read and critiqued them, and we used readability formulas to verify our qualitative approach. Three out of five readability formulas agreed that our resulting "little books" represented the targeted level. Readability analyses revealed a gradual increase in overall difficulty from level to level. An example of the gradual increase of quantitative and qualitative text features across piloted preprimer books corresponding to the readiness, preprimer A, preprimer B, and preprimer C text levels may be found in Table 4.

Readability Analysis. To confirm the readability of the passages, we subjected each selection to multiple readability formulae. These included the Flesch-Kincaid Reading Ease[35], the Spache Readability Formula, using the formula and the chart methods;[36] the Harris- Jacobson;[37] the Wheeler-Smith readability formula;[38] and the Fry Formula for Estimating Readability.[39] Each readability level was calculated by hand. Two formulae (Spache and Harris-Jacobson) used a combination of sentence length and specific word lists to determine readability, so their estimates were usually very similar. Next, the passages were subjected to measures of text complexity suggested by Common Core: Advantage/TASA Open Standard

(ATOS)[40], Degrees of Reading Power (DRP)[41], and Lexiles[42]. We also paid special attention to length, to ensure that each successive passage was longer than the easier one before it. Other guidelines for constructing informal reading inventories were followed: specifically, those suggested by Johnson et al. (1987), Lipson and Wixson (1997), and Stauffer, Abrams, and Pikulski (1978). Table 5 presents the readability characteristics of the PALS Plus passages.

Field testing. While all of the preprimer and primer passages met the readability criteria, we field-tested multiple passages at each level in grades one, two, and three in three separate large-scale pilots involving a total of over 4,000 students. We asked a subsample of these students to read three passages in succession, each one on a more difficult level than the one before. We then checked to make sure that the number of errors students made while reading the passages increased in accord with incremental increases in the difficulty of the text itself. Accuracy scores for oral reading were computed by dividing the number of words read correctly by the total number of words per passage. We used these accuracy scores to compare passages with each other. Where two passages were piloted on a given level, we chose the superior passage at each level. Superior passages had (a) higher correlations with previously established PALS 1–3 passages

Table 4 Text Features Across Form A Preprimer Level Texts

Text Feature	Readiness	Preprimer A	Preprimer B	Preprimer C
# Words	15	35	53	102
# Sentences	5	10	13	14
# Words per sentence	3	3.5	4.1	7.3
# Sentences per page	1	2	2.6	2.3
# Pages per book	5	5	5	6
# Two-syllable words	0	2	5	8
% Decodable and high-frequency words	58%	60%	74%	77%
# Consonant blends and digraphs, inflections, and ambiguous vowel sounds	1	8	10	15
% Grade-level vocabulary	100% K	100% K	95% K 5% 1st	94% K 4% 1st 2% Other
Pattern	One word change per page	Two-word change	Dialogue repetition	Question, guessing, answer
Characters and objects	Cat, dog, mouse, man	Bears, birds, cows, rabbits	Pig, fox, barn	Pig, dog, cat, fox, apple, ball, orange, yo-yo, drum
Actions	Run	Sleep, eat, hop, fly	Run under, run into	Ask, saying, guessing
Ideas	Simple	Simple	Setting-dependent	Dialogue-dependent
Predictability	High	High	Medium high	Medium
Plot	Simple recount	Simple description	Simple narrative: problem, climax, resolution	Complex narrative: problem, Q & A events, resolution

Table 5 Readability Characteristics of PALS Plus Passages

Level	Form	Title	Words	Sentences	Words/ Sentence	Spache	Harris-Jacobson	Wheeler-Smith	Fry	F-K Reading Ease	Lexile*	ATOS*	DRP*
Primer	A	Zack the Monkey	114	18	6.3	1.4	1.3	4.4 (P)	NA	100	230L	NA	NA
	B	A Bear Cub in Spring	120	18	6.7	1.7	1.7	7.2 (P)	1st	100	280L	NA	NA
	C	What is a Pet?	107	15	7.1	2.1	1.8	8.9 (1st)	1st	100	340L	NA	NA
1st	A	Dressed for Winter	143	21	6.8	1.8	1.9	7.5 (P)	1st	100	400L	NA	NA
	B	Where do Animals Live?	137	20	6.9	1.9	1.4	11.0 (1st)	1st	97	410L	NA	NA
	C	Animal Coverings	141	20	7.1	1.9	1.9	16.5 (2nd)	2nd	85	460L (2nd-3rd)	NA	NA
2nd	A	Country Music	198	21	9.4	2.9	2.7	18.6 (2nd)	3rd	94	590L (2nd-3rd)	3.6 (2nd-3rd)	46 (2nd-3rd)
	B	Nature's Magician	198	23	8.6	2.9	2.4	18.2 (2nd)	3rd	84	490L (2nd-3rd)	3.6 (2nd-3rd)	45 (2nd-3rd)
	C	Deep in the Ocean	197	21	9.2	2.9	2.8	15.7 (2nd)	3rd	94	550L (2nd-3rd)	3.6 (2nd-3rd)	45 (2nd-3rd)
3rd	A	Ocean Cities	232	23	10.1	3.9	3.7	26.2 (3rd)	4th	84	640L (2nd-3rd)	4.9 (2nd-3rd)	50 (2nd-3rd)
	B	The World of Birds	231	24	9.6	4.1	3.8	21.7 (3rd)	3rd	90	660L (2nd-3rd)	4.1 (2nd-3rd)	49 (2nd-3rd)
	C	Clever Creatures	226	22	10.3	3.7	3.7	26.4 (3rd)	5th	83	670L (2nd-3rd)	4.9 (2nd-3rd)	52 (2nd-3rd)
4th	A	Animal Forecasters	286	26	11	4.9	4.6	31.8 (4th)	6th	75.3	760L (2nd-5th)	5.8 (4th-5th)	58 (4th-8th)
	B	Animals of the Night	289	26	11.1	5.9	4.6	31.1 (4th)	5/6th	81	750L (2nd-5th)	5.8 (4th-5th)	54 (2nd-5th)
	C	Helping Paws	296	25	11.8	4.9	4.6	34.5 (4th)	7th	78.6	810L (2nd-5th)	5.2 (4th-5th)	53 (2nd-5th)
5th	A	Miniature Marvels	286	23	12.5	5.8	5.9	33.7 (4th)	5th	80.4	880L (4th-5th)	6.5 (4th-5th)	55 (4th-5th)
	B	Fossils	286	23	12.4	5.5	5.2	NA	6th	75.7	830L (4th-5th)	7.2 (6th-8th)	58 (4th-8th)
	C	Alaskan Journeys	298	23	12.9	5.7	5.9	29.6 (4th)	5th	82.9	880L (4th-5th)	5.8 (4th-5th)	55 (4th-5th)
6th	A	Space Dogs	294	22	13.4	6.9	6.9	NA	6th	72.8	940L (4th-8th)	7.4 (6th-8th)	61 (6th-8th)
	B	Sloth for a Day	298	21	14.2	6.5	6.4	33.3 (4th)	6th	82	900L (4th-5th)	7.1 (6th-8th)	56 (4th-5th)
	C	Hope for Habitats	301	21	14.2	6.6	6.7	NA	6th	69.3	970L (4th-8th)	8.8 (6th-8th)	61 (6th-8th)
7th	A	Sharks	301	20	15.1	NA	8.0	NA	8th	73.4	1060L (6th-10th)	9.0 (6th-8th)	65 (6th-10th)
	B	Jellyfish	287	19	15.2	NA	7.8	NA	8th	68.2	1060L (6th-10th)	8.8 (6th-8th)	61 (6th-8th)
	C	Penguins	300	20	15.0	NA	7.6	NA	8th	67.7	1060L (6th-10th)	8.9 (6th-8th)	65 (6th-10th)
8th	A	Tornados	326	19	17.2	NA	8.0	NA	10th	61.0	1100L (6th-10th)	9.6 (6th-8th)	65 (6th-10th)
	B	Lightning	329	19	17.3	NA	8.0	NA	9th	64.0	1070L (6th-10th)	9.9 (6th-10th)	66 (6th-10th)
	C	Volcanoes	325	19	17.1	NA	7.8	NA	9th	61.0	1090L (6th-10th)	9.9 (6th-10th)	64 (6th-10th)

**Note. Ranges in parentheses are Common Core bands.*

(equivalent forms), (b) higher correlations with the word lists, (c) a better progression from one level to the next, and (d) more positive teacher feedback.

Fine tuning. After the passages were selected, minor modifications were made to improve them, based on student performance and teacher feedback. For example, based on cross-level oral reading accuracy scores, it appeared that the original versions of the Form A preprimer B passage and the preprimer C passage were close in difficulty. We then analyzed the oral reading records (running records) generated from those passages and identified certain words that were problematic for everyone who read them. As a result of these analyses, selected words were changed in both passages to provide a better gradation of difficulty from one passage to the next.

After these modifications, a second, smaller pilot was conducted with 262 students, to retest the modified passages. Between 96% and 98% of first-, second-, and third-grade students who could read 15 or more words on the PALS word lists could also read with at least 90% accuracy the PALS passage corresponding to the same grade level.

In the spring of 2001, 2004, and 2005, we conducted additional pilot testing to confirm the readability of selected passages and to establish their degree of equivalence to PALS passages from other forms. Altogether, more than 6,000 students in grades one through three were asked to read the new passages aloud while their teachers took running records to note their accuracy. Feedback was elicited from all teachers regarding the suitability of each passage for their grade levels, the coherence of the text and illustrations, the clarity of directions, and the ease of administration and scoring. Illustrations were modified according to teacher feedback, as were the directions and rubrics for administration and scoring. In these pilots, we examined the extent to which passages from different forms are of similar difficulty by simply checking mean scores for students reading both forms of a given passage. We further examined the extent to which passages from preprimer to sixth grade appropriately increase in level of difficulty by examining the scores of students who read multiple passages (e.g., preprimer, primer, and first-grade; or second-, third-, and fourth-grade passages).

Expansion of Passages to Grades 7 and 8. The procedure used to create the seventh and eighth grade passages was very similar to the development of the original eleven readiness through sixth grade passages. For each additional grade-level, three science-related topics were chosen from among the State Standards and researched. Passage-writers referred to the corresponding grade-level word lists and incorporated grade-level word control into passages whenever possible. General guidelines as to length, syntactic complexity, and word frequency were followed to maintain the progression of difficulty established in the existing passages up through sixth grade.

The passages were subjected to the same readability formulae used for the primer through sixth grade passages. Because of the more advanced text structure of the seventh and eighth grade passages, they also were put through extensive linguistic cohesion analysis using COH-METRIX[43] software. Specifically, the passages were calibrated for narrativity, syntactic complexity, lexical density, and deep and referential cohesion. These additional analyses of text complexity and cohesion can be found in Table 5 of the Appendix. Each passage was piloted in four assessment windows, during which students were asked to read two to three successive passages. Again, the number of pilot participants for the passages ranged from 4,150 (in fall 2013) to 8,860 (in fall 2012).

After each pilot, the student data were analyzed for evidence of "stair steps" or increments of difficulty—notably, a decrease in word-reading accuracy for each increment in text difficulty and an increase in the amount of time it took students to read the text with each increment in text difficulty. Each word was also analyzed to see if there were particular words everyone got wrong no matter what their reading level was.

After each analysis, the wording of the passages was modified and then re-piloted. These steps were repeated until there was a decrease in accuracy and an increase in time for each incremental step-up in text difficulty. The end result was four new passages with empirically established increments of text difficulty associated with the seventh and eighth grades, using a variety of quantitative readability formulae as well as qualitative linguistic analyses of text complexity and empirically determined evidence of text difficulty from the students' reading accuracy and reading rate.

Fluency

Another step in developing the oral reading task was to expand the characterization of acceptable oral reading to include aspects of reading fluency beyond the accurate recognition of words in context. To develop a simple rating scale for fluency that incorporated aspects of pacing, phrasing, and expression, we adapted the four-point fluency rating scale used by the National Assessment of Educational Progress (NAEP).[44] The NAEP oral reading fluency scale rates the syntactical appropriateness of phrasing and meaningfulness of expression. For screening purposes, we reduced NAEP's four-point scale to three points by combining levels 1 and 2, the most dysfluent levels according to NAEP. A comparison of the NAEP Oral Reading Fluency Scale and the PALS Plus Oral Reading Fluency Scale may be seen in Table 6.

Reading Rate

Many assessments measure reading rate because it is a quick and easily quantifiable aspect of reading fluency, and many include benchmarks derived from norms that compare children's reading rates to other children in the same grade. However in PALS Plus, reading rate serves a different purpose. Rather than to compare children to other children, reading rate is used in PALS Plus to help teachers match children to text. Rather than focusing on a comparative norm for *grade levels*, PALS Plus focuses on minimal reading rates necessary to read successfully at each *instructional reading level*,

Table 6 A Comparison of Fluency Rating Scales: NAEP and PALS Plus

NAEP Oral Reading Fluency Scale Description	PALS Plus Fluency Rating Guide
Level 4 Reading primarily in larger, meaningful phrase groups. Although some regressions, repetitions, and deviations from text may be present, these do not appear to detract from the overall structure of the story. Preservation of the author's syntax is consistent. Some or most of the story is read with expressive interpretation.	**Level 3** Meaningful phrase groups; expressive, fluent
Level 3 Reads primarily in three- or four-word phrase groups. Some smaller groupings may be present. However, the majority of phrasing seems appropriate and preserves the syntax of the author. Little or no expressive interpretation is present.	**Level 2** Awkward phrase groups; moderate pacing; little/no expression
Level 2 Reads primarily in two-word phrases with some three- or four-word groupings. Some word-by-word reading may be present. Word groupings may seem awkward and unrelated to larger context of sentence or passage. Reads primarily word by word.	**Level 1** Word-by-word; laborious, monotone
Level 1 Occasional two-word or three-word phrases may occur, but these are infrequent and/or do not preserve meaningful syntax.	

NAEP's oral reading fluency scale is from the U.S. Department of Education, National Center for Educational Statistics, Listening to Children Read Aloud: Oral Fluency, *1 (1), Washington, D.C., 1995.*

primer through eighth. Instead of using reading rates to label children in categories of risk, PALS Plus uses reading rates diagnostically to ensure an optimal designation for an appropriate instructional reading level.

Minimal rates were derived from analyses of students' reading rates on instructional reading levels, not grade levels. We looked at instructional reading levels as determined by accuracy of word reading in isolation and in context, as well as by comprehension scores. We removed outliers (using the 1.5 interquartile range) and plotted descriptive statistics and dispersion of reading rates for each instructional reading level across the grades. We then examined the lower end of each rate distribution for each reading level for each grade to come up with a minimal reading rate for each instructional reading level. Reading rates are measured on PALS Plus in words per minute, calculated by multiplying the number of words in the passage by 60 and dividing by the total reading time in seconds.

In grades 1–3, students who do not meet the *minimal reading rate* for a given reading level receive an asterisk on their score report indicating that their reading speed is too slow for that reading level and teachers are encouraged to administer easier passage levels. In grades 4–8, if students do not meet the minimal reading rate for a given reading level, their instructional reading level is bumped back to a level that does meet the minimal rate. On PALS Plus, reading rates are used diagnostically to match students to the optimal level of text difficulty.

Comprehension

Comprehension of what we read is why we read. Students who are good at monitoring their own comprehension know when they understand what they have read and when they do not. Students who are not good at monitoring their comprehension may not be so aware of their deficiency. The research base on reading comprehension suggests that text comprehension can be improved by instruction that helps students use specific comprehension strategies to monitor their understanding.

The PALS Plus comprehension questions provide an opportunity for teachers to explore their students' comprehension. By asking students questions directly following their reading, teachers may assess the degree to which students understand what they read and, if they do not, where the breakdown in understanding occurred.

The comprehension questions for earlier versions of PALS consisted of open-ended questions that were written according to recommended guidelines for constructing an Informal Reading Inventory.[45] According to these guidelines, questions should follow the order of the text and should contain a balance of factual, main idea, inference, and vocabulary questions. Questions that can be answered by relying on background knowledge should be eliminated or kept to a minimum.

We piloted comprehension questions in Spring and Fall 2001, Spring 2003, and Spring 2004 with the same students who participated in the Oral Reading in Context pilots. Teachers evaluated the questions and, based on their feedback, we added, eliminated, or changed the wording of some. Feedback also indicated that open-ended questions were difficult to score. Questions arose regarding the assignment of half-points or quarter-points, and many teachers felt insecure about probing a student for more precise answers. As a result, PALS comprehension questions were rewritten into a multiple-choice format to reduce scoring error. Because the Virginia SOL test for Reading at the end of third grade also contains multiple-choice questions about passages just read, we reasoned that it would be beneficial for students to experience the multiple-choice format earlier in the grades in a more supportive context (because the PALS questions are administered in a one-on-one setting).

Expansion of Comprehension Questions to Grades 7 and 8. Comprehension questions were also developed for each of the additional passages at grades 7 and 8 and piloted in an iterative fashion along with the passage reading. Approximately six additional

comprehension items were piloted to allow for the elimination of poorly performing items. In addition, different forms of multiple choice responses were piloted (e.g. single answer, two answers, search and locate, paraphrase, etc.). Questions included a balance of literal and inferential questions and at least two vocabulary questions per passage.

Participants were the same participants as previously described for the passage reading in grades 4–8. After each pilot, student responses to comprehension questions as well as the comprehension format were reviewed and modified as necessary. Items were reviewed using item difficulty, discrimination, and overall item-to-total correlation. Items that did not perform well on all characteristics were flagged, reviewed, and modified. Table 6 in the Appendix summarizes the item analyses for the comprehension questions for the seventh and eighth grade passages.

