EXPLORATION OF THE CONSORTIUM ON READING EXCELLENCE PHONICS SURVEY

An Instrument for Assessing Primary-Grade Students' Phonics Knowledge

ABSTRACT

This article examines the Consortium on Reading Excellence–Phonics Survey (CORE-PS), an informal, inexpensive, and widely disseminated assessment tool that is used to determine primary-grade students' knowledge of and abilities to apply key alphabetic and phonics understandings to decode a mix of real and pseudo-words. Evidence is reported of the extent to which the CORE-PS meets the following psychometric criteria: test retest, internal consistency, and interrater reliability and face, content, construct, consequential, and criterion validity. Findings suggest that the CORE-PS provides an inexpensive, acceptably reliable and valid assessment of primary-grade students' decoding and reading phonics knowledge. Limitations for K–3 students on the alphabetic section of the CORE-PS are noted and discussed and future directions for research with the CORE-PS are presented.

D. Ray Reutzel
UTAH STATE UNIVERSITY

Lorilynn Brandt
UTAH VALLEY UNIVERSITY

Parker C. Fawson
UNIVERSITY OF KENTUCKY

Cindy D. Jones
UTAH STATE UNIVERSITY

THE preponderance of current empirical evidence has and continues to support the effectiveness of early phonics instruction for helping young readers succeed (Adams, 1990; Anderson, Hiebert, Scott, & Wilkinson, 1985; Bond & Dykstra, 1967; Camilli, Vargas, & Yurecko, 2003; Chall, 1967, 1983; National Institute of Child Health and Human Development [NICHD], 2000; National Institute for Literacy, 2008; Snow, Burns, & Griffin, 1998; Stuebing, Barth, Cirino, Francis, & Fletcher, 2008). The goal of beginning reading instruction is to help students move as quickly as possible toward comprehension of a broad range of complex and

THE ELEMENTARY SCHOOL JOURNAL VOLUME 115, NUMBER 1
 0013-5984/2014/11501-0003 $10.00

content-rich texts (National Governors Association Center for Best Practices, Council of Chief State School Officers, 2010). Phonics instruction is a gateway toward achieving that end because it helps students acquire the necessary prerequisite skills to decode unfamiliar words encountered in increasingly complex texts (Norman & Calfee, 2004). Findings of the National Reading Panel's meta-analysis concluded that systematic phonics instruction helps all children learn to read with greater success than nonsystematic or no phonics instruction (NICHD, 2000). Indeed, it is difficult if not impossible for young students to learn to read an alphabetic language without phonics knowledge and skill (Ehri, 2009; Pressley, 2006). The purpose of this study was to explore the validity and reliability of the scores obtained from administrations of the Consortium on Reading Excellence–Phonics Survey, 2nd edition (CORE, 2008).

The assessment of decoding automaticity using measures of accuracy and rate has become common practice in classrooms across the nation (Cummings, Dewey, Latimer, & Good, 2011; Deneey, 2010; Hudson, Torgesen, Lane, & Turner, 2012; Murray, Munger, & Clonan, 2012). Assessment of decoding automaticity is often accomplished by using curriculum-based measurement (CBM) of oral reading fluency (Deno, 1985). Many current scholars have criticized the exclusive use of oral reading fluency CBMs to monitor the progress of students' reading growth (Deneey, 2010; Kuhn, Schwanenflugel, & Meisinger, 2010; Murray et al., 2012). Consequently, researchers and reading scholars have called for the use of an expanded set of assessments for understanding students' growth in reading (Apel, 2011; Deneey, 2010; Hudson et al., 2012; Kuhn et al., 2010; Murray et al., 2012).

In a component-based reading assessment model (Rathvon, 2004), phonics knowledge is considered one of several important sublexical processes often measured by the use of oral reading fluency scores that are used to assess automatic word decoding (Morris et al., 2012; Rathvon, 2004; Stahl & McKenna, 2013). Consequently, teachers, coaches, and administrators seeking to troubleshoot low oral reading fluency scores have sought affordable and curriculum-sensitive assessments of students' phonics knowledge and skills (Rathvon, 2004; Stahl & McKenna, 2013).

Assessing Phonics Knowledge

Assessment of phonics knowledge has long been widely recognized as an important means for identifying readers who struggle (Ehri, Nunes, Stahl, & Willows, 2001). Currently, available phonics assessments evidence several theoretical and practical shortcomings for the purposes of helping teachers identify students' mastery of and/or specific gaps in students' phonics knowledge (Stahl & McKenna, 2013).