Level B: Alphabetics

Alphabetics includes two important aspects of alphabet knowledge and concept of word in text. The two alphabetic tasks consist of (a) letter recognition or naming and (b) recognition of letter-sound relationship. Both tasks emphasize alphabet recognition and phonemic awareness, which are the two best predictors of how easily children will learn to read in the first two years of instruction. Phonemic awareness is tested specifically in Level C tasks, but pronouncing letter sounds in isolation, expected at Level B, also requires explicit awareness of individual phonemes. Since the sounds are produced in response to printed letters, however, letter-sound recognition is primarily a phonics task that entails awareness of individual phonemes.[46] The Concept-of-Word task included in Level B is the culmination of alphabet knowledge and phonemic awareness in a real reading context (see p. 20). Research has demonstrated that the ability to fully segment all the phonemes within words follows concept-of-word attainment.[47]

Alphabet Recognition

The single best predictor, on its own, of early reading achievement is accurate, rapid naming of the letters of the alphabet.[48] In the first PALS cohort, 52,660 kindergarten and first-grade children were individually asked to name all of the letters of the alphabet, in both upper and lower case.[49] Children were asked to name a series of 26 randomly presented letters, first in upper case, then again in lower case. Item analyses from that statewide sample demonstrated ceiling effects for upper-case recognition among first graders. Since upper-case recognition and lower-case recognition were significantly and highly correlated (r = .94 for the kindergarten sample and .83 for first grade), and no ceiling effects occurred for lower-case letters, PALS was revised to include alphabet recognition for lower-case letters only. Teacher feedback from subsequent administrations also prompted a change in the order of letter presentation. Previously, the first alphabet item encountered was a lower-case *b*, a letter frequently confused with lower-case *d*. In PALS Plus, the first item encountered is an *m*.

Letter Sounds

Pronouncing the sounds represented by individual letters in isolation is difficult for young children and requires explicit awareness of individual phonemes. Since young children recognize upper-case letters more accurately than lower-case letters, PALS assesses knowledge of grapheme-phoneme correspondences using upper-case letters only. Originally, all of the upper-case letters were used with the exception of *X* and *Q*, since neither of these letters can be pronounced in isolation. We substituted *Qu* for *Q* and *Sh* for *X*. In the most recent version we replaced *Qu* with *Ch*, a more frequently occurring digraph, and we replaced *M* with *Th*. *M* became the letter used as an example in the directions.

In the Letter Sounds task, teachers ask children to touch each letter and say the sound it represents. Teachers may ask for the alternate sound for a letter that has two sounds. Only the lax, or short vowel sound, for each vowel is scored as correct, and only the hard sound for *C* and *G* is considered correct.

Ten statewide administrations of the Letter Sounds task confirm the power of this simple task to identify students who need additional instruction in phoneme-grapheme correspondences.

Concept of Word

Concept of word refers to the fledgling reader's ability to match spoken words to written words as he or she reads, as indicated by the accuracy of the child's finger-pointing to individual words as they are spoken.[50] Concept of word attainment is a water- shed event in learning to read.[51] It is the integration of alphabet recognition, letter sounds, initial phoneme segmentation, and word boundaries in text.

Research has shown that a stable concept of word in text facilitates a child's awareness of the individual sounds within words. Until a child can point to individual words accurately within a line of text, he or she will be unable to learn new words while reading or to attend effectively to letter-sound cues at the beginning of words in running text.[52] Concept of word is included in Level B (Alphabetics) of PALS because of its significance in the learning-to-read process. A solid concept of word differentiates emergent readers from beginning readers and is addressed in first grade English SOL 1.5b.

In 1997, 34,848 kindergarten students and 3,586 first-grade students were administered a Concept of Word task. Qualitative feedback from the field indicated that some children were unfamiliar with the content of the text used that year, which featured a farmer and a goat. Although familiarity with the story content would not have affected the outcome of the measure, the content was changed in subsequent versions of PALS to public domain folk rhymes, presented in a book format, one line to a page.

Multiple rhymes were field-tested with 1,405 end-of-year kindergartners and first-graders in Spring and Fall 2001, 1,779 kindergartners in Fall 2003, and 1,280 kindergartners in Spring 2004. Rhymes were selected for use if they received positive feedback from the pilot teachers and yielded reliability coefficients in the range of .80 or higher.

Words from the nursery rhyme are post-tested after the finger-pointing exercise, to see if any words were "picked up" in the process. The COW Word List sub- score at the end of kindergarten is highly correlated with the Word Recognition in Isolation score a child receives on the test administered at the beginning of first grade.

Level C: Phonemic Awareness

Phonemic awareness refers to the ability to pay attention to, identify, and manipulate phonemic segments in speech-sound units that roughly correspond to an alphabetic orthography. This awareness develops gradually over time and has a reciprocal relationship to reading. Children who have phonemic awareness learn to read more easily than children who do not. At the same time, instruction in alphabetic coding increases a child's phonological awareness.

Level C includes two measures of phonological awareness at the phoneme level: (a) a phoneme blending task (Blending) and (b) a segmenting task (Sound-to-Letter). The individual tasks and items in the phonological awareness portion of PALS were selected to represent three attributes of measurement. First, the tasks and items selected needed to represent a sufficient range of difficulty to avoid floor and ceiling effects. Previous research has demonstrated that some phonological awareness tasks are easier than others.[53] Second, the tasks selected needed to have a strong predictive relationship to reading outcomes.[54] Third, the tasks selected needed to assess two kinds of phonological awareness: (a) speech analysis at the phonemic level and (b) the transfer of phonemic awareness to letters. The latter assesses the utility of phonemic awareness in learning an alphabetic orthography.[55]

Blending

The Blending task is a phonological processing task. The task requires a student to use information from

the sound structure of speech to retrieve words. When administering this task, the teacher vocalizes specific sounds and asks the student to put them together and identify a word. The teacher slowly stretches out each separate phoneme. For example, the teacher might say "/s/ /a/ /t/" and the student responds by blending the sounds together to produce the word "sat." Previous research on phoneme blending indicates that a student's performance on blending tasks is predictive of how well he or she will read several years later.[56] Careful consideration was given to the individual items comprising the phonological awareness tasks. All words selected are in the core vocabulary of first-grade children and were listed in the only comprehensive corpus of first-grade children's speaking vocabulary.[57] Items were arranged in a developmental sequence; that is, items progress from easy to more difficult in terms of number of phonemes and phoneme location. For example, the first set of items consists of 2-phoneme words, the next set consists of 3-phoneme words, and the two final sets consist of 4-phoneme words. Previous research has demonstrated the developmental nature of phonological awareness tasks related to the number and location of phonemes.[58]

Further consideration was given to the linguistic complexity of the items. Matters pertaining to coarticulation, place of articulation, manner of articulation, and phonological ambiguity were taken into account. For example, on the Blending task, the first set of 2-phoneme words all begin with continuants, contain clear front vowels, and maintain respectable phonological distance in the progression from one item to the next. For example, lax (or "short") vowel sounds closest in place of articulation are not placed next to each other (e.g., short /e/ and short /a/). The first set of 4-phoneme words begin with high-frequency blends. All beginning blends also start with continuants. Care was taken to avoid phonologically ambiguous vowel sounds and to maintain phonological distance between contiguous items. The same linguistic characteristics were considered for the last set of items, 4-phoneme words with ending blends where there are slightly closer vowel contrasts.

Sound-to-Letter

The Sound-to-Letter task assesses two kinds of knowledge necessary for learning to read: (a) speech analysis at the level of the phoneme and (b) the ability to concretize phonemic awareness and apply it to an alphabetic code (F. R. Vellutino, personal communication, May 15, 2000). The Sound-to-Letter task is designed to measure a child's ability to segment spoken words into their constituent phonemes, as well as the use of that ability in the child's learning an alphabetic orthography. The task requires the child to provide the initial letter for a word presented orally. If a child cannot do so, he or she is asked to say the sound with which the word starts. If the child is unable to articulate the beginning phoneme, he or she is asked to give another word that begins with the same sound. The sequence of the Sound-to-Letter items follows a developmental progression from easy to more difficult; children are first asked to segment the beginning phoneme, then the final phoneme, and finally, the phoneme in the middle.[59] Previous research suggests that difficulty with this type of phonological coding is related to difficulty with alphabetic retrieval and could impair written word learning (e.g., Vellutino & Scanlon, 1987).

Linguistic complexity was also considered for the items in the Sound-to-Letter task. Since the Sound-to-Letter task gradually increases the difficulty of the items by varying the location of the phoneme, all items were limited to 3-phoneme words. The easiest set, the beginning phoneme set, contains 5 beginning continuants, 2 bilabials, and 3 stops (2 alveolar and 1 velar). The second set of items, the ending phoneme set, follows a similar scheme. The last and hardest set of items, the middle phoneme set, contains 4 tense ("long") and 6 lax ("short") vowel sounds in the middle. Each vowel has 1 tense and 1 lax exemplar. Again, care was taken to avoid phonologically ambiguous vowel sounds and to maintain phonological distance between contiguous items.

The two phonological awareness tasks were field-tested in three school divisions with 180 students

in spring of 2000. Pilot testing resulted in item changes (e.g., the removal of the word "food" from the Sound-to-Letter task) and instruction modifications designed to make the task clearer to students.

Feedback from the Field

In addition to the formal feedback solicited during the pilot studies, the PALS office continually seeks informal feedback from the field. During many screening windows, for example, the PALS office posts a survey on the PALS website (pals.virginia.edu) to seek feedback from teachers in the field. Response rates to questions posted on the surveys have ranged from 200 to 800 teachers. On one survey, teachers were asked to rate PALS tasks on (a) the ease of administration and scoring, (b) the clarity of directions, and (c) the information gained from screening. Open-ended comments were also invited. The results from the survey and qualitative comments from the field were consistent with comments received through the toll-free phone line, (888) UVA-PALS. That is, most teachers rated the PALS tasks good (4) to excellent (5) on a rating scale of 1 to 5.

On a second survey, teachers were asked to rate the impact of the PALS assessment on their teaching. Eighty-eight percent (577 out of 652, and 559 out of 638) of the classroom teachers who responded to questions about the "value added" by PALS reported that the assessment provided useful information and reliably identified students who needed extra help in reading. Seventy-five percent (479 out of 643) of the classroom teachers who responded to the question about the impact of PALS reported that the PALS assessment had a positive impact on their teaching. In addition, 2,011 teachers responded to a brief survey designed primarily to assess the usefulness of various PALS reports and website features. Between 71% and 80% of respondents rated class reports, class summary sheets, score history reports, and student summary reports as "very useful;" 2% or fewer of respondents rated any of these reports as "not useful."

Outside Review

The Code of Fair Testing Practices in Education (1988) defines the obligations of professionals who undertake the process of creating an assessment instrument. Included among these obligations are

Table 7 PALS Advisory Review Panel

Denise Pilgrim
Coordinator of Instruction
Charlottesville City, VA

Barbara Jackson
Elementary Principal
Appomattox, VA

Tisha Hayes
Assistant Professor, University of Virginia
Charlottesville, VA

Sandra Mitchell
Associate Superintendent for Instruction
Fauquier County, VA

Christine Gergely (Retired)
Reading Specialist
Hampton City, VA

Mary Maschal
Director of Elementary Education
Hanover County, VA

Linda Bland
Language Arts Supervisor
Harrisonburg City, VA

Laura Justice
Professor, Ohio State University
Columbus, Ohio

Jim Heywood (Retired)
Director, Office of Elementary Education
Virginia Department of Education

Table 8 External Reviewers

Dr. Nicholas Bankson
Professor of Communication Sciences & Disorders
James Madison University
Harrisonburg, Virginia

Dr. Susan Brady
Professor of Psychology
University of Rhode Island &
Haskins Laboratories
New Haven, Connecticut

Dr. Francine Johnston
Associate Professor of Reading
University of North Carolina-Greensboro

Dr. Frank Vellutino
Professor of Psychology &
Director, Child Research & Study Center
State University of New York at Albany

procedures that minimize the potential for bias or stereotyping. The potential for bias can be minimized if assessment tools are carefully evaluated.[60] Procedures that protect against inappropriate instrument content include the use of an advisory review panel and an external evaluation.

Advisory Review Panel

To evaluate the appropriateness of PALS content, we sought opinions about PALS from outside reviewers. Members of the advisory review panel and their affiliations are listed in Table 7. In addition, the Virginia Department of Education (VDOE) invited primary grade teachers, reading specialists, speech-language pathologists, instructional coordinators, special educators, and school administrators to serve on an advisory committee. Committee members were asked to review the content of the PALS 1–3 assessment, including student materials, the teacher's manual, and the directions for administration and scoring. The review committee was further asked to suggest changes or deletions of items and to provide feedback from their school or division. Suggestions about the PALS website were also solicited.

External Review

In addition to the opinions of the advisory review panel, the Virginia Department of Education (VDOE) sought the opinion of several external reviewers (listed in Table 8), all of whom were national experts in the fields of reading, communication sciences, or psychology. The first PALS technical manual and report[61] detailing the psychometric qualities of PALS and first-year results, as well as PALS materials and teacher's manuals, were sent to prominent researchers. Their charge was to determine the technical soundness of PALS as a valid and reliable instrument for the EIRI. Their opinions were presented to VDOE in March 1999. The judgments of these reviewers were favorable; copies of the reviews can be obtained from the Virginia Department of Education. An additional, independent review of PALS can be found in *Early Reading Assessment* (Rathvon, 2004, pp. 250–261).

Section IV

Establishing Summed Score Criteria and Benchmarks

In the following sections, we describe the process through which benchmarks were established for Entry Level tasks (Word Recognition and Spelling).

Decisions regarding PALS benchmarks were theoretically and empirically driven, and have been informed by data from several sources:

- Seventeen years of research with struggling readers in the Commonwealth of Virginia;
- statewide PALS data from successive cohorts of Virginia's EIRI;
- data gathered from pilot and field tests conducted with approximately 8,000 first-, second-, and third-grade students in the Commonwealth of Virginia;
- data gathered from pilot and field tests conducted with approximately 9,000 students in grades four through eight in the Commonwealth of Virginia.

Benchmarks reflect raw scores for each PALS task, based on the available data sources. The sum of these benchmark scores for the Entry Level tasks equals the summed score criterion for each grade. These benchmarks are reevaluated based on analyses of each year's statewide PALS results and data from ongoing pilot studies and field tests.

In November 2002 we conducted a formal standard-setting procedure to verify PALS benchmarks for grades 1–3. Standard setting refers to the process used by instrument developers to help establish, or in this case to verify, benchmarks or levels of performance that reflect 'minimal competence.' In standard setting, expert judges evaluate each individual task or item and state whether they believe that the student who is minimally competent would respond correctly. In the case of PALS, we assembled panels of experts in reading from throughout the Commonwealth. One panel of 20 judges was invited for each grade level, K through three. Each panel of judges spent a full day in Charlottesville evaluating individual entry level task items from all PALS materials.

We evaluated standard-setting judges' mean scores for PALS tasks against two sources of information: our current benchmarks, and statewide data from the most recent screening windows. In virtually all cases, standard-setting judges' scores were comparable to current benchmarks (i.e., within one standard deviation), and, moreover, fell at approximately the bottom quartile, which has traditionally been the approximate range of students identified for school intervention by PALS. For these reasons, we decided that standard-setting judges' evaluations supported PALS benchmarks, with the exception of one spelling list, which we describe in further detail in the Spelling section of this Technical Reference.

Word Recognition

Benchmarks for word recognition were determined using the construct of functional reading levels.[62] There are three functional reading levels: (a) the independent level, (b) the instructional level, and (c) the frustration level.

The construct of functional reading levels postulates that a student's *independent level* is the point at which he or she operates with few, if any, mistakes. Any errors that do exist are usually careless ones that are easily self-corrected. In reading or spelling word lists, this level corresponds to an error rate of 10% or fewer. In reading words in context, this level corresponds to an accuracy rate of 98% or greater. Since they do not require instructional guidance, students

with this level of competency can and should read independently.

A student's *instructional level* is the level at which he or she needs instructional guidance and can learn from teaching. At this level, students already have some background knowledge, but not enough to function independently without coaching. At the instructional level, students err about 25% of the time on tests of word recognition and spelling (Powell, 1971). In reading actual texts, their error rate does not exceed 10% of the words in running context. Research suggests that if students struggle to read more than 10% of the words in context, then they are unlikely to benefit from instruction using text at that level.[63]

The *frustration level* is reached when students miss 50% or more of the items on a test or list, or more than 10% of the words in running context. Reading at a frustration level is too laborious and flawed for the student to derive information, meaning, or enjoyment. Using the theoretical construct of instructional level, the benchmark for each graded word list was set at 15 words (about 75% accuracy) for word recognition in isolation, and at 90% for oral reading, or word recognition in context. Both benchmarks were lowered for preprimer readers at the beginning of first grade, since beginning first graders are just getting off the ground.[64] This construct was confirmed empirically in the Spring 2001 pilot; 97% to 100% of the students in grades 1–3 who read 15 or more words on the Word Recognition in Isolation task read the corresponding grade-level text with 90% or greater accuracy. In the Fall 2013 pilot in grades 4–8, 99% of the students who read 15 or more words on the Word Recognition task read the corresponding level text with 90% or greater accuracy.

Spelling

Benchmark scores and criteria for the Entry Level Spelling task were also theoretically and empirically determined for each grade level. First, we surveyed teachers, as well as reading researchers, teacher educators, and members of the advisory board, to establish a set of curricular assumptions for students at both the beginning and the end of the school year. Second, we reviewed existing research on trends and norms for the acquisition of specific phonic/spelling features across the grades. Finally, we reviewed the Virginia Standards of Learning (SOL) for English (grades one through eight) that related to phonics, spelling, and morphology. All of this information was condensed and developed into a rubric for scoring the presence or absence of specific phonics/spelling features, the total number of words spelled correctly, and a total score that was the sum of both.

Spelling Samples. Spelling samples from the K-3 pilot corpus (n = 2,405) and the 4–8 pilot corpus (n = 8,860) were scored for the presence or absence of specific phonic/spelling/morphological features, regardless of whether the whole word was spelled correctly. The total number of words spelled correctly was also recorded, as well as a composite score representing the sum of the feature score and the total number of words correct. These three variables were entered into the database as Feature Score, Total Correct Score, and Total Spelling Score. Next, teams of raters classified each spelling sample according to stages of developmental word knowledge.[65] These stages received categorical names, designating groups of students who shared apparent mastery of certain phonics/spelling/morphological features but "used but confused" others.[66] Disagreements were resolved and one spelling stage was established for each sample. Stages were assigned a code and entered into the database along with the Feature Score, Total Correct Score, and Total Spelling Score. The correlations between stage designations, as determined by qualitative featural analyses and by the numerical Total Spelling Score, were high and significant for each grade level (r = .84 to .95, $p < .01$).

Qualitative Benchmarks. Next, for each grade level, we settled on qualitative benchmarks based on developmental spelling research. For example, feedback from teachers and previous research on PALS confirmed that rising first graders who subsequently

learn to read without difficulty start out as Early Letter Name-Alphabetic Spellers and already represent beginning and ending sounds in their spelling. This expectation is in keeping with findings from the *Early Childhood Longitudinal Study: Kindergarten Class of 1998–99*,[67] which reports that,

as they are finishing kindergarten, nearly all the first-time kindergartners are recognizing their letters (94%), and nearly three out of four children (72%) understand the letter-sound relationship at the beginning and about half (52%) understand the letter-sound relationship at the ending of word (p. 12).

Similarly, research has shown that most upcoming second graders have already learned beginning and ending consonant sounds, short vowels, and a good many high-frequency consonant digraphs and blends. Developmental spelling research refers to such students as Late Letter Name-Alphabetic Spellers (Bear et al., 2004). Teachers also concur that advancing third graders have mastered most of the basic correspondences between single letters and sounds as well as letter patterns representing basic phonic elements such as short vowels and consonant blends. Students who read independently and well by the end of third grade begin that year as Early Within Word Spellers, who can read and spell long vowels and read silently at a second grade level. Teachers of students in the upper elementary and middle grades expect their minimally competent students to be able to represent common long-vowel patterns and other vowel sounds by grade 4; to double consonants or drop an e at appropriate syllable junctures by grade 5; to use and understand the meanings of common affixes by grade 6; and to use and understand spelling-meaning connections among derivationally related words by the middle grades. Using this research and theory, we determined general target points for students entering grades 1–8 that were in keeping with the Virginia SOLs, developmental spelling theory, and our data sources.

Quantitative Benchmarks. Next we conducted a series of investigations using the Statistical Package for the Social Sciences (SPSS) "explore" and "frequencies" options.[68] We examined the distributions of scores for each developmental spelling stage within each grade. We determined the exact range of scores associated with each spelling stage and the means, median, and mode for each phonic/spelling feature and total Feature Score. We looked at the range of Total Correct and Total Spelling Scores. We analyzed the pilot database by grade level, by stages within and across grade levels, and by quartiles. We also contrasted various subgroups of students from our validity and reliability analyses: students who

- could read grade-level text versus those who could not;
- could read at least 15 words on their grade-level word list versus those who could not;
- scored at the bottom quartile versus those who did not.

After establishing the range of scores associated with successful reading at the end of each grade, we looked for the intersection of theory and fact. That is, quantitative benchmarks and criterion-referenced scores were selected that validated theoretical and expert-teacher expectations for reading and spelling at each grade level.

Adjustment. As mentioned previously, the standard setting process in November 2002 prompted one change in spelling benchmarks for grades 1 and 2. Standard-setting judges who evaluated the spring first-grade spelling list and judges who evaluated the fall second-grade spelling list (two separate panels of judges working on different days to evaluate the same spelling list) agreed that minimally competent students would score higher on this spelling list than current PALS benchmarks. This finding prompted us to re-examine statewide data, and to conduct our own word-by-word and feature-by-feature review of these spelling lists with attention to the developmental spelling literature.

Based on these reviews, we adjusted the benchmarks for spring first-grade spelling and fall second-grade spelling from 18 to 20. This consequently raised the Entry Level Summed Score criteria for these

screening windows by two points as well (from 33 to 35). The new benchmarks and summed score criteria resulted in more consistent and stable identification rates in both first and second grade.

Summed Score Benchmarks

The sum of the scores for Word Recognition and Spelling equals the Entry Level Summed Score for each grade level. In addition, the Letter Sounds task is included in the Entry Level Summed Score in the fall of the first grade.

Benchmarks and the summed score criterion for Level B (Alphabetics) are the same as they have been for previous cohorts of Virginia's EIRI for the Alphabet Recognition and Letter Sounds tasks.[69] Various sources of information were consulted, including consensus opinions of primary school teachers on the advisory board, the National Center for Educational Statistics (NCES), and the means and standard deviations of students not in the bottom quartile in our statewide samples. That is, to establish the benchmark for Alphabet Recognition and Letter Sounds, we took the mean score on these tasks for all students scoring above the first quartile and subtracted a standard deviation from that mean. Then, new benchmarks were determined for the Concept-of-Word task and included in the summed score for Level B. Benchmarks for the Concept-of-Word task were determined in the same way that those for Spelling were derived: by examining distributions and correlations with other core variables in pilot samples, by previous research in literacy acquisition, and through the expert consensus of the PALS advisory board.

Benchmarks for Level C tasks were extrapolated from

- scores generated from previous statewide screenings in grades one, two, and three;
- data generated in the pilot samples;
- the extensive research base on developmental expectations for phonemic awareness (e.g., Smith et al., 1995).

Since the only students who performed the Blending and Sound-to-Letter tasks were those identified through the EIRI as needing additional instruction, this sample was positively skewed. Nevertheless, we examined the means and standard deviations, the median, the mode, and other measures of central tendency for each task.

Benchmarks and Discriminant Analysis (DA)

To verify PALS benchmarks statistically, we subject statewide and pilot data annually to discriminant analyses (DA). DA helps us assess the extent to which PALS variables reliably discriminate between groups of students who are or are not identified as needing additional services based on their PALS Entry Level Summed Score. The primary goal of DA is to isolate statistically the dimensions on which groups differ based on a set of variables (i.e., PALS subtask scores).

Since the inception of PALS, discriminant function analyses based on the PALS subtasks included in the Entry Level Summed Score have classified 93% to 99% of students correctly as Identified or Not-identified. This suggests that the combination of Word Recognition and Spelling scores (and, in fall of first grade, Letter Sounds scores as well) produces a discriminant function (a linear combination of these variables) that classifies students as Identified or Not-identified, using mathematical measures to isolate the dimensions that distinguish the groups. The abstract (or mathematical) classifications have consistently demonstrated a very high correspondence to PALS classification. PALS also has an Area Under the Curve (AUC) statistic, an indicator of overall diagnostic accuracy from a Receiver Operating Characteristic (ROC) curve of .92 indicating high classification accuracy in identifying students as "at risk for reading difficulty" and "not at risk for reading difficulty." All of these analyses provide evidence of the validity of PALS as an early reading assessment that reliably identifies students in need of additional instruction.

Section V

Technical Adequacy

In this chapter, we provide an overview of the demographic characteristics of students who have made up the PALS pilots and statewide samples, and then describe the technical adequacy of PALS Plus in terms of validity and reliability.

Standards for test construction, evaluation, and documentation, as outlined in the Standards for Educational and Psychological Testing[70] were carefully followed throughout the development of PALS. Special efforts were made to satisfy all the major criteria for acquiring and reporting technical data (cf. Invernizzi, Landrum, Howell, & Warley, 2005). In addition, we have attended carefully to the assessment criteria spelled out in various policy initiatives (e.g., Reading First, No Child Left Behind, Race to the Top, etc.). Specifically, Reading First guidelines suggest that assessment tools must serve four assessment purposes: (a) screening, (b) diagnosis, (c) progress monitoring, and (d) outcome evaluation. Moreover, states are encouraged to use assessments that target five core reading areas: (a) phonemic awareness, (b) phonics, (c) fluency, (d) vocabulary, and (e) comprehension.

In general, PALS Plus provides an assessment tool that clearly meets screening and diagnostic assessment purposes and the mid-year assessment provides for the use of PALS Plus as a progress monitoring tool. Originally designed as a screening tool for identifying children who were behind in the acquisition of important literacy fundamentals, PALS was not designed to serve as an assessment of outcomes. The diagnostic aim of PALS is readily apparent in the leveled nature of the PALS tasks, in which students proceed to increasingly focused diagnostic tasks (Levels B and C) if they do not meet benchmarks at the broader levels. PALS' focus on the core reading areas identified by policy initiatives is evident in its direct and instructionally relevant assessment of these literacy fundamentals (displayed previously in the PALS Conceptual Framework in Table 2). It assesses these core areas by means of various tasks: Word Recognition in Isolation, Spelling, Letter Sounds, and Oral Reading in Context (including accuracy, fluency, rate, and comprehension), and at the more diagnostic levels using Alphabet Recognition, Concept of Word, Blending, and Sound-to-Letter.

Broad Representation of Students

The tasks, items, and benchmarks in PALS Plus are derived from almost two decades of research, during which we evaluated PALS scores from over 500,000 students in grades one, two, and three in schools that participated in Virginia's EIRI between Fall 1997 and 2006. The first nine cohorts of the EIRI provide nine statewide samples representing a diverse population.[71] Table 9 lists the total number of students screened with PALS 1–3 in the sixteenth cohort (school year 2012–2013) of Virginia's EIRI by gender, free or reduced price lunch (FRPL), race/ethnicity, and grade level.

In our pilot and field tests, we work to ensure that pilot samples approximate statewide school enrollments in terms of gender, race/ethnicity, and socioeconomic status (SES). Table 10 summarizes the demographics of two pilot samples. For each demographic category, the percentage of the total pilot sample is compared to the percentage in the total statewide enrollment. With the possible exception of including slightly more students from higher poverty areas, the pilot samples generally mirrored the demographics of statewide enrollment.

Table 9 Demographics of Virginia's 17th Cohort Screened With PALS 1–3 (School Year 2013–14)

Demographic Category		Grade 1	Grade 2	Grade 3	Totals
GENDER	Males	42,977 (51.3%)	37,155 (51.4%)	17,824 (52.3%)	97,956 (51.5%)
	Females	40,776 (48.7%)	35,062 (48.6%)	16,249 (47.7%)	92,087 (48.5%)
ECONOMIC STATUS	Eligible for FRPL	40,536 (48.4%)	36,277 (50.2%)	19,321 (56.7%)	96,134 (50.6%)
	Not eligible for FRPL	42,528 (50.8%)	35,339 (48.9%)	14,404 (42.3%)	92,271 (48.6%)
	Unknown economic status	689 (0.8%)	601 (0.8%)	348 (1.0%)	1,638 (0.9%)
RACE/ETHNICITY	Black	20,775 (24.8%)	18,641 (25.8%)	10,521 (30.9%)	49,937 (26.3%)
	White	43,023 (51.4%)	36,935 (51.1%)	17,053 (50.0%)	97,011 (51.0%)
	Hispanic	11,683 (13.9%)	9,970 (13.8%)	4,164 (12.2%)	25,817 (13.6%)
	American Indian or Alaskan Native	347 (0.4%)	274 (0.4%)	98 (0.3%)	719 (0.4%)
	Asian	3,622 (4.3%)	2,814 (3.9%)	837 (2.5%)	7,273 (3.8%)
	Native Hawaiian or Other Pacific Islander	181 (0.2%)	130 (0.2%)	43 (0.1%)	354 (0.2%)
	Two or more races	4,122 (4.9%)	3,453 (4.8%)	1,357 (4.0%)	8,932 (4.7%)

FRPL = Free or reduced price lunch.

Table 10 Pilot Sample Demographics Compared to Statewide Enrollment: Spring 2004 (n = 6,392) and Spring 2005 (n = 452)

Demographic Category		Spring 2004 Pilot	2003–04 Statewide Enrollment	Spring 2005 Pilot	2004–05 Statewide Enrollment
GENDER	Males	50.9%	51.5%	44.0%	51.5%
	Females	49.1%	48.5%	56.0%	48.5%
SES	Low FRPL	23.3%	30.8%	19.9%	31.6%
	Med-Low FRPL	23.6%	25.6%	27.2%	25.2%
	Med-High FRPL	27.2%	22.5%	26.1%	22.7%
	High FRPL	25.9%	20.5%	26.8%	20.4%
RACE/ETHNICITY	Black	26.6%	26.9%	28.2%	26.7%
	White	65.8%	60.4%	63.3%	59.7%
	Hispanic	3.7%	6.5%	4.4%	7.0%
	Asian/Pacific Islander	2.1%	4.6%	2.4%	4.8%
	American Indian/ Alaska Native	0.6%	0.5%	0.9%	0.3%
	Ethnicity Not Listed	1.2%	1.1%	0.7%	1.5%

FRPL = Free or reduced price lunch.

Pilot and Field Testing for Grades 1–3

Data on the development, refinement, and technical adequacy of PALS 1–3 items and scoring procedures were obtained from an initial pilot conducted in Spring 2000, from large-scale pilots conducted in Spring 2001, Fall 2001, Spring 2003, Spring 2004, and Spring 2005, and from statewide data collected on first- through third-graders since the fall of 2000. Taken together, complete pilot samples include data from 13,021 students in grades one through three, while PALS 1–3 statewide samples included approximately 140,000 to 160,000 students' scores each year since 2000–01. A summary of the participants in grades 1–3 pilot studies appears in Table 11.

Two points are important to reiterate regarding these PALS scores and changes to the EIRI since its inception. First, PALS 1–3 refers to the version of PALS developed and first used in Fall 2000 in response to the expansion of the EIRI from a K–1 to a K–3 initiative. Data documenting the technical adequacy of PALS presented in this report are drawn from statewide and pilot samples both before and after this expansion. The time frame for each data collection effort is indicated in the title of each table.

Second, beginning with the Spring 2002 screening, participating divisions were required to screen all students in the spring to identify those who would receive intervention through the EIRI during the following school year. Prior to Spring 2002, fall had been the mandatory screening period. This switch, made in large part to assist schools in beginning each school year with students already identified for EIRI funded services, means that from Spring 2002 forward, the spring screening window represents the most comprehensive data set for students who have been administered PALS.

Table 11 Pilot Participation Summary: Spring 2000 through Spring 2005

	Grade	# Schools	# Teachers	# Students
S 2000	1	5	15	214
	2	5	15	187
	3	5	14	184
	Totals	*	**44**	**585**
S 2001	1	38	45	802
	2	30	32	609
	3	32	39	706
	Totals	*	**116**	**2,117**
F 2001	1	48	63	992
	2	33	42	609
	3	30	41	536
	Totals	*	**146**	**2,137**
F 2002	1	22	38	185
	2	20	34	165
	3	16	21	104
	Totals	*	**93**	**454**
S 2003	1	23	31	274
	2	32	41	336
	3	13	15	184
	Totals	*	**87**	**794**
S 2004	1	51	249	3,368
	2	51	223	3,024
	Totals		**472**	**6,392**
S 2005	1	73	196	200
	2	88	243	248
	Totals	*	**439**	**448**
Grand Totals				
		*	**1,397**	**12,927**

**Totals are not provided for number of schools because many teachers from different grade levels at the same schools participated.*

Pilot and Field Testing for Grades 4–8

The extension of PALS to grades 4–8 was accomplished in an iterative process of pilot testing between Fall 2012 and Fall 2013 at four different points in time: Fall 2012, Mid-Year 2012, Spring 2013, and Fall 2013. Pilot participants included students in grades 4–8 from 74 to 279 different schools, depending on the testing window. Pilot tests were administered by 246 to 807 classroom teachers, again depending on the testing window.

The total number of student participants ranged from 4,175 in Fall 2013 to 9,565 in Fall 2012. Table 12 shows the sample sizes for each of the four pilots with grades 4–8. We based our final results on the Fall 2013 sample. Table 13 shows the breakdown of students in the pilot sample by gender and race/ethnicity.

Summary Statistics

Students screened for Virginia's EIRI with PALS 1–3 are identified as in need of additional services based on their Entry Level Summed Score, which is the sum of two subtask scores: Word Recognition and Spelling. In the case of first-grade fall assessment only, the Entry Level Summed Score is the sum of three subtask scores: Word Recognition, Spelling, and Letter Sounds. Table 14 reports the number and percent of students identified as in need of additional instruction for recent statewide samples. Table 15 summarizes descriptive data for the Entry Level Summed Score for grades one, two, and three for statewide samples from 2011, 2012, and 2013. Note that relatively few third-graders have spring scores, because spring screening is optional for that grade.

Table 12 Pilot Sample Sizes for Grades 4–8

	Fall 2012 Pilot	Mid-Year 2013 Pilot	Spring 2013 Pilot	Fall 2013 Pilot
Schools	279	238	231	74
Teachers	807	674	625	246
Grade 4	4,250	2,657	2,571	2,116
Grade 5	3,634	2,070	1,777	1,839
Grade 6	685	338	306	118
Grade 7	531	244	190	56
Grade 8	465	182	143	46
Total Students	9,565	5,491	4,987	4,175

Table 13 Pilot Sample for Grades 4–8 by Gender and Race/Ethnicity

Demographic Category		Fall 2012 Pilot	Mid-Year 2013 Pilot	Spring 2013 Pilot	Fall 2013 Pilot
Gender	Males	51.1	52.2	51.6	50.5
	Females	48.9	47.8	48.4	49.5
Race/Ethnicity	Black	36.7	36.3	34.9	29.4
	White	50.8	51.1	52.1	57.3
	Hispanic	7.4	8.0	8.3	8.4
	American Indian or Alaskan Native	0.7	0.7	0.4	0.3
	Asian	1.9	1.6	1.9	2.0
	Two or more races	2.5	2.3	2.4	2.7

FRPL = Free or reduced price lunch.

Table 14 Number and Percent of Students Identified: Spring 2012 Through Spring 2014

	Grade	Screened	Identified
2012	1	80,746	10,020 (12.4%)
	2	71,449	8,659 (12.1%)
	3	16,603	3,783 (22.8%)
	Total	**168,798**	**22,462**
2013	1	82,220	12,243 (14.9%)
	2	72,436	11,956 (16.5%)
	3	19,249	5,059 (26.3%)
	Total	**173,905**	**29,258**
2014	1	83,753	12,860 (15.4%)
	2	72,217	11,946 (16.5%)
	3	18,879	5,356 (28.4%)
	Total	**174,849**	**30,162**

Table 15 Mean (sd) Entry Level Summed Score by Identification (ID) Status: Spring 2012 Through Spring 2014

	Grade	ID	Not ID
2012	1	23.93 (8.50)	55.73 (9.98)
	2	38.99 (12.52)	70.64 (6.00)
	3	47.00 (15.36)	77.81 (5.54)
2013	1	23.94 (8.42)	53.14 (9.66)
	2	40.70 (11.81)	68.65 (6.44)
	3	47.64 (14.04)	78.02 (5.69)
2014	1	23.60 (8.60)	54.10 (9.60)
	2	39.30 (12.46)	68.70 (6.33)
	3	47.77 (14.67)	77.08 (5.75)

Spring identification rates for third- graders appear higher because this sample is skewed toward a much higher percentage of previously identified students. In Table 15, the discrepancy between means for Identified and Not-identified groups highlights the clear distinction between these groups.

We examine and summarize PALS 1–3 scores from the Entry Level and Level A, B, and C tasks each year for indices of central tendency, internal consistency, and item reliability. We also conduct factor analyses and discriminant function analyses to assess the validity of PALS tasks. The following sections contain a brief description of the technical adequacy of PALS 1–3 in terms of reliability (the consistency of scores) and validity (the extent to which PALS 1–3 is supported as a true measure of the construct of reading).