First, many informal phonics assessments and surveys currently available for classroom use, such as the Informal Phonics Inventory (McKenna & Stahl, 2009), the Names Test (Cunningham, 1990; Mather, Sammons, & Schwartz, 2006), or the Z test (McKenna & Stahl, 2009), do not reflect current theoretical understandings of the components involved in processing orthographic information. Apel (2011) divides orthographic knowledge into two components: sight words (mental graphic representations) and knowledge of spelling patterns stored in memory. In theoretical and practical terms, this definition means that currently available classroom phonics assessments need to measure students' abilities to read both real and pseudo-words

(Apel, 2010). Currently available informal phonics assessments and surveys do not consistently do so.

Second, currently available phonics assessments are often not sequenced to assess developmental progressions of how students acquire phonics knowledge over time to support word recognition (Apel, 2011; Rathvon, 2004; Stahl & McKenna, 2013). Stahl and McKenna (2013) state that a good phonics assessment "will provide teachers with the knowledge of which skills have been mastered, which require review or consolidation, and which call for explicit instruction" (p. 21). Ehri (1987, 1992, 1995, 2005, 2009) describes the development of word recognition using a four-stage model. During the first or *pre-alphabetic* stage, early readers evidence little awareness of letters and phonemes in print. In the second or *partial alphabetic* stage, young students begin to understand how selected letters and sounds relate. In the third or *full alphabetic* stage, students largely master the content of alphabetic knowledge. As students initially move into the full alphabetic stage, they apply their mastery of alphabetic knowledge to phonologically recode the sounds and letters in unfamiliar words. Phonological recoding is considered the central accomplishment at the full alphabetic stage (Gough & Hillinger, 1980). In the fourth stage, the *consolidated alphabetic* stage, students acquire knowledge of larger orthographic units, spelling patterns, or word chunks that reoccur within and across words and use this knowledge to recognize words by analogizing. Organized developmentally, classroom-administered phonics assessments would provide teachers with insights into students' progress along the word-recognition development continuum.

Third, cost-to-benefit is important to consider when selecting assessments for introduction into and use in the school infrastructure (McBride, Ysseldyke, Milone, & Stickney, 2010; Snow & Van Hemel, 2008). For example, norm-referenced, standardized decoding tests such as the Woodcock Reading Mastery Test, 3rd edition (Woodcock 2011) are often quite expensive to purchase and require substantial training and time to administer, making them less accessible and affordable for classroom teachers. Another norm-referenced test, the Test of Word Reading Efficiency (TOWRE), charges in excess of several hundred dollars to purchase the assessments and necessary ancillary supplies plus the ongoing cost of purchasing students' test recording forms at approximating $1.50 per student. The cost of acquiring and continuously using assessment instruments such as these, as psychometrically sound as they are, often compels cash strapped teachers and schools to look elsewhere for inexpensive, informal assessments.

Fourth, the results of administering norm-referenced, standardized reading achievement test batteries are often reported to teachers as total test or subtest aggregate scores. Reporting aggregate scores or interpolated scores in summary fashion to teachers provides few insights into individual student response profiles tied to item content, thus limiting the potential of these tests to inform teachers' instruction or grouping decisions.

Fifth, there is often some disparity between the content (scope and sequence) of the typical phonics curriculum taught in classrooms and the content of decoding assessment items found on norm-referenced reading-achievement tests (Stahl & McKenna, 2013). Stahl, Duffy-Hester, and Doughterty-Stahl (2006) explain that the typical sequence of early phonics instruction begins with teaching young students to recognize and name the letters of the alphabet and associate these letters with individual sounds or phonemes heard in spoken words (alphabetic principle). Next,

students are taught to apply their knowledge of single letter names and letter sounds to spellings of letter sounds represented by more than one letter (e.g., ck, ch, ea, ai) and simple single-syllable, within-word spelling patterns such as consonant-vowel-consonant (cvc) and vowel-consonant-e (vce) orthographic patterns.

Eventually, early phonics instruction advances to teaching young students to decode multiple-syllable words using syllabic and affix spelling patterns. In this more advanced stage of phonics instruction in elementary schools, students learn to recognize words through analogizing. Without a close articulation between the content of reading achievement decoding test items and the decoding curriculum (scope and sequence) taught in the classroom, norm-referenced achievement tests have limited value for assessing the decoding curriculum taught in elementary classrooms and for helping teachers make necessary decisions about the placement of students into intervention groups or the content of the decoding instruction to be provided in these groups (Stahl & McKenna, 2013).

Finally, and perhaps most importantly, many phonics assessments accessible to classroom teachers lack evidence for score reliability and validity. Much of the current focus on the quality of beginning reading assessments, including decoding assessments, has been stimulated by the No Child Left Behind Act (2001) and other legislative requirements that reading assessments provide evidence of technical adequacy. The National Association for the Education of Young Children (NAEYC, 2003) position statement on the testing of young children states that policy makers and educators should use valid and reliable assessments to make judgments about young students' reading progress.

Apel (2011) echoes the need to develop reliable and valid measures of young readers' orthographic knowledge development and processing. For example, to assess students' orthographic knowledge in spelling, informal criterion-referenced approaches such as the Primary and Elementary Qualitative Spelling Inventories have been recommended as one of the most useful, valid, and reliable informal classroom assessments available (Bear, Invernizzi, Templeton, & Johnston, 2008; Sterbinsky, 2007). Qualitative spelling assessments, however, require students to produce (encode) rather than recognize (decode) orthographic units. Thus, valid and reliable decoding measures, that assess recognition rather than production tasks, are greatly needed. Apel (2011) asserts, "Not only will researchers benefit from well-developed and accepted measures of orthographic knowledge, but practitioners will, too . . . practitioners will need assessment tools that allow them to determine whether their students are struggling with orthographic knowledge and then help them plan their instruction or interventions accordingly" (p. 599).

Because of the shortcomings previously discussed here, school administrators, coaches, teachers, teacher educators, and even researchers have sought alternative phonics assessments. One such assessment that has grown in popularity and use in schools and teacher education programs across the nation is the Consortium on Reading Excellence–Phonics Survey, 2nd edition (CORE, 2008).

Description of the CORE Phonics Survey (CORE-PS)

The CORE Phonics Survey, 2nd edition is a criterion-referenced (CR) mastery measure available in English and Spanish (CORE, 2008) (Stahl & McKenna, 2013). The content of criterion-referenced assessments typically aligns well with grade-level cur-

riculum (Stahl & McKenna, 2013). The CORE Phonics Survey manual states, "The *CORE phonics survey* . . . assesses the phonics and phonics-related skills that have a high rate of application in beginning reading" (CORE, 1999, p. 63). A mastery measure assesses discrete skills that are "useful when it is important to monitor a skill that is taught in isolation or while troubleshooting a particular area that is giving a student difficulty" (Stahl & McKenna, 2013, p. 6). The CORE Phonics Survey manual states, "This test is a mastery test. It is expected that students will ultimately get all items correct" (CORE, 1999, p. 64). Developers also assert that the organization and content of the CORE Phonics Survey allow teachers to inspect student item-level responses to inform phonics instructional decision making and can be used as a tool for placing students into targeted decoding intervention groups (CORE, 2008).

The CORE-PS, 1st edition (CORE, 1999) originally consisted of 12 subtests; the twelfth subtest assessed students' phonics knowledge through spelling. In the second edition, published in 2008, the spelling subtest was dropped, limiting the CORE-PS (CORE, 2008) to an assessment of students' phonics knowledge through the use of recognition not production tasks. The CORE-PS, 2nd edition (CORE, 2008) also consists of a series of 12 subtests (A–L) addressing alphabetic knowledge and reading and decoding components. The 12 subtests of the CORE-PS, 2nd edition are described in Table 1.

In the alphabetic knowledge section of the CORE-PS, 2nd edition there are four subtests, A–D, with a single item for each subtest and a total of four scored items for all four tests. In the reading and decoding section of the CORE-PS, 2nd edition, there are eight subtests, E–L, with a total of 30 scored items.

The CORE-PS requires roughly 10–15 minutes administration time per student. Although the CORE-PS can be used in grades K–8 to assess phonics knowledge mastery, practically and developmentally it is used most frequently and appropriately in grades K–3. The purpose of the CORE-PS, according to its developers, is to monitor students' acquisition of reading phonics knowledge to a level of mastery.

Items are scored correct or incorrect and totaled ($N = 34$) using the CORE-PS Record Form for each subtest. No ceiling or basal score information is available for the CORE-PS, 2nd edition. The administration guide provides a matrix suggesting when each subtest should be administered during the school year (fall, winter, spring) for grades K–3 and up.

Need for Analysis of Psychometric Properties of CORE-PS

A recent Internet search of Google, MSN, and Yahoo yielded over 200,000 items that specifically referenced the use of the CORE-PS. Many of these Internet sites were school or district webpages that actively promote the use of the CORE-PS. Other CORE-PS Internet hits were college or university webpages where preservice teacher candidates and in-service graduate students are trained to use the CORE-PS as a part of their teacher preparation program and/or as a part of their postgraduate professional development. The CORE-PS was also widely used in many past federally funded Reading First state projects. In addition, several prominent nationally published reading teacher education textbooks such as *Teaching Children to Read: The Teacher Makes the Difference*, 6th edition (Reutzel & Cooter, 2012) or *Early Reading Assessment* (Rathvon, 2004) have recommended the use of the CORE-PS as a tool for assessing students' decoding abilities. Why has the CORE-PS become so widely ad-