Reliability

Reliability coefficients provide information about the consistency with which a test (or subtest) measures a given construct. Reliability may be assessed by comparing the scores of individuals taking the same test on different occasions (test-retest reliability), taking equivalent forms of the test (equivalent forms reliability), or, when it is not practical to assess individuals on two separate occasions, by examining the internal consistency of the scale (e.g., split-half reliability). Reliability evaluates the error of measurement or the "true score" variance. We assess two aspects of PALS' reliability: internal consistency (subtask reliability), and the consistency and accuracy of scoring (inter-rater reliability). Internal consistency was assessed using Cronbach's alpha, an index of internal consistency based on the average correlation of subtasks within a screening instrument;[72] these are reported in the following sections. Inter-rater reliability was assessed by having tasks scored and tabulated by multiple raters.

Subtask Reliability

Reliabilities for PALS subtasks were determined for grade, gender, SES, and ethnicity using data generated from statewide samples for the years 1998–99 and 1999–2000 (during which time the previous version of PALS was used for grades K and 1). Task reliabilities were determined using Cronbach's alpha; Table 16 displays the alpha coefficients for the Summed Score tasks for the first-grade sample by gender, SES, and race/ethnicity, based on state-

wide samples from the fall and spring of 1998–99 and 1999–2000. Alpha coefficients are acceptable across the two-year period, ranging from .66 to .88, with a mean alpha coefficient of .80 and a median coefficient of .81 for all segments of the sample. The consistency of the coefficients for all demographic subgroups indicates that the Summed Score tasks for PALS were stable and reliable across a broad representation of students.

Expansion of the EIRI from a K–1 initiative to a K–3 initiative in Fall 2000 required changes in the structure of PALS 1–3 and demanded that reliabilities be examined differently. Beginning in Fall 2000, PALS 1–3 tasks were structured in such a way that students only proceeded beyond the two Entry Level tasks (Spelling and Word Recognition) if they failed to meet the benchmark on the Entry Level Summed Score.

Now there are only two scores included in the Entry Level Summed Score. The range of scores for one of these tasks (Word Recognition) is highly restricted by its very nature. Because the purpose of the Word Recognition task is to screen out students at a minimal competence level, there is a ceiling effect, with most students' scores clustered at the top of this distribution. Students not scoring near the top of the scale generally are those identified for additional instruction. Thus a more reasonable estimate of the reliability for these tasks is computed using Cronbach's alpha separately on each of the two individual scales—Word Recognition and Spelling. Reliability estimates on the Word Recognition task in pilot data from the Spring 2000, Spring 2001, Fall 2001, and Spring 2004 are presented in Table 17. Table 18 presents the reliability measures (coefficient alpha) for the words lists by form, for grades 4–8.

Table 16 Summed Score Task Reliability (Cronbach's Alpha) Across Demographic Categories: First-Grade, Entire Sample, 1998–2000

	Entire Sample	Female	Male	SES 4	SES 3	SES 2	SES 1
Fall 1998	.83	.83	.84	.82	*	.85	.84
Spring 1999	.82	.79	.83	*	.83	.74	.84
Fall 1999	.78	.77	.78	.76	.76	.80	.75
Spring 2000	.76	.72	.79	.75	.81	.76	.73

	Region I	Region II	Region III	Region IV	Region V	Region VI	Region VII	Region VIII
Fall 1998	.83	.83	.79	*	.82	.78	.86	.82
Spring 1999	.85	.76	.81	*	*	.82	.84	.84
Fall 1999	.75	.75	.77	*	*	.77	*	*
Spring 2000	.66	.79	.84	.83	.87	.73	.76	.67

	African American	Asian & Pacific Islander	Caucasian	Hispanic	Native American	Other
Fall 1998	.82	*	.83	.84	*	.82
Spring 1999	.82	*	.80	*	*	.85
Fall 1999	.77	*	.76	*	*	
Spring 2000	.76	.87	.73	.82	*	.88

** = too few cases to compute Cronbach's alpha. SES based on quartiles of free lunch at the school level. SES 1 > 55% free lunch; SES 2 = 36–55% free lunch; SES 3 = 18–35% free lunch; SES 4 = 0–17% free lunch.*

Table 17 Reliability Coefficients for Word Recognition in Isolation Task for Pilot Samples

	Cronbach's alpha (*n*)			
Word List	**Spring 2000**	**Spring 2001**	**Fall 2001**	**Spring 2004**
Preprimer	n/a	.96 (*n* = 486)	.92 (*n* = 617)	.83 (*n* = 315)
Primer	.91 (*n* = 77)	.94 (*n* = 25)	.91 (*n* = 369)	.86 (*n* = 699)
Grade 1	.93 (*n* = 224)	.90 (*n* = 54)	.88 (*n* = 409)	.79 (*n* = 1,188)
Grade 2	.91 (*n* = 223)	.87 (*n* = 93)	.91 (*n* = 223)	.86 (*n* = 1,674)
Grade 3	.87 (*n* = 222)	.81 (*n* = 109)	.86 (*n* = 295)	.86 (*n* = 1,747)
Grade 4	—	—	—	.88 (*n* = 1,379)
Grade 5	—	—	—	.83 (*n* = 513)
Grade 6	—	—	—	.87 (*n* = 190)

Table 18 Reliability Measures (coefficient alpha) for Word Recognition in Isolation Task by Form: Grades 4–8

	Form A		Form B	
Word List	**alpha**	***n***	**alpha**	***n***
Grade 4	.91	542	.90	1,800
Grade 5	.83	442	.84	1,527
Grade 6	.84	371	.86	1,311
Grade 7	.85	262	.79	905
Grade 8	.79	131	.76	545

To assess the reliability of the Word Recognition task further, a set of individual item scores is randomly collected from the statewide sample. In Fall 2001, over 4,000 such scores were collected for the preprimer, grade one, and grade two word lists, since these three word list scores form part of the Entry Level Summed Score for first, second, and third grade, respectively. Table 19 presents Cronbach's alpha for this subsample.

Reliability coefficients for the Spelling task were also consistently high across pilot and statewide samples. Table 20 presents Cronbach's alpha computed for first-, second-, and third-grade spelling lists for the Spring 2001, Fall 2001, Spring 2003, and Spring 2004 pilot samples. Table 21 presents similar reliability coefficients for the spelling task administered in grades 4–8.

Inter-rater Reliability

Inter-rater reliability coefficients provide evidence that individuals score a particular task in the same way. To determine the inter-rater reliability of PALS 1–3, scores for various PALS 1–3 tasks from two different raters (or scorers) have been compared. These inter-rater reliabilities are summarized in Table 22. Early estimates of inter-rater reliability of various PALS tasks, based on pilot samples prior to the leveling of PALS 1–3 in Fall 2000, are also included in Table 22. These include Fall 1997 and Spring 1999 estimates for Spelling, Alphabet Recognition, and Letter Sounds, each of which were .98 or .99.

Table 19 Reliability Coefficients for Word Recognition in Isolation Task for Statewide Subsample, Fall 2001

	Cronbach's alpha (*n*)	
Word List	***n***	**alpha**
Preprimer	4,668	.93
Grade 1	4,541	.92
Grade 2	4,387	.93

Table 20 Reliability Coefficients for Spelling Task for Pilot Samples

	Cronbach's alpha (*n*)				
Spelling List	**Spring 2001**	**Fall 2001**	**Spring 2003**	**Spring 2004**	**Spring 2005**
Grade 1	.86 (*n* = 324)	.86 (*n* = 463)	.93 (*n* = 1,401)	.92 (*n* = 1,485)	—
Grade 2	.92 (*n* = 302)	.89 (*n* = 286)	.94 (*n* = 1,122)	.92 (*n* = 1,404)	—
Grade 3	.92 (*n* = 267)	.92 (*n* = 269)	.89 (*n* = 455)	—	—
Additional Spelling Words (syllable juncture, affixes)	—	—	—	—	.88 (*n* = 60)

Table 21 Reliability Measures (coefficient alpha) for Spelling Task by Form: Grades 4–8

		Form A		Form B	
		alpha	***n***	**alpha**	***n***
SAMPLE	Elementary School	.97	596	.92	1,905
	Middle School	.95	40	.91	129
GENDER	Male	.97	316	.96	967
	Female	.96	331	.96	978
RACE/ ETHNICITY	White	.97	304	.96	1,040
	Non-White	.96	343	.96	905

In Fall 2000, inter-rater reliability was calculated on the scores for 478 students in five schools by having independent raters score PALS tasks simultaneously with teachers administering them. For the Word Recognition in Isolation, Oral Reading in Context, and Spelling tasks, the classroom teacher administered and scored the appropriate sections of PALS 1–3, following the same directions provided in the 2000–01 Teacher's Manual. In each case another rater, trained in the administration and scoring of PALS 1–3, rated the student's performance alongside the classroom teacher. For the Blending and Sound-to-Letter tasks, two raters, trained by the PALS office in the administration and scoring of those two tasks, participated. One rater administered the task and scored the student's performance, while

Table 22 Inter-rater Reliabilities Expressed as Pearson Correlation Coefficients for PALS Tasks

PALS Task	Date	Correlation (n)
Entry-Level Tasks		
Word Recognition in Isolation	Fall 2000	Preprimer: .99 ($n = 51$) Primer: .99 ($n = 52$) Grade 1: .98 ($n = 45$) Grade 2: .98 ($n = 63$) Grade 3: .98 ($n = 46$)
Spelling	Fall 1997	K & 1: .99 ($n = 130$)
	Spring 1999	K & 1: .99 ($n = 154$)
	Fall 2000	Total: .99 ($n = 214$)
	Fall 2001	Grade 1: .99 ($n = 375$) Grade 2: .99 ($n = 276$) Grade 3: .99 ($n = 257$)
Leval A Tasks		
Oral Reading in Context	Fall 2000	Primer: .94 ($n = 36$) Grade 1: .97 ($n = 43$) Grade 2: .96 ($n = 50$) Grade 3: .98 ($n = 72$)
	Fall 2002	Readiness: .74 ($n = 33$) Preprimer A: .77 ($n = 32$) Preprimer B: .63 ($n = 29$) Preprimer C: .83 ($n = 29$) Primer: .97 ($n = 18$) Grade 1: .97 ($n = 21$) Grade 2: .85 ($n = 38$) Grade 3: .81 ($n = 78$)
Level B Tasks		
Alphabet Recognition	Fall 1997	K & 1: .99 ($n = 122$)
	Spring 1999	K & 1: .99 ($n = 154$)
Letter Sounds	Fall 1997	K & 1: .99 ($n = 121$)
	Spring 1999	K & 1: .98 ($n = 154$)
Concept of Word	Fall 2001	Total: .97 ($n = 110$)
Level C Tasks		
Blending	Fall 2000	Total: .97 ($n = 55$)
Sound-to-Letter	Fall 2000	Total: .94 ($n = 55$)

$p < .01$ for all correlations

the second rater watched and simultaneously scored the student's performance. In each setting, the two raters were instructed not to compare or alter their scoring based on that of the other rater. After testing was complete, the two scores were compared and inter-rater reliability was determined using Pearson correlation coefficients. Correlations ranged from .936 to .997 ($p < .01$). Table 22 lists the correlations for Word Recognition in Isolation, Oral Reading in Context, Spelling, Blending, and Sound-to-Letter tasks. These inter-rater reliability coefficients are high and significant, indicating that the tasks on PALS 1–3 can be scored accurately and reliably.

In Fall 2001, inter-rater reliability coefficients were also calculated for spelling lists in grades one, two, and three ($r = .99$ in each case) and for Concept of Word ($r = .97$). In Fall 2002, an examination of inter-rater reliability on Oral Reading in Context accuracy yielded coefficients ranging from .81 to .97 across all passages from the primer through third-grade level, and from .63 to .83 across preprimer levels; these should be interpreted with caution given the relatively small *n* associated with any one passage.

Table 23 Test-retest Reliabilities for Entry Level Tasks, Pilot Sample, Fall 2002

Grade	Entry Level Task	Correlation (n)
1	Letter Sounds	.90 ($n = 77$)
	Spelling	.92 ($n = 77$)
	Preprimer Word List	.89 ($n = 77$)
	Sum Score	.96 ($n = 77$)
2	Spelling	.89 ($n = 59$)
	1st Grade Word List	.88 ($n = 59$)
	Sum Score	.92 ($n = 59$)
3	Spelling	.95 ($n = 68$)
	2nd Grade Word List	.93 ($n = 68$)
	Sum Score	.97 ($n = 68$)

$p < .001$ in all cases.

In summary, inter-rater reliability estimates for PALS 1–3 Entry Level Tasks have been consistently high, ranging from .98 to .99. Inter-rater reliability coefficients for Level A tasks ranged from .81 to .97 for primer through third-grade passages and from .63 to .83 for preprimer passages. Inter-rater reliability coefficients for PALS 1–3 Level B and Level C tasks have also been high, ranging from .94 to .99.

Test-retest Reliability

To examine the stability of PALS scores, we assessed test-retest reliability in a small sample ($n = 204$) in Fall 2002. Participating teachers were asked to randomly select 5 students from their class rosters and administer PALS 1–3 tasks a second time at least one week, but no more than two weeks, after the initial screening was completed. These reliability estimates, expressed as Pearson correlation coefficients in Table 23, are all high and significant, suggesting that PALS 1–3 tasks are stable over a brief period of time.

Validity

In general terms, validity refers to the extent to which one can trust that a test measures what it is intended to measure. But a test is not said to be valid or not valid in isolation. Instead, a test must be assessed for evidence of validity in relation to the specific purpose for which it is used with a given population. Thus for PALS 1–3, three types of validity have been assessed through our examination of pilot and statewide PALS data over the past nine years. In the following sections, we describe evidence of PALS' (a) content validity, (b) construct validity, and (c) criterion-related validity, both predictive and concurrent. Finally, to provide further evidence of validity, we assess the differential item functioning of PALS tasks for different groupings of students.

Content Validity

Content validity is the degree to which the sample of items and tasks provides a relevant and representative sample of the content addressed.[73] The content addressed in PALS 1–3 is reading.

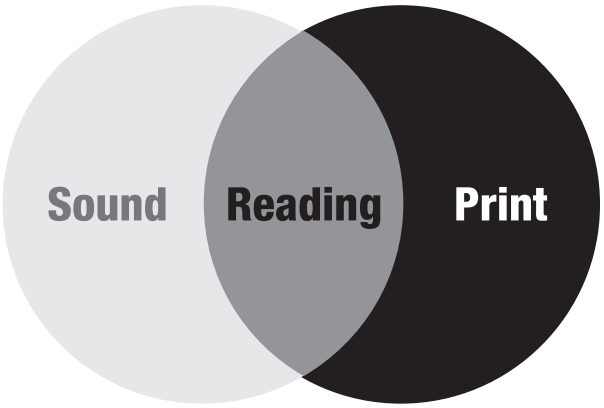
Reading is defined as fast and accurate recognition of written words such that there are cognitive resources left over to simultaneously group those words into meaningful grammatical units for understanding text. The National Reading Panel notes that a fluent reader is "one who can perform multiple tasks—such as word recognition and comprehension—at the same time."[74]

Researchers who study eye movements during reading have demonstrated that fluent readers are able to take in more information about words in a single fixation than are non-fluent readers.[75] Not only are they better at recognizing a word in a single fixation, but they also demonstrate fewer regressions back to look at the word again after having read on further. Word knowledge and practice allows fluent readers to recognize words automatically and to group them into meaningful phrases. As children's reading experiences widen and their knowledge of letter patterns expands, there is a gradual but continuous increase in word recognition and reading speed. Reading speed and fluency facilitate reading comprehension by freeing cognitive resources for interpretation.[76]

To ensure that PALS 1–3 has ample content validity, we took special care to select tasks shown by research to be essential to reading comprehension and to select words that are appropriate for each grade level being assessed. Entry Level tasks represent the fundamental orthographic word knowledge necessary for fluent reading in context.

The Level A task, Oral Reading in Context, provides opportunities for teachers to assess aspects of reading fluency and to determine an instructional reading level by calculating the proportion of words read accurately in the passage. Teachers are also provided a simple rubric for rating other aspects of oral reading fluency, such as reading rate and expression.[77] To ensure that students are not focusing solely on fluency at the expense of comprehension, questions are provided to probe their understanding.

Figure 1 PALS Theoretical Model



Level B Alphabetic tasks were chosen to provide a straightforward assessment of alphabet knowledge. To assess alphabet recognition, all 26 letters of the alphabet were included. To assess knowledge of letter sounds, all letters were included except *Q* and *X*, which are too difficult to pronounce in isolation. To assess the application and utility of the alphabetic code, a Concept-of-Word task was included in order to demonstrate a child's ability to use the alphabetic code to coordinate speech with printed word boundaries.

Level C tasks assess phonological awareness, the basic understanding that speech sounds can be segmented and clustered in variously sized units. The unit assessed in Level C is the phoneme. Phonemic awareness is the ability to pay attention to, identify, and manipulate the smallest units of speech-sounds, which correspond roughly to an alphabetic orthography. Researchers have assessed phoneme awareness in children by asking them to categorize spoken words by beginning sounds (e.g., *man* and *moon* go together because they both start with /*m*/), or by segmenting spoken words into individual phonemes (e.g., *man* = /*m*/ /*a*/ /*n*/), or by blending individual speech sounds to form a recognizable word (e.g., /*m*/ + /*a*/ + /*n*/ = *man*). What all these tasks have in common is the necessary focus on the underlying structure of the spoken word. This focus on speech sounds is needed to learn letter

sounds and to apply them to reading and writing. Level C of PALS 1–3 includes two measures of phoneme awareness: (1) phoneme blending and (2) segmenting sounds and matching them to letters. These two tasks assess two kinds of phonological awareness: (1) speech analysis at the phonemic level and (2) the transfer of phonemic awareness to letters. We took special care to use words previous researchers have determined to be in the core speaking vocabulary of first grade children.[78] We gave further consideration to the linguistic complexity of each sound.

Construct Validity

Construct validity refers to the degree to which the underlying traits of an assessment can be identified and the extent to which these traits reflect the theoretical model on which the assessment was based.[79] The theoretical model on which PALS was originally based is illustrated in Figure 1. It depicts the original purpose of PALS, which was designed to assess children's knowledge of speech sounds, knowledge of print, and ability to perform tasks that required the wedding of the two. The pronunciation of letter sounds, the ability to match letters and letter patterns to speech segments to produce a spelling, and the ability to recognize words in and out of context all require the application of knowledge of both sound and print.