Table 1. The CORE-PS (2nd ed.) Components, Dimensions, and Tasks

Components and Dimensions	Tasks
Alphabet knowledge:	
Subtest A: Uppercase letter recognition	Point to a matrix of uppercase alphabet letters; students say name of letters ($N = 1$)
Subtest B: Lowercase letter recognition	Point to a matrix of lowercase alphabet letters; students say name of letters ($N = 1$)
Subtest C: Consonant sounds identification	Point to a line of consonant letters and ask students to say the sound associated with letter to which the examiner points ($N = 1$)
Subtest D: Long and short vowel sounds identification	Point to a line of vowel letters and ask students to say the sound associated with a letter to which the examiner points ($N = 2$ [long] [short])
Reading and decoding knowledge:	
Subtest E: Consonant-vowel-consonant (cvc) single-syllable word spelling pattern	Blend letter sounds to pronounce real and pseudo-words ($N = 3$) (examples: sip, mat, nop, dit)
Subtest F: Single-syllable words with consonant blends and short vowel spelling patterns	Blend letter sounds to pronounce real and pseudo-words ($N = 3$) (examples: stop, trap, nask, dilt)
Subtest G: Single-syllable words with consonant digraphs/tri-graphs and short vowel spelling patterns	Blend letter sounds to pronounce real and pseudo-words ($N = 3$) (examples: chop, match, shom, phid)
Subtest H: Single-syllable words with "r" controlled vowel spelling patterns	Blend letter sounds to pronounce real and pseudo-words ($N = 3$) (examples: form, bird, gorf, murd)
Subtest I: Single-syllable words with long vowel sounds	Blend letter sounds to pronounce real and pseudo-words ($N = 3$) (examples: tape, paid, hine, soat)
Subtest J: Single-syllable words with variant vowel sound spellings	Blend letter sounds to pronounce real and pseudo-words ($N = 3$) (examples: moon, hawk, fout, zoy)
Subtest K: Single-syllable words with low frequency vowel and consonant spellings	Blend letter sounds to pronounce real and pseudo-words ($N = 3$) (examples: kneel, cent, dimb, wrep)
Subtest L: Multiple-syllable words	Blend letter sounds to accurately pronounce real and pseudo-words ($N = 8$) (examples: consent, admire, menu, railways, timbut, morkle, fauntoon)
Total all subtest items	$N = 34$

opted by teacher education programs, schools, and teachers to assess students' reading decoding abilities? The obvious answer seems to be that the CORE-PS addresses many of the shortcomings previously discussed with currently available and accessible phonics tests. An analysis of the CORE-PS, 2nd edition's content, structure, and characteristics reveals why and how this popular decoding assessment tool effectively addresses many of the previously discussed theoretical, practical, and economic concerns about currently available phonics assessment instruments.

The "CORE Handbook of Assessing Reading: Multiple Measures for Kindergarten through Eighth" (CORE, 2008) in which the CORE-PS, 2nd edition is published was carefully consulted to determine what, if any, reliability or validity evidence is available or reported. Finding nothing in these sources, we placed telephone calls to the Consortium on Reading Excellence (CORE), and then to the publisher, Area Press, in Novato, CA. Both sources, CORE and the publisher, confirmed they had no data on the validity and reliability of the CORE-PS and could not provide any technical information for this assessment that is being used to assess knowledge, to monitor mastery, and to design instruction for literally thousands of students.

Next, we conducted a comprehensive review of the literature related to the CORE-PS, 1st and 2nd editions, to determine if research had been conducted and reported

on their validity or reliability. No evidence was located of the validity or the reliability of CORE-PS scores.

Purpose of This Study

There are several justifications for conducting this exploratory study of the validity and reliability of the CORE-PS, 2nd edition's scores. First, the CORE-PS, 2nd edition is widely used nationally in schools and universities to assess and make daily judgments about the status and progress of students' decoding knowledge. Second, the CORE-PS provides a theoretically grounded assessment of the two definitional components of orthographic knowledge: (1) using real words to assess specific mental representations of words represented in long-term memory, for example, sight words; and (2) using pseudo-words to assess individuals' knowledge of spelling patterns and rules that facilitate the more laborious processes of word analysis (Apel, 2011; Cunningham, Nathan, & Raher, 2011).

To guide the conduct of this study, we used an argument-based approach described by Kane (1992, 2006) that stresses the importance of making inferences from test results that are based on evidence outlined in interpretive arguments. There are two major components to Kane's approach: the interpretive argument and the validity argument (Kane, 2006). The interpretive argument focuses on inferences and assumptions leading to statements and decisions that can be made from assessment results. The validity argument evaluates the interpretive argument as a whole, and the inferences and assumptions in the interpretive argument specifically using appropriate evidence (Cronbach, 1988). Below is the set of arguments around which our validation study of the CORE-PS, 2nd edition was based: (A_0) The CORE-PS should be grounded in descriptions of school-based phonics curricula. The scale should be useful for their assessment. (A_1) The CORE-PS should be consistent with current research and theory on the development of students' phonological and orthographic knowledge. The scale should be useful for their assessment. (A_2) The CORE-PS should help teachers determine students' mastery of phonics content knowledge. (A_3) The CORE-PS should provide teachers with valuable information for making decisions about phonics instruction. (A_4) CORE-PS should demonstrate score stability. (A_5) CORE-PS should demonstrate internal consistency. (A_6) Multiple raters should rate students' phonics knowledge similarly when using the CORE-PS. (A_7) CORE-PS items should demonstrate reasonably close fit with expert opinion and national standards related to phonics instruction. (A_8) CORE-PS items should demonstrate adequate fit with the construct of phonics. (A_9) Because phonics knowledge is an integral component of decoding automaticity, the CORE-PS should allow teachers to investigate underlying phonics issues related to students' inadequate oral reading fluency. (A_{10}) Because phonics knowledge is one of several components associated with decoding automaticity, CORE-PS scores should be moderately to strongly associated with measures of reading fluency.

Research Questions

The research questions for this measurement study of the CORE-PS, 2nd edition were as follows: (1) What is the evidence for the CORE-PS, 2nd edition's test-retest, internal consistency, and interrater reliability? (2) What is the evidence for the

Table 2. CORE-PS Study Sample Demographics

Demographic Category	Total	K	1	2	3
N	592	35	206	252	99
Race:					
African American	7	0	2	2	3
Asian	7	1	1	3	2
Caucasian	314	19	110	135	50
Hispanic	226	14	79	97	36
Native American	7	0	3	2	2
Pacific Islander	31	1	11	13	6
Gender:					
Female	302	18	105	129	50
Male	290	17	101	123	49
Low SES	290	17	102	124	47
English language learner	80	5	28	34	13
Special services	47	3	16	20	8

CORE-PS, 2nd edition's face, content, construct, consequential, and concurrent validity with oral reading fluency scores?

Method

Sample

CORE-PS, 2nd edition (CORE, 2008) scores were obtained from a convenience sample of 592 K–3 elementary students in two western U.S. school districts and four elementary schools. CORE-PS scores were drawn from the primary grades (K–3) where classroom phonics instruction is recommended as evidence-based practice (NICHD, 2000; NIFL, 2008). The demographic characteristics of the sample with regard to gender, ethnicity, socioeconomic status, English language learner classification, and students receiving special services are reported in Table 2. An additional 170 K–3 students' scores for the test-retest and 129 K–3 students' scores for the criterion validity analysis were randomly selected by grade-level strata from two other demographically similar schools in one of the two school districts from which the original sample of 592 K–3 students' was drawn.

Reliability Study Procedures

Test-retest reliability. To investigate the reliability of the CORE- PS, we first explored test-retest and internal consistency estimates of reliability. Test-retest reliability study data were collected from a random sample of 170 K–3 students in two schools by a group of nine undergraduate and two graduate students who were trained in a university literacy clinic by a member of the research team. The CORE-PS was administered twice to these 170 grade K–3 students, separated by 2 weeks in time. Test-retest descriptive statistics and Pearson's *r* coefficients were calculated using SPSS v. 21 for Mac (SPSS, 2011).

Internal consistency reliability. Internal consistency reliability study data were collected from 592 grade K–3 students in four local elementary schools that were using the CORE-PS, 2nd edition. Literacy coaches in the schools already described collected internal consistency reliability study data and were trained by a member of

the research team. Internal consistency descriptive statistics and Cronbach's alpha coefficients were calculated using SPSS v. 21 for Mac (SPSS, 2011).

Interrater reliability. To examine the reliability of raters' scoring of the CORE-PS, 25 students, grades 1–3 (1st = 9, 2nd = 8, 3rd = 8), were randomly selected and were administered the CORE-PS twice and videotaped. Two raters, trained literacy coaches, independently viewed and scored these 25 students' two videotaped CORE-PS testing-retesting occasions for a G theory analysis. Using SPSS v. 21 for Mac (SPSS, 2011), a fully crossed, two-facet rater by occasion *G* theory study was used to assess interrater reliability, and a *D* study was used to assess the facets of rater and measurement occasion to reduce error variance. A Cohen's kappa estimate for rater pairs was also reported.

Validity Study Procedures

Face validity. To explore the face validity of the CORE-PS, expert opinion of three nationally recognized phonics instruction authorities was solicited. Each of these experts had published nationally disseminated books on phonics instruction and assessment. Expert evaluations of the face validity of the CORE-PS were analyzed for statements about strengths and weaknesses by the research team.

Content validity. To explore the content validity of the CORE-PS, two members of the research team made comparisons between each of the phonics concepts assessed by the CORE-PS and those listed in the Reading Foundational Skills-Phoneme-Grapheme Correspondences recommendations of the Common Core State Standards (NGA & CCSSO, 2010). The Common Core reading foundation standards were developed using evidence based on the report of the National Reading Panel (NICHD, 2000) and *Preventing Reading Difficulties in Young Children* (Snow et al., 1998). In addition, international performance standards on the *Progress in International Literacy Study* (PIRLS) along with "school to work" program standards were used as the basis for constructing the Common Core State Standards (NGA & CCSSO, 2010). Content validity of the CORE-PS was examined through a content-analysis approach (Neuendorf, 2002) in which researchers compared the assessment content of the CORE-PS with the Reading Foundational Skill-Phoneme-Grapheme Correspondences recommendations of the Common Core State Standards (NGA & CCSSO, 2010).