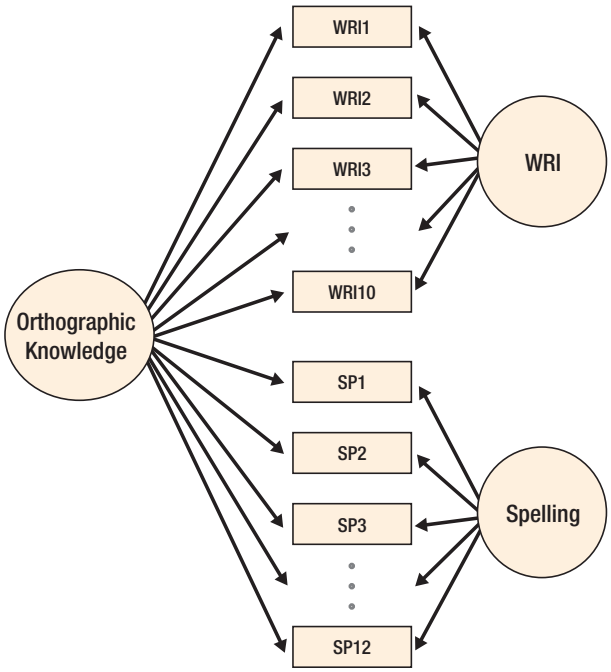
We tested the theoretical model illustrated in Figure 1 in two ways. First, we conducted principal components analyses (PCA) on PALS data to verify the underlying factor structure. Second, we conducted discriminant analyses (DA) on PALS data to determine the extent to which group membership (i.e., Identified versus Not-identified for additional instruction) could be predicted accurately from PALS subtask scores.

Principal Components Analysis (PCA). Factor analysis for the first statewide PALS sample yielded one factor with an eigenvalue of 5.20. The same unitary factor was also found using kindergarten data only (eigenvalue of 4.92) and first-grade data only (eigenvalue of 4.05). The one-factor solution suggested that PALS was measuring a unitary trait: reading, or the marriage between sound and print. In the first EIRI cohort, the single PALS factor accounted for 58% to 74% of the total variance in the children's scores on all the tasks in both the phonological awareness and literacy screening components of PALS for the entire sample, and separately for kindergarten and for first grade.[80]

A unitary factor was replicated using second- and third-year PALS results (1998–99; 1999–2000). Principal components analysis (PCA) consistently yielded a single factor with an eigenvalue greater than one for the entire sample and for each grade level. Factor loadings from the second and third year were similar to the first: five core variables (Rhyme, Sound, Alphabet Recognition, Letter Sounds, and Spelling) defined the construct. Factor loadings for Letter Sounds and Spelling were consistently large and accounted for most of the construct. This pattern stayed the same for the entire sample and for each grade level. Word Recognition, given only to first- graders, also loaded onto the single factor in first grade.[81]

In Fall 2000, PALS 1–3 was streamlined into a more efficient screening tool to accommodate the EIRI expansion to third grade. First grade tasks with the largest factor loadings (Letter Sounds, Spelling, and Word Recognition) were subsumed into Entry Level. Other tasks that discriminated between children who did and did not require additional instruction were moved to subsequent levels, which became increasingly more diagnostic in nature. Children were routed to each subsequent level based on grade-level criteria for minimal competence on the level before. Principal component analyses each year have consistently yielded a single factor for each level of tasks: Entry Level, Alphabetics (Level B), and Phonemic Awareness (Level C). For Entry Level tasks (Word Recognition in Isolation and Spelling), this unitary factor accounted for 79% to 85% of the variance in Entry Level Summed Scores for grades one through three in Spring 2001 and Fall 2001 statewide samples. In all cases, both Entry Level tasks

Figure 2 Bifactor Structure of PALS 1–3



(Word Recognition and Spelling) loaded heavily on this unitary factor (loadings ranging from .89 to .94). In the case of fall first grade, wherein the Entry Level Summed Score consists of three scores (Word Recognition, Spelling, and Letter Sounds), all three scores loaded heavily (.88 or greater in Fall 2001) on this single factor. We repeat these PCAs each year with statewide data, and have consistently replicated these single-factor solutions. The consistency of these PCA results and the factor structure that has emerged supports that PALS Entry Level is indeed associated with a unitary factor that is consistent with the construct of reading.

Factor Analysis. More recently, research[82] was conducted using both exploratory and confirmatory factor analyses (CFA) to investigate the factor structure of PALS 1–3. Item-level data from the Word Recognition in Isolation (WRI) and Spelling tasks were analyzed using a large sample (n = 14,993) of second-grade students from Virginia. Alternative factor models were tested (i.e., a one-factor model, a two-correlated factor model, and a bi-factor model) using an exploratory sample with item parcels. Results indicated that the bi-factor model (see Figure 2) best represented the data as evidenced by a comparison of model fit indices. Using a confirmatory, hold-out sample, CFA model fit indices were good as well (RMSEA = .02, CFI = .99, TLI = .99, SRMR = .02) providing additional evidence of the bi-factor model's generalizability. A general overarching Orthographic Knowledge factor accounted for a large proportion of the variability in the WRI and spelling tasks. As a measure of reliability, coefficient omega hierarchical[83] (ωh) was more than adequate, ωh =.88, and 88% of the variance in the PALS 1–3 overall summed score is attributable to a general factor of Orthographic Knowledge.

Discriminant Analyses (DA). The purpose of discriminant analysis is to determine whether test scores can discriminate accurately between groups of subjects if their group identity is removed. Because PALS is designed as a screening tool to identify students in need of additional reading instruction, we test this hypothesis each year on PALS data by determining the extent to which a combination of PALS subtest scores accurately predicts membership in Identified and Not-identified groups.

Since the inception of PALS 1–3 in Fall 2000, we have conducted discriminant analyses on statewide data during each screening window using the subtask scores that make up the Entry Level Summed Score—Word Recognition and Spelling. Letter Sounds is included in the Entry Level Summed Score in Fall of 1st grade only. These analyses have consistently yielded a function that is statistically significant (as indicated by a statistically significant Wilks' lambda for the discriminant function) in differentiating groups of students. The discriminant functions have also accurately classified between 94% and 98% of students as Identified or Not-identified at grades 1, 2 and 3 over the course of the past three years. Table 24 summarizes the discriminant analysis results for three school years.

Together, the results of our PCA and DA analyses indicate that PALS 1–3 assesses a single general construct associated with beginning reading, and

further, that the combination of variables making up the PALS subtasks discriminates reliably between groups of students who are or are not identified as needing additional reading instruction.

Intercorrelations among PALS Tasks. A third source of evidence for a test's construct validity may be found in the intercorrelations among its subtests. We examined the intercorrelations among PALS 1–3 task scores to assess the relationships among PALS tasks and, further, to verify that the pattern of intercorrelations is consistent across grade levels and among student subgroups (e.g., SES levels or ethnicity categories). High intercorrelations (above .80) are consistently obtained among PALS 1–3 Summed Scores in the fall of grades two and three and in the spring of the year before ($p < .001$). The correlation between spring kindergarten and fall first grade is medium-high (.60 to .79). For all three grades, the correlation between PALS Summed Scores at yearly intervals, from fall to fall, is medium-high and significant ($p < .001$).

At all grade levels, medium-high (.60 to .79) to high (above .80) intercorrelations are consistently obtained between the Spelling and Word Recognition tasks and the PALS 1–3 Entry Level Summed Score.

Table 24 Discriminant Analysis Results for Entry Level Tasks and Identification Status: Statewide Samples

	Grade	Wilk's lambda*	Students Classified Accurately
2008	1	0.39	96%
	2	0.32	96%
	3	0.30	98%
2009	1	0.40	96%
	2	0.35	97%
	3	0.31	95%
2010	1	0.39	96%
	2	0.32	98%
	3	0.33	94%

$p < .001$ in all cases.

In addition, high intercorrelations are consistently obtained between the Letter Sounds and Concept of Word tasks and the Level B (Alphabetics) Summed Score in all three grades. Letter Sounds is highly correlated with the Entry Level Summed Score in the fall of first grade, as is the Level B Summed Score. All of these intercorrelations are significant ($p < .001$).

Medium (.40 to .59) to medium-high (.60 to .79) correlations are consistently obtained between Level B (Alphabetics) and the Entry Level Summed Score and between Concept of Word and the Entry Level Summed Score at all three grade levels. Medium-high intercorrelations are obtained between Alphabet Recognition and the Level B Summed Score at all three grades and with the Entry Level Summed Score in grade one. The Blending and Sound-to-Letter tasks from Level C (Phonemic Awareness) are intercorrelated in the medium-high range with each other and with the Level B (Alphabetics) Summed Score in grades two and three, but only Sound-to-Letter is correlated with the Level B Summed Score in grade one.

Medium correlations (.40 to .59) are obtained between the Blending and Sound-to-Letter tasks and the Entry Level Summed Score in grade three, but only Sound-to-Letter is correlated with the Entry Level Summed Score at this magnitude for grades one and two. Only medium correlations are obtained for Sound-to-Letter and Blending in grade one. Further, while Letter Sounds is highly correlated with the Entry Level Summed Score in grade one, Letter Sounds is only moderately correlated with the Entry Level Summed Score in grades two and three. At these grade levels, only students not meeting the Entry Level Summed Score are administered Letter Sounds.

Medium correlations (.40 to .59) are also obtained in all three grades among the preprimer passage- reading accuracy scores, recognition of the grade-level word lists, and the Entry Level Summed Score. At the first and third-grade level, reading comprehension is moderately correlated with the Level B (Alphabetics) Summed Score and, for third

grade, with the Blending task as well. Reading accuracy scores for the readiness passage and the Level B Summed Score are moderately correlated in grade one, as are many of the preprimer reading accuracy scores across all three grades. All of these intercorrelations are significant ($p < .001$).

Low intercorrelations are consistently obtained for the Blending task and the Entry Level Summed Score and the Level B (Alphabetics) Summed Score for grade one. Nevertheless, these correlations are significant ($p < .001$).

Intercorrelations among PALS 1–3 tasks and between each task and the PALS 1–3 Entry Level Summed Score are also calculated for demographic categories, including gender, SES, and race/ethnicity. Correlational patterns are examined for consistency across demographic groups. For each demographic category, intercorrelations among tasks and between each task and the overall Summed Score are similar to the correlational pattern for the entire statewide sample. That is, high correlations (in the range of .80 to .99) are generally high for all demographic categories and for the entire sample; medium-high correlations (in the range of .60 to .79) are generally medium-high for all demographic categories and for the entire sample; medium correlations (in the range of .40 to .59) are generally medium for all demo- graphic categories and for the entire sample; and low correlations (in the range of .00 to .39) are generally low for all demographic categories and for the entire sample. This pattern of consistency suggests that the tasks on PALS 1–3 behave in a similar manner for all students, regardless of gender, SES, or race/ethnicity.

Criterion-related Validity

Criterion-related validity refers to the extent to which assessment scores are related to one or more outcome criteria.[84] There are two types of criterion-related validity: predictive, where an assessment is used to predict future performance; and concurrent, where assessment results are compared to a different criterion assessed at approximately the same time. Both forms of validity have been assessed for PALS 1–3. During the 2000–01 school year, PALS 1–3 scores from the fall screening were compared to a number of criteria assessed later in the spring (predictive validity). PALS 1–3 scores obtained during Spring 2001 were also compared to a number of other measures also obtained that spring (concurrent validity). A summary of the predictive and concurrent validity studies conducted on PALS 1–3 is shown in Table 25.

Predictive validity is a form of criterion-related validity in which one assessment is used to predict future performance on another assessment conducted later in time. The predictive validity for PALS 1–3 was examined during the 2000–01 school year by testing the hypothesis that higher Entry Level Summed Scores on the fall administration of PALS 1–3 would be associated with higher scores on another reading test administered to the same students at the end of the school year. This hypothesis was tested with two different outcome criteria: (a) the Stanford Achievement Test,[85] and (b) the third-grade Virginia Standards of Learning (SOL) reading test, both administered in Spring 2001. These two were selected as outcome measures because they are both required by the Virginia Department of Education in alternate grades, beginning in grade three. We assessed predictive validity by examining correlation coefficients between PALS scores and Stanford-9 and SOL reading test scores, and further by conducting regression analyses. The resulting R^2 provides an index of the amount of variability in the outcome measure (i.e., Stanford-9 or SOL scores) that can be predicted based on its relationship to the predictors (PALS Entry Level task scores).

Because the Stanford-9 is a norm-referenced test while PALS 1–3 is an informal, criterion-referenced assessment, high correlation coefficients were not expected. Although the SOL reading component is a criterion-referenced test, it primarily measures reading comprehension, while the Entry Level of PALS

Table 25 Overview of Criterion-related Validity Studies

Validity Study	Date	Grade	*n*
Predictive			
PALS 1–3 Entry Level Summed Score (fall) with Stanford-9 Total Reading scaled score (spring)	Fall 2000 Spring 2001	1	739
		2	766
PALS 1–3 Entry Level Summed Score (fall) with Standards of Learning (SOL) Reading (spring)	Fall 2000 Spring 2001	3	277
Concurrent			
PALS 1–3 passage accuracy and Qualitative Reading Inventory (QRI-I) passage accuracy	Spring 2001	1	65
PALS 1–3 Entry Level Summed Score and Developmental Reading Assessment (DRA) Spring 2001 instructional reading level	Spring 2001	1	104
		2	61
		3	32
PALS 1–3 passages and California Achievement Test (CAT/5)	Spring 2001	1	195
PALS 1–3 Entry Level Summed Score with Stanford-9 Total Reading scaled score	Spring 2001	1	174
		2	50
PALS 1–3 Entry Level Summed Score with Standards of Learning (SOL) Reading	Spring 2001	3	283

1–3 primarily measures orthographic competence. Nevertheless, we expected PALS 1–3 Entry Level Summed Scores to explain a significant amount of variance in the end-of-year SOL reading assessment.

PALS 1–3 and Stanford-9. To assess the predictive validity of PALS 1–3 relative to the Stanford-9, Fall 2000 PALS scores for 739 first-graders and 766 second-graders were compared to Stanford-9 scores collected at the end of the school year (Spring 2001). Bivariate correlations between the fall PALS Entry Level Summed Score and the spring Stanford-9 Total Reading scaled score were medium-high and significant for both first- and second-grade samples (r = .73 and .63, $p < .01$). In regression equations, the adjusted R^2 was .53 for first grade and .34 for second grade, indicating that in a conservative model corrected for estimated shrinkage upon replication, 53% and 34% of the variance in Stanford-9 total scale scores in the spring (for first- and second-graders, respectively) could be predicted based on their relationship to PALS Entry Level Summed Scores from the previous fall.

First and second grade PALS Entry Level Summed Scores from Fall 2000 also predicted a significant amount of variance in subtests of the end-of-year Stanford-9 (scaled scores for Word Study Skills, Vocabulary, and Reading Comprehension) for both grades ($p < .001$). For example, the adjusted R^2 for Stanford-9 Reading Comprehension was .50 for first grade and .25 for second grade. The relationship between second-grade spring Stanford-9 Reading Comprehension scale scores and fall PALS 1–3 results suggests that the fall PALS screening is statistically significant in predicting end-of-year reading achievement and explains approximately one-half of the variance of Stanford-9 Reading Comprehension scores obtained at the end of first grade and about one-quarter of the variance of Stanford-9 Reading Comprehension scores obtained at the end of second grade.

PALS 1–3 and SOL. For third grade, the predictive validity of PALS 1–3 was assessed by examining Standards of Learning (SOL) reading test scores for 34,750 third-graders who were screened with PALS at the beginning of the year in Fall 2011. Correlation and regression analyses were conducted to test the hypothesis that greater Entry Level Summed Scores for PALS in the fall, would be associated with greater SOL scores for reading in the spring. The bivariate correlation between the fall PALS Entry Level Summed Score and the spring SOL Total Reading score was .50 ($p < .001$). A regression equation using Fall 2000 PALS Entry Level Summed Scores as the predictor and Spring 2001 SOL Total Reading scores as the dependent variable resulted in an R^2 value of .25, indicating that 25% of the variability in spring SOL Total Reading scores could be predicted by the fall PALS Entry Level Summed Scores. These data indicate that scores on PALS administered at the beginning of third grade are significant predictors

of end-of-year reading achievement on the SOL test and explain roughly one-third of the variance in the SOL reading score.

PALS 1–3 Scores from Spring to Fall. In addition to assessing the predictive validity of PALS with other measures taken at a later time, we also examine the extent to which PALS spring scores are predictive of students' PALS scores in the following fall assessment. Across the time frame from spring to fall, we assess the relationship between first-graders' spring PALS scores and their fall (second grade) PALS scores, and between second-graders' spring PALS scores and their fall (third grade) PALS scores. In 2013, spring scores predicted a significant amount of the variance in PALS scores the following fall.

For first-graders screened in Spring 2013, simple correlations suggested that Word Recognition and Spelling scores were significantly related to the Entry Level Summed Scores in the fall of second grade (r = .81 and .83 respectively, $p < .001$). Further, regression analyses were used to examine the predictive relationship between PALS scores in the spring and the following fall. In these regression analyses, the combination of Word Recognition and Spelling scores yielded an R^2 of .77, suggesting that 77% of the variability in fall Entry Level Summed Scores could be predicted based on their relationship to Word Recognition and Spelling scores from the previous spring. The results of the regression analysis are summarized in Table 26. The regression fit (R^2 = .77) was good, and the overall relationship was statistically significant. Holding performance on the spring first grade Spelling task constant, a 1-point increase in performance on the first grade Word Recognition task in the spring was associated, on average, with an approximate 1.4-point gain on the Entry Level Summed Score in the fall. Similarly, holding performance on the spring first grade Word Recognition task constant, a 1-point increase in performance on the spring first grade Spelling task was associated, on average, with an approximate .76-point gain on the fall Entry Level Summed Score. Both predictors achieved statistical significance.