Construct validity. To examine the hypothesized component structure of the CORE-PS (1) alphabet knowledge, and (2) reading and decoding knowledge, we conducted a confirmatory factor analysis (CFA) with data from 592 grade K–3 students using M*Plus*, v. 7 for Mac (Muthén & Muthén, 2011).

Consequential validity. To explore the consequential validity of the CORE-PS, 2nd edition, we used a structured interview containing the following six questions: Which phonics assessments do you know about and use? How did you come to learn about the CORE Phonics Survey? Why do you use the CORE Phonics Survey? How do you use the results of the CORE Phonics Survey? If you were to no longer use the CORE Phonics Survey, what would be the consequences for your students? and, If you were to no longer use the CORE Phonics Survey, what would be the consequences for you as a teacher? Four randomly selected K–3 teachers and one literacy coach from two schools, wherein we had also collected the test-retest and criterion-validity data, met in a focus group with one of the members of the research team.

Consequential validity focus group member answers to structured interview questions were coded and analyzed by one research team member using NVivo 7 and examined for accuracy of coding and themes by one other team member (NVivo, 2006).

Criterion validity. Finally, 129 randomly selected K–3 students' CORE-PS scores in two schools were collected to examine the criterion validity of students' words correct per minute (wcpm) scores using the DIBELS Nonsense Word Fluency (NWF) in kindergarten and DIBELS Oral Reading Fluency (ORF) grades 1–3 (K = 34, 1st = 32, 2nd = 30, 3rd = 33) with the CORE-PS. The DIBELS NWF tests students' abilities to identify and blend sounds in simple three-letter pseudo-words. The NWF test was scored only for words blended or recoded in 1 minute (words correct per minute) in this study rather than counting the number of sounds identified in each pseudo-word. The DIBELS ORF tests students reading accuracy and speed (words correct per minute) when reading a grade-level passage for 1 minute. A trained cohort of DIBELS test administrators used in the school districts administered all DIBELS tests. The DIBELS NWF test shows a range of reliabilities of .79 in kindergarten to .83 in first grade using alternate forms. The DIBELS ORF test shows reliability coefficients using alternate forms of .94 (Good & Kaminski, 2002). Pearson's *r* coefficients were calculated using SPSS v. 21 for Mac to examine the criterion validity question (SPSS, 2011).

Results

The purpose of this study was to explore the reliability and validity of the CORE-PS, 2nd edition. Results are reported in two major sections: (1) reliability, and (2) validity. We begin by reporting the results of the reliability study because reliability is a necessary, but insufficient, condition to establish validity (Mislevy, 2004).

Reliability Results

Test-retest reliability. A Pearson's *r* correlation test-retest coefficient was calculated using 170 students' CORE-PS scores given 2 weeks apart. The obtained Pearson *r* test-retest correlation coefficient for the total CORE-PS scores across grade levels was .98. Test-retest Pearson *r* correlation coefficients by grade level were: K = .95, 1st = .91, 2nd = .94, and 3rd = .95.

Internal consistency reliability. To determine the internal consistency of the CORE-PS, a Cronbach's alpha coefficient was calculated for each of the 12 subtests within the two major CORE-PS sections: (1) alphabetic knowledge (subtests A–D) and (2) reading and decoding knowledge (subtests E–L). Descriptive statistics and Cronbach's alpha coefficients for each of the 12 CORE-PS subtests are found in Tables 3 and 4. Note that subtests A–D were collapsed into a single score for the alphabet knowledge section for reliability analysis since each of these four subtests contained only a single item response (see Table 1).

A Cronbach's alpha of .70 or higher is considered acceptable (Reynolds, Livingston, & Willson, 2009). Internal consistency Cronbach's alpha coefficients ranged from a low of .64 for alphabetic knowledge and skills (subtests A–D) to a high of .97 for multisyllabic words (subtest L). Grade level total CORE PS alphas ranged from a low of .95 for kindergarten to a high of .98 for second grade.