A similar pattern held for second-graders: Word Recognition and Spelling scores from the spring correlated significantly with fall Entry Level Summed Scores (r = .80 and .86 respectively, $p < .001$), and the combination of these tasks resulted in an R^2 value of .81. In other words, 81% of the variance in fall Entry Level Summed Scores could be predicted based on previous spring scores. The results of the regression analysis are summarized in Table 27, showing that

Table 26 Spring 2013 First-Grade Scores (First-Grade Word Recognition and First-Grade Spelling) Predicting Fall 2013 Second-Grade Entry Level Summed Scores

	Descriptive Statistics			Regression Coefficients			
Variables	Fall 2013 Second-Grade Entry Level Sum Score	Fall 2013 First-Grade Word List	Fall 2014 Second-Grade Spelling	B	Beta	*t*	*p*
Spring 2013 First-Grade Word List	.81	—	—	1.35	0.43	133.07	<.001
Spring 2013 First-Grade Spelling	.83	.68	—	0.76	0.51	159.58	<.001
Mean	48.23	16.07	30.99	R^2 = .77			
(sd)	13.98	5.07	10.94				

B designates raw regression coefficients; Beta designates standardized regression coefficients; t = the test size for null hypothesis that the coefficient equals zero; p is an abbreviation for probability; and sd = standard deviation.

the regression fit (R^2 = .81) was good and the overall relationship was statistically significant. Holding performance on the spring second-grade Spelling task constant, a 1-point increase in performance on the spring second-grade Word Recognition task was associated with an approximate 1.5-point gain on the fall Entry Level Summed Score, on average. Similarly, holding performance on the spring second-grade Word Recognition task constant, a 1-point increase in performance on the spring second-grade Spelling task was associated with an approximate .78-point gain on the fall Entry Level Summed Score, on average. Both predictors achieved statistical significance.

PALS 4–8 and SOL. To test the accuracy of the PALS Plus Entry Level Summed Score in identifying upper elementary and middle school students at risk for not passing Virginia's Standards of Learning (SOL) in English, we conducted a series of receiver-operating characteristic (ROC) curve analyses. ROC curve analysis is a tool for evaluating how well an assessment classifies subjects into one of two categories, in this case being at risk or not being at risk for failing the Virginia Standards of Learning (SOL) in English. The Area Under the Curve (AUC) statistic of a ROC curve analysis is an indication of overall diagnostic accuracy (AUC values of 1.00 indicate perfect classification accuracy; values of .50 indicate accuracy no better than chance). Based on guidelines suggested

Table 27 Spring 2013 Second-Grade Scores (Second-Grade Word Recognition and Second-Grade Spelling) Predicting Fall 2013 Third-Grade Entry Level Summed Scores

	Descriptive Statistics			Regression Coefficients			
Variables	Fall 2013 Third-Grade Entry Level Sum Score	Fall 2013 Second-Grade Word List	Fall 2013 Second-Grade Spelling	B	Beta	*t*	*p*
Spring 2013 Second-Grade Word List	.80	—	—	1.51	0.40	110.5	<.001
Spring 2013 Second-Grade Spelling	.86	.66	—	0.78	0.58	161.7	<.001
Mean	60.86	17.85	43.02	R^2 = .81			
(sd)	15.08	4.13	11.89				

B designates raw regression coefficients; Beta designates standardized regression coefficients; t = the test size for null hypothesis that the coefficient equals zero; p is an abbreviation for probability; and sd = standard deviation.

Table 28 ROC Curve Analysis Studies in Grades 4–8

Form	Grade(s)	Type	*n*	AUC	Discrimination*
A	4	Predictive	274	.82	Excellent
B	4	Predictive	920	.78	Acceptable
A	5	Predictive	267	.82	Excellent
B	5	Predictive	820	.74	Acceptable
A	6, 7 & 8	Predictive	42	.70	Acceptable
B	6, 7 & 8	Predictive	157	.85	Excellent

** Based on Hosmer & Lemeshow (1989)*

by Hosmer and Lemeshow (1989), PALS Plus has acceptable to excellent discriminating capabilities in grades 4 and 5 and acceptable discriminating capabilities in grades 4 through 8. Table 28 reports AUC statistics from studies using PALS PLUS and the external indicator: The Virginia SOL for English. Note that the AUC values range from .78 to .82 for the fourth grade SOLs, from .74 to .82 for the fifth grade SOLs, and .70–.85 for the sixth, seventh, and eighth grade SOLs. Figures 2 and 3 shows the AUCs for spring 2013 PALS Plus Entry Level scores for grades 4 and 5 predicting risk using spring 2014 SOLS for English.

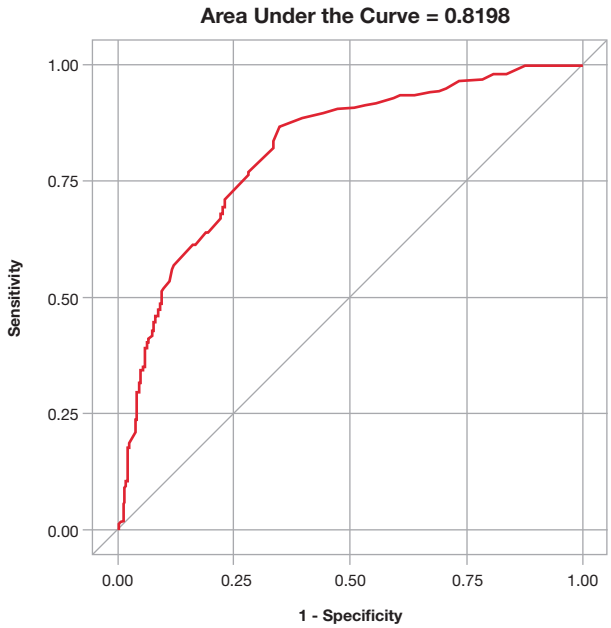
Concurrent Validity

Concurrent validity refers to the degree to which a given measure is consistent with some independent standard.[86] Concurrent validity is desirable for instruments used for diagnostic purposes or for instruments that are designed to measure a specified construct.[87] To measure the concurrent validity of PALS 1–3, the 2000–01 PALS 1–3 screening results were compared against four different independent standards. For first grade, comparisons were made using the *Qualitative Reading Inventory-II* (QRI-II),[88] the *Developmental Reading Assessment* (DRA),[89] the *Stanford-9* (1996) Total Reading scaled score, and the *California Achievement Test* (CAT/5) (1992) Total Reading scaled score.

For second grade, comparisons were made using the DRA and the second grade Stanford-9 achievement test. For third grade, PALS 1–3 was compared against the DRA and the Virginia Standards of Learning Total Reading score. For all three grades, the piloted alternative forms of PALS 1–3 tasks were compared to their corresponding tasks, administered in Fall 2000.

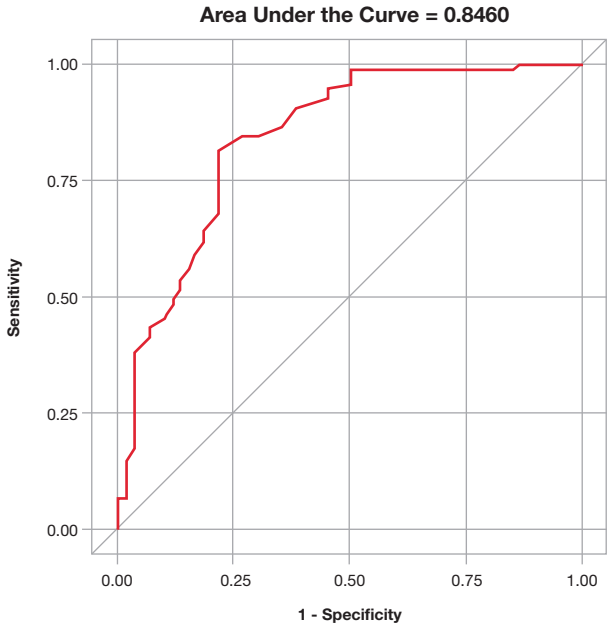
PALS 1–3 and QRI-II. A total of 65 first-grade students read orally from the PALS 1–3 passages and from the corresponding leveled passages in the QRI-II. An accuracy score for Word Recognition in Context was derived for each passage read. The bivariate correlation between a student's instructional

Figure 2 Area Under the Curve (AUC) for ES Students, Form A



AUC = .82 (n = 541)

Figure 3 Area Under the Curve (AUC) for MS Students, Form B



AUC = .85 (n = 157)

Table 29 Spearman Correlations Between PALS 1–3 and QRI-II Word Lists, Spring 2000

PALS Word List	QRI-II Word List	Correlation
Preprimer	QRI-PP	.73
Primer	QRI-P	.90
Grade 1	QRI 1	.87
Grade 2	QRI 2	.80
Grade 3	QRI 3	.80

p < .01.

reading level, as determined by the accuracy of oral reading of PALS passages and QRI-II passages at the same level, was medium-high and significant ($r = .73$, $p < .01$). Bivariate correlations were also calculated for 146 first-grade students, comparing their instructional reading level scores for oral reading on the QRI-II with their spring PALS Entry Level Summed Score. This correlation was also medium-high and significant: ($r = .73$, $p < .01$).

Medium-high and significant correlations among PALS 1–3 oral reading scores, QRI-II oral reading scores, and Spring 2001 PALS Entry Level Summed Scores indicate strong relationships among these three variables within the construct of instructional reading level, which corroborates the Spring 2000 pilot results comparing PALS 1–3 word lists to QRI-II word lists.[90] In that pilot study, 679 students in grades one through three read word lists from PALS 1–3 and corresponding word lists from QRI-II. Correlations between the PALS and QRI-II word lists ranged from .73 for the preprimer lists to .90 for the primer lists ($p < .01$). Table 29 shows the word-list correlations by grade level. The correlations between PALS 1–3 and QRI-II word lists, in combination with the significant correlations among passage and overall scores, indicate a strong relationship between PALS 1–3 and QRI-II.

PALS 1–3 and DRA. In Spring 2001, 197 first, second, and third grade students were assessed with the Developmental Reading Assessment (DRA). Students read orally from passages leveled according to increments of difficulty, and an instructional level was obtained for each student. The bivariate correlation between students' instructional reading level on the DRA and their Spring 2001 PALS Entry Level Summed Score was .82 ($p < .01$). An independent reading level was also obtained for 96 first- through third-grade students. The overall correlation between students' independent reading level on the DRA and their Spring 2001 PALS Entry Level Summed Score was .81 ($p < .01$). Significantly high correlations between students' reading level as indicated by the DRA and their Spring 2001 PALS Entry Level Summed Score demonstrate a strong relationship between the DRA assessment and PALS 1–3.

PALS 1–3 and California Achievement Test. Also in Spring 2001, 195 first-grade students were assessed with the California Achievement Test (CAT/5) and PALS 1–3. These students represented a mixed sample of Identified and Non-identified EIRI students as determined by their Fall 2000 PALS 1–3 Entry Level Summed Scores. Student performance on both assessments was compared. The bivariate correlation between the Total Reading scaled score on the CAT/5 and the PALS 1–3 Entry Level Summed Score was medium-high and significant ($r = .75$, $p < 01$). The correlation between the scaled score for Word Analysis on the CAT/5 and the PALS 1–3 Entry Level Summed Score was also medium-high and significant ($r = .67$, $p < .01$). Results on the CAT/5 Word Analysis subtest significantly correlated with those on the Spring 2001 PALS Spelling task ($r = .66$, $p < .01$). The PALS Spelling task also correlated significantly with the CAT/5 Total Reading scaled score ($r = .70$, $p < .01$). Medium-high, significant correlations among the total scores and subtest scores of the CAT/5 and PALS 1–3 indicate a considerable amount of shared variance among the measures when administered to first-graders at the end of the year.

PALS 1–3 and Stanford-9. A total of 174 first-grade students and 50 second grade students were assessed using the Stanford-9 achievement test as well as the PALS 1–3 in Spring 2001. These students had not

met the Fall 2000 PALS Entry Level Summed Score criterion and had consequently been identified as needing additional instruction in reading. Their end-of-year Stanford-9 Total Reading scaled scores were compared to their Spring 2001 PALS 1–3 performance. For first grade, the bivariate correlation between the first grade Stanford-9 Total Reading scaled score and the Spring 2001 PALS Entry Level Summed Score was medium-high ($r = .67$, $p < .01$). For second grade, the correlation between the Stanford-9 Total Reading scaled score and the Spring 2001 PALS Entry Level Summed Score was medium ($r = .57$, $p < .01$). The correlations for both grades are statistically significant and suggest a strong relationship between the Stanford-9 and PALS 1–3 when administered at the end of grades one and two to children receiving interventions funded by the EIRI.

PALS 1–3 and Virginia's SOL. The Standards of Learning (SOL) assessment is given to all third-grade children in Virginia in the spring of the school year to determine students' proficiency and ability to meet prescribed standards, including standards for reading achievement. Data were examined on 15,650 students, who were assessed with both the SOL reading component and PALS 1–3 in Spring 2012. The correlation between the SOL Total Reading score and the spring PALS Entry Level Summed Score was medium and significant: ($r = .57$, $p < .01$). In addition, the SOL Total Reading Score was significantly correlated with both the PALS Spelling ($r = .50$, $p < .01$) and the Word Recognition in Isolation task ($r = .46$, $p < .01$). These bivariate correlations indicate a significant amount of shared variance among the reading tasks on the Virginia SOL assessment and PALS 1–3 administered in the spring of third grade.

Differential Item Functioning

Differential item functioning refers to the consistency of response to specific items or tasks across groups. The Mantel-Haenszel statistic can be defined as the average factor by which the odds that members of one group will answer a question correctly exceed the corresponding odds for comparable members of another group. The Mantel-Haenszel statistic is a form of an odds ratio.[91]

To explore the consistency of responses to PALS items, we examined the responses to PALS Entry Level tasks from groups defined as Identified and Not-identified for additional instruction under EIRI, based on these students' PALS Entry Level Summed Score. Since the purpose of PALS is to identify children in need of additional instruction, individual

Table 30 Mantel-Haenszel Statistics (general association) for First Through Third Grade Identified and Not-Identified Groups

	Spring 2008		Spring 2009		Spring 2010	
PALS Task	GA*	*p*	GA*	*p*	GA*	*p*
First Grade						
1st Grade Word List	47,004	< .001	45,502	<.001	46,625	<.001
Spelling	35,817	< .001	35,511	<.001	35,656	<.001
Second Grade						
2nd Grade Word List	34,511	< .001	30,865	<.001	35,903	<.001
Spelling	45,096	< .001	44,039	<.001	45,605	<.001
Third Grade						
3rd Grade Word List	6,151	< .001	6,851	<.001	7,862	<.001
Spelling	8,399	< .001	10,164	<.001	10,414	<.001

**General association*

items within each PALS task should function differently for Identified and Not-identified groups. This was the case for first-graders' fall and spring scores from the 1998–99 and 1999–2000 samples, as well as for scores from first- through third-graders in every statewide sample since 2000. Table 30 displays the Mantel-Haenszel statistic (based on item scores) for each PALS subtask for first-, second-, and third-graders for 2008 through 2010. As can be seen, the general association statistic is significant for all PALS tasks at all grade levels.

The technical adequacy of PALS Plus has been established through pilot and field tests and statistical analyses of PALS scores for hundreds of thousands of Virginia students in grades one through eight. The reliability of individual subtasks has been documented through the use of Cronbach's alpha, item-to-total correlations, difficulty indices, and discrimination indices. Reliability coefficients for individual Entry Level tasks on PALS 1–3 have ranged from .81 to .96 and demonstrate the adequacy of their internal consistency. Differential item function analyses (DIF) using ETS classification demonstrate negligible evidence of bias for or against reference groups based on gender and race. Inter-rater reliabilities expressed as Pearson correlation coefficients have ranged from .94 to .99, demonstrating that PALS 1–3 tasks can be scored consistently across individuals. In all of these analyses, PALS Plus has been shown to be steady, reliable, and consistent among many different groups of users.

Further analyses have also supported the content, construct, and criterion-related validity of PALS 1–3. Principal components analyses, discriminant function analyses, and intercorrelations among tasks support the construct validity of PALS 1–3. Regression analyses have demonstrated the predictive relationship between PALS 1–3 Entry Level Summed Scores in the fall and Stanford-9 and SOL reading scores in the spring. Coefficients of determination have demonstrated that a significant proportion of the variability in spring Stanford-9 and SOL reading scores can be explained by the PALS 1–3 Entry Level Summed Score from nine months earlier. Similar analyses provide evidence of the concurrent validity of PALS 1–3, using the CAT/5 and the QRI for grade one; the Stanford-9 for grade two; the DRA for grades one, two, and three; and the SOL reading component for grade three.

In addition, differential item functioning analyses using the Mantel-Haenszel statistic demonstrate the consistency of responses to specific tasks across groups of Identified and Not-identified students. All of these analyses provide evidence of the validity of PALS 1–3 as an early reading assessment that reliably identifies students in need of additional instruction, and provides diagnostic information that is useful in planning that instruction.

In summary, PALS Plus provides an assessment tool with good evidence of validity that can be used reliably to screen students in grades one through eight for difficulty in reading.

Section VI

References

Abouzeid, M. P. (1986). Developmental stages of word knowledge in dyslexia (Doctoral dissertation, University of Virginia, 1986). Dissertation Abstracts International, 48, 09A2295.

Adams, M. (1990). *Beginning to read: Thinking and learning about print.* Cambridge, MA: MIT Press.

American Educational Research Association (AERA), American Psychological Association (APA), & National Council on Measurement in Education (NCME) (1999). *Standards for Educational and Psychological Testing.* Washington, DC: American Psychological Association.

Bader, L. (1998). *Bader Reading and Language Inventory* (3rd ed.). Upper Saddle River, NJ: Prentice-Hall.

Ball, E., & Blachman, B. (1991). Does phoneme awareness training in kindergarten make a difference in early word recognition and developmental spelling? *Reading Research Quarterly, 26,* 49–66.

Barnes, W. G. (1993). Word sorting: The cultivation of rules for spelling in English. *Reading Psychology: An International Quarterly, 10*, 293–307.

Barr, R., Blachowicz, C. & Wogman-Sadow, M. (1995). *Reading diagnosis for teachers: An instructional approach* (3rd ed.). White Plains, NY: Longman.

Bear, D. (1989). Why beginning reading must be word-by-word: Disfluent oral reading and orthographic development. *Visible Language, 23*(4), 353–367.

Bear, D., & Barone, D. (1998). *Developing literacy: An integrated approach to assessment and instruction*. Boston: Houghton Mifflin.

Bear, D., Invernizzi, M., Templeton, S., & Johnston, F. (2004). *Words their way: Word study for phonics, vocabulary and spelling instruction* (3rd ed.). Upper Saddle River, NJ: Merrill.

Beaver, J. (1997). *Developmental Reading Assessment*. New York: Celebrations Press.