Table 3. Descriptive Statistics for the 12 Subtests of the CORE-PS, 2nd edition

CORE-PS Subtest	Min	Mean	*SD*	Max
Kindergarten:				
A–D	38.0	73.1	9.5	83.0
E	.0	8.7	6.0	15.0
F	.0	5.8	5.3	15.0
G	.0	4.2	4.6	14.0
H	.0	3.3	4.4	15.0
I	.0	2.4	4.0	15.0
J	.0	2.2	3.8	14.0
K	.0	1.2	1.2	13.0
L	.0	1.7	1.7	23.0
Grade 1:				
A–D	55.0	81.9	3.3	85.0
E	3.0	14.0	1.8	15.0
F	.0	11.9	2.7	15.0
G	.0	11.8	3.1	15.0
H	1.0	10.2	4.4	15.0
I	.0	10.6	4.5	15.0
J	.0	9.2	3.9	15.0
K	.0	6.4	4.8	15.0
L	.0	9.2	7.8	24.0
Grade 2:				
A–D	64.0	82.4	3.2	85.0
E	1.0	14.0	2.3	15.0
F	.0	12.7	3.3	15.0
G	.0	12.6	3.9	15.0
H	.0	12.2	4.2	15.0
I	.0	12.2	4.7	15.0
J	.0	11.4	4.4	15.0
K	.0	9.7	5.3	15.0
L	.0	14.3	8.9	24.0
Grade 3:				
A–D	53.0	81.1	4.3	85.0
E	1.0	13.5	2.3	15.0
F	.0	11.6	3.3	15.0
G	1.0	11.6	3.8	15.0
H	1.0	11.2	4.4	15.0
I	.0	11.3	4.8	15.0
J	.0	9.9	4.3	15.0
K	.0	7.9	5.3	15.0
L	.0	11.3	8.8	24.0

Interrater reliability. *G* theory and *D* study analyses were used to determine test reliability by studying various factors or facets that contribute to error variance (Brennan, 1983; Grimm & Yarnold, 2000). Our interest in using *G* theory analysis was to determine interrater reliability between two trained raters and maximize score reliability and minimize score error variance for the CORE-PS. A *G* or phi coefficient of .30–.49 is considered weak, .50–.79 or higher is considered acceptable, and .80 or above is considered strong (Cohen, 1988). Results showed that *G* and phi reliability coefficients for the alphabetic knowledge section (subtests A–D) of the CORE-PS were acceptable, $G = .73$, $\varphi = .73$. On the other hand, *G* and phi reliability coefficients for the reading and decoding knowledge section (subtests E–L) of the CORE-PS were strong, $G = .96$, $\varphi = .95$ (Cohen, 1988).

A *D* study was conducted as a follow-up to a *G* study to determine ways in which one could minimize error variance while simultaneously optimizing overall instru-

Table 4. Cronbach's Alpha Coefficients for CORE-PS by Subtest and Grade Level

	Title of Subtest	Cronbach's Alpha
Survey subtest:		
Subtests A–D	Alphabetic Knowledge	.64
Subtests E–L	Reading and Decoding	.94
Subtests A–L	Total	.92
Decoding subtests (multiple items):		
Subtest E	Consonant Vowel Consonant (CVC) Spelling Pattern	.88
Subtest F	Blends with CVC	.86
Subtest G	Digraphs, Trigraphs	.91
Subtest H	R-Controlled Vowel	.92
Subtest I	Long Vowel Spellings	.94
Subtest J	Variant Vowel Spellings	.92
Subtest K	Low Frequency Spellings	.94
Subtest L	Multi-Syllable Words	.97
Grade-level analysis:		
Grade K	Total CORE PS score	.95
Grade 1	Total CORE PS score	.96
Grade 2	Total CORE PS score	.98
Grade 3	Total CORE PS score	.96
Grades K–3	Total CORE PS score	.98

ment reliability. Figure 1 depicts the results of changing the number of raters ($n = 2$) on the y-axis and the number of testing occasions ($N = 2$) as represented by solid versus dotted lines on the resulting generalizability reliability coefficients shown on the x-axis for the alphabetic knowledge section of the CORE-PS. Figure 2 depicts the results of changing the number of raters ($N = 2$) on the x-axis and the number of