Biemiller, A. (2010). *Words worth teaching: Closing the vocabulary gap.* McGraw-Hill SRA.

Betts, E. (1946). *Foundations of reading instruction with emphasis on differentiated guidance*. New York: American Book Company.

Bodrova, E., Leong, D., & Semenov, D. (1999). 100 most frequent words in books for beginning readers. Retrieved from http://www.mcrel.org/resources/literacy/road/100words

Burton, R., Hill, E., Knowlton, L., & Sutherland, K. (1999). A Reason for Spelling. Retrieved from http://www.areasonfor.com/index.htm

Byrne, B., & Fielding-Barnsley, R. (1989). Phonemic awareness and letter knowledge in the child's acquisitions of the alphabetic principle. *Journal of Educational Psychology, 81*, 313–321.

California Achievement Test (5th ed.). (1992). Monterey, CA: McGraw-Hill.

Cantrell, R. (1991). Dialect and spelling in Appalachian first grade children (Doctoral dissertation, University of Virginia, 1991). Dissertation Abstracts International, 53, 01A0112.

Carroll, J. B., Davies, P., & Richman, B. (1971). *The American Heritage word frequency book*. Boston: Houghton Mifflin.

Cary, L., & Verhaeghe, A. (1994). Promoting phonemic analysis ability among kindergartners. *Reading and Writing, 6*, 251–278.

Catts, H. W. (1993). The relationship between speech-language impairments and reading disabilities. *Journal of Speech & Hearing Research, 36*, 948–958.

Clay, M. M. (1979). *Reading: The patterning of complex behavior.* Auckland, New Zealand: Heinemann.

Code of Fair Testing Practices in Education (1988). Washington, DC: Joint Committee on Testing Practices.

Coltheart, M. (1981). The MRC psycholinguistic database. *The Quarterly Journal of Experimental Psychology, 33*(4), 497-505.

Cunningham, A. (1990). Explicit versus implicit instruction in phonemic awareness. *Journal of Experimental Child Psychology, 50*, 429–444.

Davies, M. (2008). The corpus of contemporary American English: 425 million words, 1990–present.

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

Dolby, J. L., Resnikoff, H. L., & MacMurray, E. (1963, November). A tape dictionary for linguistic experiments. In *Proceedings of the November 12-14, 1963, fall joint computer conference* (pp. 419–423). ACM.

Dolch, E. W. (1936). *A combined word list*. Boston: Ginn.

Dorans, N. (1989). Two new approaches to assessing differential item functioning: standardization and the Mantel-Haenszel method. *Applied Measurement in Education, 2,* 217–233.

Duvvuri, R., & Millard, R. T. (1995). *The educator's word frequency guide.* Brewster, NY: Touchstone Applied Science Associates.

EDL (1997). *EDL Core Vocabularies in reading, mathematics, science, and social studies*. Orlando, FL: Steck-Vaughn.

Flesch, R. (1948). A new readability yardstick. *Journal of applied psychology, 32*(3), 221.

Fry, E. (1977). Fry's readability graph: Clarifications, validity and extension to level 17. *Journal of Reading, 21*, 242–252.

Ganske, K. (1999). The developmental spelling analysis: A measure of orthographic knowledge. *Educational Assessment, 6*, 41–70.

George, D., & Mallery, P. (2001). *SPSS for Windows*. Needham Heights, MA: Allyn & Bacon.

Gill, C. E. (1980). An analysis of spelling errors in French (Doctoral dissertation, University of Virginia, 1980). Dissertation Abstracts International, 41, 09A3924.

Gill, T. P. (1985). The relationship between spelling and word recognition of first, second, and third graders (Doctoral dissertation, University of Virginia, 1985). Dissertation Abstracts International, 46, 10A2917.

Gronlund, N. E. (1985). *Measurement and evaluation in teaching.* New York: Macmillan.

Gunning, T. (1997). *Best books for beginning readers*. Boston: Allyn & Bacon.

Harris, A. J., & Sipay, E. (1985). *How to increase reading ability: A guide to developmental and remedial methods* (8th ed.). New York: Longman.

Henderson, E. (1990). *Teaching spelling* (2nd ed.). Boston: Houghton Mifflin.

Henderson, E., & Beers, J. (1980). *Developmental and cognitive aspects of learning to spell*. Newark, DE: International Reading Association.

Henderson, E. H., & Templeton, S. (1994). *Spelling and Vocabulary* (Teacher's Edition). Boston: Houghton Mifflin.

Hiebert, E. H. (1998). Text matters in learning to read (CIERA Rep. No. 1-001). Ann Arbor, MI: University of Michigan, Center for the Improvement of Early Reading Achievement.

Hoffman, J. V. (2001). Words: On words in leveled texts for beginning readers. Paper presented at the National Reading Conference, San Antonio, TX, December 2001.

Hoffman, J. V., & Isaacs, M. E. (1991). Developing fluency through restructuring the task of guided oral reading. *Theory into Practice, 30*, 185–194.

Holland, P. W., & Thayer, D. T. (1988). Differential item performance and the Mantel-Haenszel procedure. *Test validity*, 129–145.

Huang, F. (2014). Using a bifactor model to assess the factor structure of the Phonological Awareness Literacy Screening for Grades One through Three. *Journal of Psychoeducational Assessment.* doi: 10.1177/0734282914525026.

Invernizzi, M. (1992). The vowel and what follows: A phonological frame of orthographic analysis. In S. Templeton & D. Bear (Eds.), *Development of orthographic knowledge and the foundation of literacy* (pp.105–136). Hillsdale, NJ: Lawrence Erlbaum and Associates, Inc.

Invernizzi, M., Abouzeid, M., & Gill, T. (1994). Using students' invented spelling as a guide for spelling instruction that emphasizes word study. *The Elementary School Journal, 95*, 155–167.

Invernizzi, M., Juel, C., Rosemary, C., & Richards, H. (1997). At-risk readers and community volunteers: A three year perspective. *Scientific Studies of Reading, 1*, 277–300.

Invernizzi, M. A., Landrum, T. J., Howell, J. L., & Warley, H. P. (2005). Toward the peaceful coexistence of test developers, policy makers, and teachers in an era of accountability. *Reading Teacher, 58*(7), 2–10.

Invernizzi, M., & Meier, J. D. (2000). *Phonological Awareness Literacy Screening-Grades 1–3 Teacher's Manual*. Charlottesville, VA: University Printing Services.

Invernizzi, M., Meier, J. D., Swank, L., & Juel, C. (1997). *PALS: Phonological Awareness Literacy Screening*. Charlottesville, VA: University Printing Services.

Invernizzi, M., Robey, R., & Moon, T. (1999). Phonological Awareness Literacy Screening (PALS) 1997–1998: Description of Sample, First-Year Results, Task Analyses, and Revisions. Technical manual and report prepared for the Virginia Department of Education. Charlottesville, VA: University Printing Services.

Invernizzi, M., Robey, R., & Moon, T. (2000). Phonological Awareness Literacy Screening (PALS) 1998–1999: Description of Sample & Second-Year Results. Technical manual and report prepared for the Virginia Department of Education. Charlottesville, VA: University Printing Services.

Johnson, M. S., Kress, R. A., & Pikulski, J. J. (1987). *Informal reading inventories* (2nd ed). Newark, DE: International Reading Association.

Johnston, F., Invernizzi, M., & Juel, C. (1998). *Book Buddies: Guidelines for volunteer tutors of emergent and early readers.* New York: Guilford Publications.

Juel, C. (1988). Learning to read and write: A longitudinal study of fifty-four children from first through fourth grades. *Journal of Educational Psychology, 80*, 437-447.

Juel, C., & Minden-Cupp, C. (2000). Learning to read words: Linguistic units and instructional strategies. *Reading Research Quarterly, 35,* 458–492.

Leslie, L., & Caldwell, J. (1995). *Qualitative Reading Inventory-II*. New York: HarperCollins

Lipson, M., & Wixson, K. (1997). *Assessment and instruction of reading and writing disability: An interactive approach.* New York: Longman.

Lyon, R. (1998). Overview of reading and literacy initiatives. Retrieved January 15, 2002, from National Institute of Health, National Institutes of Child Health and Human Development, Child Development and Behavior Branch Web site: http://156.40.88.3/publications/pubs/jeffords.htm

McBride-Chang, C. (1998). The development of invented spelling. *Early Education & Development, 9*, 147–160.

McConkie, G. W., & Zola, D. (1987). *Eye movement techniques in studying differences among developing readers*. Cambridge: Bolt Beranek and Newman.

McLoughlin, J., & Lewis, R. (2001). *Assessing students with special needs.* Upper Saddle River, NJ: Merrill/Prentice Hall.

Mehrens, W., & Lehmann, I. (1987). *Using standardized tests in education.* New York: Longman.

Metzoff, J. (1998). *Critical thinking about research*. Washington, DC: American Psychological Association.

Microsoft office [Computer software]. (2000). Redman, WA: Microsoft Corporation.

Moe, A. J., Hopkins, C. J., & Rush, R. T. (1982). *The vocabulary of first grade children*. Springfield, IL: Charles C. Thomas.

Morris, D. (1992). What constitutes at risk: Screening children for first grade reading intervention. In W. A. Second (Ed.), *Best Practices in School-Speech Pathology* (pp.17–38). San Antonio, TX: Psychological Corporation.

Morris, D. (1993). The relationship between children's concept of word in text and phoneme awareness in learning to read: A longitudinal study. *Research in the Teaching of English, 27*, 133–154.

Morris, D. (1999a). Preventing reading failure in the primary grades. In T. Shanahan & F. V. Rodriguez-Brown (Eds.), *National Reading Conference Yearbook, 48*, (pp.17–38).

Morris, D. (1999b). *The Howard Street Tutoring Manual*. New York: Guilford Press.

Morris, D., Bloodgood, J., Lomax, R., & Perney, J. (2003). Developmental steps in learning to read: A longitudinal study in kindergarten and first grade. *Reading Research Quarterly, 38, 302–328.*

National Reading Panel (2000). Teaching children to read: An evidence-based assessment of the scientific research literature on reading and its implications for reading instruction. Reports to the subgroups (NIH Publication No. 00-4754). Washington, DC: U.S. Department of Health and Human Services.

O'Connor, R. E., & Jenkins, J. R. (1995). Improving the generalization of sound/symbol knowledge: Teaching spelling to Kindergarten children with disabilities. *Journal of Special Education, 29*, 255–275.

Perfetti, C. (2007). Reading ability: Lexical quality to comprehension. *Scientific studies of reading, 11*(4), 357–383.

Perney, J., Morris, D., & Carter, S. (1997). Factorial and predictive validity of first graders' scores on the Early Reading Screening Instrument. *Psychological Reports, 81*, 207–210.

Peterson, B. L. (1988). Characteristics of texts that support beginning readers (Doctoral dissertation, Ohio State University, 1988). Dissertation Abstracts International, 49, 8A2105.

Powell, W. R. (1971). The validity of the instructional reading level. In R. E. Leibert (Ed.), *Diagnostic viewpoints in reading* (pp 121–133). Newark, DE: International Reading Association.

Rathvon, N. (2004). *Early reading assessment: A practitioner's handbook*. New York: The Guilford Press.

Renaissance Learning, Inc. (n.d.). Advantage-TASA Open Standard Readability Formula for Books (ATOS), Available from http://www.renlearn.com

Richardson, E., & DiBenedetto, B. (1985). *Decoding skills test.* Los Angeles, CA: Western Psychological Services.

Rinsland, H. D. (1945). *A basic vocabulary of elementary school children.* New York: Macmillan.

Roberts, E. (1992). The evolution of the young child's concept of word as a unit of spoken and written language. *Reading Research Quarterly, 27*, 124–138.

Santa, C., & Hoien, T. (1999). An assessment of early steps: A program for early intervention of reading problems. *Reading Research Quarterly, 34*, 54–79.

Scarborough, H. S. (1998). Early identification of children at risk for reading disabilities: Phonological awareness and some other promising predictors. In B. K. Shapiro, P. J. Accardo, & A. J. Capute (Eds.), *Specific reading disability: A view of the spectrum* (pp. 75–199). Timonium, MD: York Press.

Scarborough, H. S. (2000). Predictive and causal links between language and literacy development: Current knowledge and future direction. Paper presented at the Workshop on Emergent and Early Literacy: Current Status and Research Direction, Rockville, MD, 2000.

Schlagal, R. C. (1989). Informal and qualitative assessment of spelling. *The Pointer, 30*(2), 37–41.

Schreiber, P. A. (1987). Prosody and structure in children's syntactic processing. In R. Horowitz & S. J. Samuels (Eds.), *Comprehending oral and written language*. New York: Academic Press.

Shanker, J. L., & Ekwall, E. E. (2000). *Ekwall/Shanker reading inventory* (4th ed.). Boston: Allyn & Bacon.

Smith, S. B., Simmons, D. C., & Kameenui, E. J. (1995). Synthesis of research on phonological awareness: Principles and implications for reading acquisition (Technical Rep. No. 21). [Eugene, OR]: University of Oregon, National Center to Improve the Tools of Educators.

Snow, C., Burns, M., & Griffin, P. (Eds.). (1998). *Preventing reading difficulties in young children*. Washington, DC: National Academy Press.

Spache, G. D. (1974). *Good reading for poor readers* (Rev. 9th ed.). Champaign, IL: Garrard Publishing.

Stanford Achievement Test (9th ed.). (1996). San Antonio, TX: Harcourt Brace.

Stauffer, R., Abrams, J., & Pikulski, J. (1978). *Diagnosis, correction, and prevention of reading disabilities*. New York: Harper & Row.

Stieglitz, E. (1997). *The Stieglitz Informal Reading Inventory* (2nd ed.). Needham Heights, MA: Allyn & Bacon.

Stenner, A. J., Horabin, I., Smith, D. R., & Smith, M. (1988). The lexile framework. *Durham, NC: MetaMetrics*.

Supts. Memo No. 92 (2007). Early Intervention Reading Initiative—Application Process for the 2007–2008 school year. Retrieved June 27, 2007, from Commonwealth of Virginia Department of Education Web site: http://www.pen.k12.va.us/VDOE/suptsmemos/2007/info92.html

Templeton, S. (1983). Using the spelling/meaning connection to develop word knowledge in older students. *Journal of Reading, 27*, 8–14.

Templeton, S., & Bear, D. (1992). *Development of orthographic knowledge and the foundations of literacy: A memorial festschrift for Edmund Hardcastle Henderson*. Hillsdale, NJ: Lawrence Erlbaum and Associates, Inc.

Torgesen, J. K., & Davis, C. (1996). Individual difference variables that predict responses to training in phonological awareness. *Journal of Experimental Child Psychology, 63*, 1–21.

Torgesen, J. K., & Wagner, R. K. (1998). Alternative diagnostic approaches for specific developmental reading disabilities. *Learning Disabilities Research & Practice, 31*, 220–232.

Torgesen, J. K., Wagner, R. K., Rashotte, C.A., Burgess, S., & Hecht, S. (1997). Contributions of phonological awareness and rapid automatic naming ability to the growth of word-reading skills in second to fifth grade children. *Scientific Studies of Reading*, Vol. 1(2), 161–185.

Touchstone Applied Science Associates. (1979-1991). Degrees of reading power (DRP): An effectiveness measure of reading. Available from http://www.questarai.com.

U.S. Department of Education (1995). National Center for Educational Statistics: Listening to Children Read Aloud: Oral Fluency, 1(1). Washington, DC.

Vellutino, F., & Scanlon, D. (1987). Phonological coding, phonological awareness, and reading ability: Evidence from a longitudinal and experimental study. *Merrill-Palmer Quarterly, 33*, 321–363.

Vellutino, F. R., Scanlon, D. M., Sipay, E. R., Small, S. G., Pratt, A., Chen, R., & Denckla, M. B. (1996). Cognitive profiles of difficult-to-remediate and readily remediated poor readers: Early intervention as a vehicle for distinguishing between cognitive and experiential deficits as a basic cause of specific reading disability. *Journal of Educational Psychology, 88*, 601–638.

Viise, N. (1992). A comparison of child and adult spelling (Doctoral dissertation, University of Virginia, 1992). Dissertation Abstracts International, 54, 5A1745.

Virginia Department of Education. (1995). Standards of learning for Virginia public schools. Richmond, VA: Commonwealth of Virginia Board of Education.

Wagner, R. K., Torgesen, J. K., Laughon, P., Simmons, K., & Rashotte, C. A. (1993). Development of young readers' phonological processing abilities. *Journal of Educational Psychology, 85*, 83–103.

Weaver, B. (2000). *Leveling books K-6: Matching readers to text.* Newark, DE: International Reading Association.

West, J., Denton, K., & Reaney, L. (2001). Early childhood longitudinal study: Kindergarten class of 1998–1999. National Center for Education Statistics (NCES 2001-023). U.S. Department of Education. Washington, DC: U.S. Government Printing Office.

Wheeler, L. R., & Smith, E. H. (1954). A practical errors of normal and disabled students on achievement levels one through four: Instructional implications. *Bulletin of the Orton Society, 40*, 138–151.

Zinbarg, R. E., Revelle, W., Yovel, I., & Li, W. (2005). Cronbach's α, Revelle's β, and McDonald's ωH: Their relations with each other and two alternative conceptualizations of reliability. *Psychometrika, 70*, 123–133.

Zutell, J. B. (1975). Spelling strategies of primary school children and their relationship to the Piagetian concept of decentration (Doctoral dissertation, University of Virginia, 1975). Dissertation Abstracts International, 36, 8A5030.

Section VII

Endnotes

[1] Standards for Educational and Psychological Testing (1999), prepared jointly by the American Educational Research Association, the American Psychological Association, and the National Council on Measurement in Education.

[2] Virginia Department of Education, 1995, p. 61.

[3] Catts, 1993; Lyon, 1998; Scarborough, 1998, 2000; Torgesen & Wagner, 1998; Torgesen, Wagner, Rashotte, Burgess, & Hecht, 1997; Vellutino et al., 1996.

[4] Invernizzi, Juel, Rosemary, & Richards, 1997; Perney, Morris, & Carter, 1997; Santa & Hoien, 1999.

[5] Bodrova, Leong, & Semenov, 1999.

[6] Henderson, 1990.

[7] Burton, Hill, Knowlton, & Sutherland, 1999.

[8] Dolch, 1936.

[9] Rinsland, 1945.

[10] Henderson & Templeton, 1994.

[11] Leslie & Caldwell, 1995.

[12] Stieglitz, 1997.

[13] Bader, 1998.

[14] Richardson & Benedetto, 1985.

[15] Shanker & Ekwall, 2000.

[16] Johnston et al., 1998.

[17] Morris, 1999b.

[18] Carroll, Davies, & Richman, 1971.

[19] Davies, 2008.

[20] Duvvuri & Millard, 1995.

[21] Dale & O'Rourke, 1981.

[22] Biemiller, 2010.

[23] Dolby, Resnikoff & MacMurray, 1963.

[24] Coltheart, 1981.

[25] Holland & Thayer, 1988.

[26] McBride-Chang, 1998.

[27] Torgesen & Davis, 1996.

[28] Perfetti, 2007.

[29] Henderson, 1990.

[30] Templeton & Bear, 1992.

[31] Abouzeid, 1986; Barnes, 1993; Bear, 1989; Cantrell, 1991; Ganske, 1999; Gill, C.E., 1980; Gill, T.P., 1985; Henderson, 1990; Henderson & Beers, 1980; Invernizzi, 1992; Schlagal, 1989; Templeton, 1983; Viise, 1992; Zutell, 1975.

[32] Johnson, Kress, & Pikulski, 1987.

[33] Hoffman & Isaacs, 1991; Morris, 1999a.

[34] Gunning, 1997; Hiebert, 1998; Hoffman, 2001; Peterson, 1988; Weaver, 2000.

[35] Flesch, 1948.

[36] Spache, 1974.

[37] Harris & Sipay, 1985.

[38] Wheeler & Smith, 1954.

[39] Fry, 1977.

[40] Renaissance Learning, Inc. (n.d.)

[41] Touchstone Applied Science Associates, 1979-1991.

[42] Stenner, Horabin, Smith, & Smith, 1988.

[43] McNamara, Louwerse, Cai, & Graesser, 2005.

[44] U.S. Department of Education, 1995.

[45] Johnson et al., 1987; Stauffer et al.,1978.

[46] National Reading Panel, 2000.

[47] Morris, 1993; Morris, Bloodgood, Lomax, & Perney, 2002.

[48] Adams, 1990; Snow, Burns, & Griffin, 1998.

[49] Invernizzi, Meier, Swank, & Juel, 1997.

[50] Clay, 1979; Henderson & Beers, 1980; Morris, 1992; Roberts, 1992.

[51] Henderson & Beers, 1980.

[52] Morris, 1993; Morris et al., 2002.

[53] Smith, Simmons, & Kameenui, 1995.

[54] Vellutino et al., 1996.

[55] Vellutino & Scanlon, 1987.

[56] Wagner, Torgesen, Laughon, Simmons, & Rashotte, 1993.

[57] Moe, Hopkins, & Rush, 1982.

[58] Ball & Blachman, 1991; Byrne & Fielding-Barnsley, 1989; Cary & Verhaeghe, 1994; O'Connor & Jenkins, 1995.

[59] Ball & Blachman, 1991; Cunningham, 1990; O'Connor & Jenkins, 1995.

[60] McLoughlin & Lewis, 2001.

[61] Invernizzi, Robey, & Moon, 1999.

[62] Bear & Barone, 1998; Betts, 1946; Stauffer et al., 1978.

[63] Barr, Blachowicz, & Wogman-Sadow, 1995; Juel, 1988.

[64] Morris, 1999b.

[65] Bear, Invernizzi, Templeton, & Johnston, 2004.

[66] Invernizzi, Abouzeid, & Gill, 1994.

[67] West, Denton, & Reaney, 2001

[68] George & Mallery, 2001.

[69] Invernizzi et al., 1999.

[70] American Educational Research Association et al., 1999.

[71] Invernizzi, Robey, & Moon, 2000.

[72] Mehrens & Lehmann, 1987.

[73] Gronlund, 1985.

[74] National Reading Panel, 2000, pp. 3–8.

[75] McConkie & Zola, 1987.

[76] Schreiber, 1987.

[77] U.S. Department of Education, 1995.

[78] Moe et al., 1982.

[79] Gronlund, 1985.

[80] Invernizzi et al., 1999.

[81] Invernizzi et al., 2000.

[82] Huang, 2014.

[83] Zinbarg et al., 2005

[84] American Educational Research Association et al., 1999.

[85] Stanford-9, 1996.

[86] Metzoff, 1998.

[87] American Educational Research Association et al., 1999.

[88] Leslie & Caldwell, 1995.

[89] Beaver, 1997.

[90] Invernizzi & Meier, 2000.

[91] Dorans, 1989.

Section VIII

Appendix: Expansion to Grades 7 and 8

Table 1 Form A. Item level characteristics for word recognition in isolation task by grade level.

Grade 7 Word Lists		Difficulty	item-total *r*	Discrimination	DIF	
					Gender	Race/Ethn
1	chariot	.82	.41	.35	A	A
2	statesman	.78	.37	.36	A	A
3	indigestion	.39	.56	.67	A	A
4	temptation	.75	.40	.36	A	B-
5	typhoon	.68	.52	.62	B-	A
6	longitude	.49	.47	.62	A	A
7	mercury	.81	.41	.40	A	A
8	conviction	.66	.56	.68	A	A
9	abroad	.41	.48	.64	A	B-
10	periscope	.50	.55	.71	A	A
11	symphony	.70	.51	.57	B+	A
12	institution	.67	.54	.60	A	A
13	meteorite	.47	.59	.77	A	B+
14	desirable	.52	.56	.68	A	A
15	uncertainty	.38	.57	.65	A	A
16	adhesive	.35	.59	.70	A	A
17	peninsula	.83	.33	.26	A	A
18	mutiny	.55	.52	.65	A	A
19	industrious	.45	.55	.69	A	A
20	masculine	.36	.62	.75	A	A

Table 1 (Continued)

	Grade 8 Word Lists	Difficulty	item-total *r*	Discrimination	DIF	
					Gender	Race/Ethn
1	custody	.84	.43	.30	A	B-
2	wealthiest	.78	.39	.36	A	A
3	stratosphere	.69	.54	.59	A	A
4	ambassador	.73	.46	.48	A	B+
5	pathetic	.79	.41	.38	A	A
6	constable	.76	.22	.21	A	A
7	reliance	.79	.47	.42	A	A
8	assumption	.60	.54	.63	A	A
9	ferocity	.60	.51	.53	A	A
10	adrift	.89	.28	.19	A	A
11	substantial	.45	.35	.41	A	A
12	excavation	.43	.55	.66	A	A
13	convincingly	.44	.45	.50	A	A
14	miscellaneous	.31	.52	.57	A	A
15	accommodation	.56	.50	.55	A	A
16	prosperous	.53	.56	.68	A	A
17	juror	.73	.41	.39	A	B-
18	consolation	.72	.38	.38	A	A
19	brigade	.46	.52	.63	B-	A
20	haphazard	.27	.34	.34	A	A

Notes. Difficulty (p) indicates the proportion of respondents who correctly responded to the item. Item-total r (correlation) is also known as the point biserial correlation. Discrimination indices approximately .20 and above were considered adequate. For item-total correlations, r >.30 was considered adequate. DIF = differential item functioning. Focus group is female, White (+ favors the focus group). Reference group is male, nonWhite (- favors the reference group). DIF characteristics are based on ETS classifications: A= negligible, B = moderate, C = for investigation.

Table 2 Form B. Item level characteristics for word recognition in isolation task by grade level.

Grade 7 Word Lists		Difficulty	item-total *r*	Discrimination	DIF	
					Gender	Race/Ethn
1	antler	.92	.32	.17	A	A
2	frequency	.87	.39	.28	A	A
3	sanitary	.90	.35	.21	A	A
4	indigestion	.36	.56	.66	A	A
5	cavern	.81	.40	.33	A	B+
6	particle	.79	.38	.32	A	A
7	circulate	.69	.48	.50	A	A
8	publication	.70	.43	.47	A	A
9	cultural	.61	.55	.62	A	A
10	abroad	.40	.51	.59	A	A
11	deputy	.78	.36	.32	A	B+
12	complexion	.77	.46	.43	A	A
13	residence	.80	.45	.38	A	A
14	geologist	.64	.50	.56	A	A
15	uncertainty	.47	.46	.56	A	A
16	bristle	.69	.39	.39	A	A
17	circumference	.21	.45	.42	A	A
18	peninsula	.85	.34	.25	A	A
19	industrious	.44	.55	.67	A	B-
20	masculine	.34	.52	.60	A	A

Table 2 (Continued)

Grade 8 Word Lists		Difficulty	item-total *r*	Discrimination	DIF	
					Gender	Race/Ethn
1	habitation	.74	.47	.49	A	A
2	pacify	.47	.49	.62	A	A
3	violate	.85	.37	.29	A	A
4	persistence	.73	.40	.43	B+	A
5	ambassador	.74	.53	.56	A	A
6	quota	.77	.43	.45	A	B+
7	prosperity	.43	.49	.61	A	A
8	defendant	.69	.34	.38	A	A
9	revelation	.79	.38	.35	A	B-
10	assumption	.54	.53	.68	A	B+
11	superb	.61	.37	.45	A	A
12	serenity	.68	.48	.55	B+	A
13	counterfeit	.44	.38	.50	A	A
14	embassy	.32	.35	.42	A	A
15	optimism	.44	.43	.52	A	B+
16	dissatisfied	.39	.43	.47	A	A
17	juror	.75	.40	.41	A	A
18	consolation	.74	.39	.40	A	A
19	brigade	.33	.45	.54	B-	A
20	haphazard	.45	.32	.35	A	A

Notes. Difficulty (p) indicates the proportion of respondents who correctly responded to the item. Item-total r (correlation) is also known as the point biserial correlation. Discrimination indices approximately .20 and above were considered adequate. For item-total correlations, r >.30 was considered adequate. DIF = differential item functioning. Focus group is female, White (+ favors the focus group). Reference group is male, nonWhite (- favors the reference group). DIF characteristics are based on ETS classifications: A= negligible, B = moderate, C = for investigation. \

Table 3 Form A. Item level characteristics for spelling words (bonus point followed by feature score) by grade level.

	Word	Difficulty	item-total *r*	Discrimination	DIF	
					Gender	Race/Ethn
1	slide	.89	.57	.35	A	A
		.90	.54	.32	A	A
2	brave	.94	.45	.20	B+	B+
		.95	.43	.17	B+	A
3	drive	.93	.48	.23	A	A
		.95	.42	.15	A	A
4	shade	.87	.56	.40	A	A
		.89	.53	.34	A	A
5	float	.75	.68	.68	A	A
		.76	.67	.65	A	A
6	clean	.91	.58	.32	A	A
		.91	.57	.31	A	A
7	paint	.85	.68	.52	A	B-
		.86	.65	.47	A	B-
8	flight	.69	.72	.83	A	A
		.69	.72	.83	A	A
9	start	.92	.50	.24	A	A
		.95	.43	.16	A	A
10	hurt	.75	.64	.66	A	A
		.75	.63	.65	A	A
11	shelf	.81	.62	.55	A	B+
		.83	.62	.50	A	B+
12	cork	.60	.51	.63	A	B+
		.74	.48	.51	A	B+
13	caught	.46	.67	.86	A	A
		.47	.67	.86	A	A
14	grouch	.61	.67	.81	A	A
		.65	.64	.77	A	A
15	spoil	.71	.66	.71	A	A
		.74	.69	.72	A	A

Table 3 (Continued)

	Word	Difficulty	item-total *r*	Discrimination	DIF	
					Gender	Race/Ethn
16	stood	.64	.70	.83	A	A
		.64	.71	.83	B-	A
17	noises	.60	.69	.84	A	A
		.66	.65	.76	B-	A
18	traced	.57	.71	.92	A	A
		.65	.68	.85	A	A
19	posing	.63	.44	.52	A	A
		.66	.43	.49	A	A
20	striped	.62	.65	.80	B-	B+
		.70	.62	.70	A	A
21	lazily	.24	.52	.62	A	A
		.26	.51	.62	A	A
22	youthful	.45	.58	.76	B+	A
		.68	.50	.54	B+	A
23	misgivings	.30	.44	.52	A	A
		.49	.33	.44	A	B-
24	stiffness	.46	.62	.78	A	A
		.66	.63	.74	B+	A
25	simplicity	.16	.44	.46	A	A
		.34	.57	.75	A	A
26	resign	.25	.54	.68	A	B-
		.36	.55	.76	A	B-
27	divinity	.20	.46	.51	A	B+
		.42	.54	.73	A	A
28	omission	.23	.47	.52	B-	A
		.27	.49	.56	B-	A

Notes. Difficulty (p) indicates the proportion of respondents who correctly responded to the item. Item-total r (correlation) is also known as the point biserial correlation. Discrimination indices approximately .20 and above were considered adequate. For item-total correlations, r >.30 was considered adequate. DIF = differential item functioning. Focus group is female, White (+ favors the focus group). Reference group is male, nonWhite (- favors the reference group). DIF characteristics are based on ETS classifications: A= negligible, B = moderate, C = for investigation.

Table 4 Form B. Item level characteristics for spelling words (bonus point followed by feature score) by grade level.

	Word	Difficulty	item-total *r*	Discrimination	DIF	
					Gender	Race/Ethn
1	skate	.22	.48	.56	A	A
		.26	.50	.62	A	A
2	glide	.32	.50	.66	A	A
		.38	.45	.61	A	A
3	rope	.26	.48	.57	B+	A
		.28	.48	.59	B+	A
4	shape	.15	.36	.34	A	A
		.48	.41	.50	A	A
5	soap	.23	.45	.53	A	A
		.45	.41	.52	A	A
6	dream	.90	.53	.29	A	A
		.93	.52	.24	A	A
7	snail	.84	.54	.42	A	B-
		.85	.54	.41	A	B-
8	tight	.95	.42	.14	A	A
		.95	.42	.14	A	A
9	sharp	.94	.50	.18	A	A
		.95	.49	.17	A	A
10	silk	.83	.59	.48	A	A
		.83	.59	.47	A	A
11	thorn	.91	.59	.29	B-	A
		.92	.59	.27	A	A
12	burn	.87	.63	.40	A	A
		.88	.61	.37	A	A
13	bought	.72	.70	.76	A	A
		.72	.69	.75	A	A
14	voice	.89	.53	.32	A	A
		.92	.53	.25	B+	A
15	pool	.87	.60	.41	A	B-
		.90	.56	.33	A	B-

Table 4 (Continued)

	Word	Difficulty	item-total r	Discrimination	DIF	
					Gender	Race/Ethn
16	mouth	.82	.48	.39	C+	A
		.88	.50	.32	B+	B-
17	noises	.81	.68	.57	A	A
		.82	.68	.56	A	A
18	copies	.70	.69	.74	B-	A
		.71	.69	.72	B-	A
19	waving	.79	.66	.59	A	A
		.83	.63	.50	A	A
20	bullies	.89	.60	.34	B-	A
		.90	.60	.32	C-	A
21	lazily	.62	.60	.72	A	A
		.65	.59	.68	A	A
22	mistreatment	.59	.57	.69	B-	A
		.60	.56	.68	C-	A
23	stillness	.55	.61	.78	B-	A
		.56	.61	.78	B-	A
24	truthful	.25	.49	.60	A	A
		.27	.50	.62	A	B-
25	simplicity	.18	.46	.48	C+	A
		.23	.46	.53	C+	A
26	fasten	.45	.55	.73	A	B+
		.65	.36	.45	A	A
27	submission	.66	.59	.67	B-	B-
		.75	.60	.58	B-	A
28	serenity	.55	.52	.64	A	A
		.68	.40	.44	A	A

Notes. Difficulty (p) indicates the proportion of respondents who correctly responded to the item. Item-total r (correlation) is also known as the point biserial correlation. Discrimination indices approximately .20 and above were considered adequate. For item-total correlations, r >.30 was considered adequate. DIF = differential item functioning. Focus group is female, White (+ favors the focus group). Reference group is male, nonWhite (- favors the reference group). DIF characteristics are based on ETS classifications: A= negligible, B = moderate, C = for investigation.

Table 5 Text complexity and cohesion of oral reading passages

	Narrativity		Syntactic Simplicity		Word Concreteness		Referential Cohesion		Deep Cohesion	
	7th	8th	7th	8th	7th	8th	7th	8th	7th	8th
Form A	36	18	50	42	97	95	68	72	65	90
Form B	36	29	50	43	96	96	63	78	67	85

Note. All values are percentiles as reported by the Coh-Metrix Text Easability Assessor.

Table 6 Item level characteristics for comprehension questions by form and grade level

	Grade 7 Question	Difficulty	item-total *r*	Discrimination	DIF	
					Gender	Race/Ethn
Form A	1	.60	.45	.49	A	A
	2	.63	.48	.59	A	A
	3	.88	.42	.29	A	A
	4	.72	.57	.54	A	A
	5	.79	.48	.41	A	A
	6	.26	.49	.45	A	A
	Grade 8 Question	**Difficulty**	**item-total *r***	**Discrimination**	**DIF**	
					Gender	**Race/Ethn**
Form A	1	.85	.51	.35	A	A
	2	.77	.48	.39	A	A
	3	.74	.49	.51	A	A
	4	.72	.47	.41	A	A
	5	.44	.43	.46	B-	A
	6	.76	.49	.49	A	A
	Grade 7 Question	**Difficulty**	**item-total *r***	**Discrimination**	**DIF**	
					Gender	**Race/Ethn**
Form B	1	.53	.46	.49	A	A
	2	.86	.36	.24	A	A
	3	.70	.50	.52	A	A
	4	.64	.44	.44	B+	B-
	5	.69	.56	.59	A	A
	6	.77	.47	.45	A	A
	Grade 8 Question	**Difficulty**	**item-total *r***	**Discrimination**	**DIF**	
					Gender	**Race/Ethn**
Form B	1	.79	.46	.36	A	A
	2	.66	.50	.49	A	A
	3	.59	.53	.58	B-	A
	4	.46	.46	.47	A	A
	5	.80	.39	.37	A	A
	6	.76	.42	.38	A	A

Notes. Difficulty (p) indicates the proportion of respondents who correctly responded to the item. Item-total r (correlation) is also known as the point biserial correlation. Discrimination indices approximately .20 and above were considered adequate. For item-total correlations, r >.30 was considered adequate. DIF = differential item functioning. Focus group is female, White (+ favors the focus group). Reference group is male, non-White (- favors the reference group). DIF characteristics are based on ETS classifications: A= negligible, B = moderate, C = for investigation.