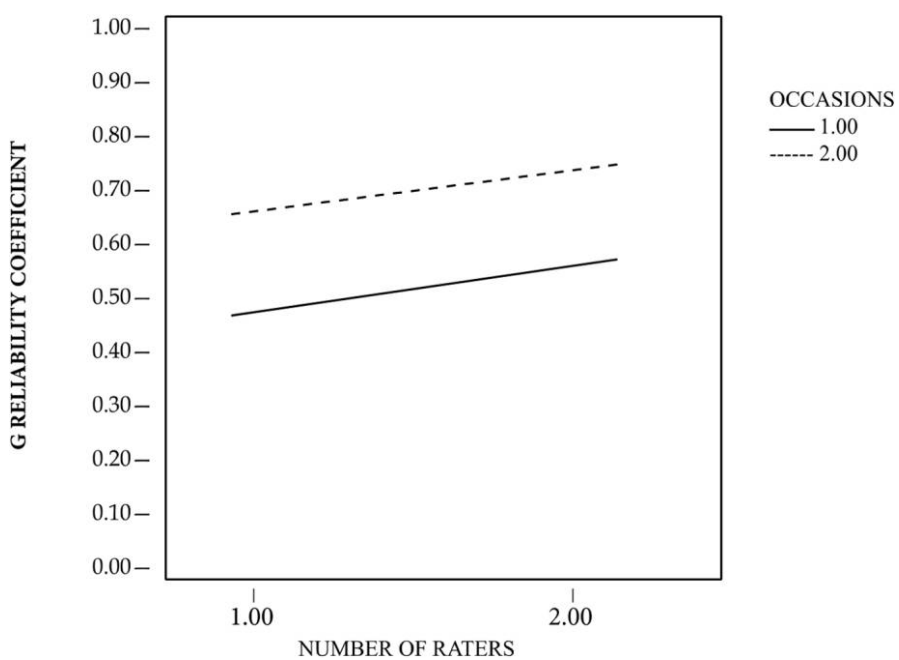

Figure 1. *D* study effects of changing CORE-PS raters or occasions on the Alphabetic Knowledge section scores.

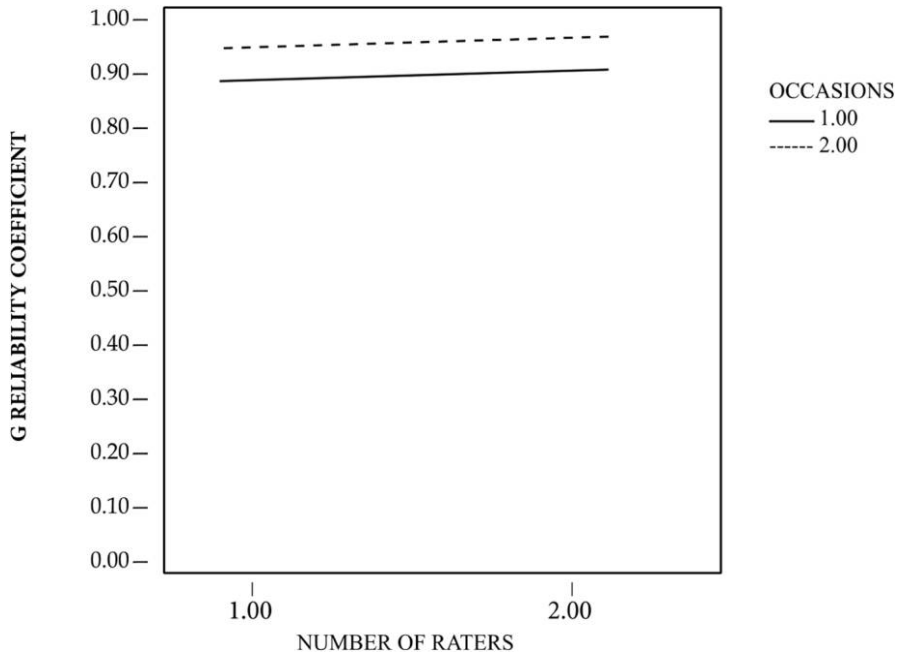

Figure 2. *D* study effects of changing Core-PS raters or occasions on the Reading and Decoding Knowledge section scores.

testing occasions ($N = 2$) as represented by solid versus dotted lines on the resulting generalizability reliability coefficients shown on the y-axis for the reading and decoding section of the CORE-PS. The *D* study (see Figs. 1 and 2) showed that increasing the number of testing occasions from one to two occasions reduced error variance and increased score reliability more than an increase in the number of raters.

A Cohen's kappa for the rater pair that scored the 25 students' CORE-PS in this study was .85—a very strong indicator of interrater agreement. A desirable level of kappa for research purposes should be between .60 and .70, and Banerjee, Capozzoli, McSweeney, and Sinha (1999) report that achieving a Cohen's kappa of .75+ "indicates excellent agreement beyond chance" (Neuendorf, 2002, p. 143).

Validity Results

In this section, we report the results of five separate estimates of the CORE-PS's validity: face, content, construct, consequential, and criterion-related.

Face validity. Three phonics experts in the field of reading were asked to examine the CORE-PS for face validity. Two of the three responded to our request. Both experts indicated that overall the CORE-PS was an adequate measure of students' phonics knowledge as defined by nationally recognized and published phonics advocates. Expert 1 noted that the CORE-PS failed to assess students' knowledge of common phonograms, for example, *ight, ick*, or *all*, and also failed to assess students' knowledge of contractions.

Expert 2 indicated that the CORE-PS was "not an extensive diagnostic assessment, but it does represent the most important instructional elements involved in phonics." Expert 2 also commented that "all pseudo-words used in the test should, in my opinion, represent the structure of real words. A pseudo-word such as 'nik' does not

Table 5. Multiple Goodness-of-Fit Tests for a Confirmatory Factor Analysis of the CORE-PS

Test	Fit
Chi-square	1,688.51*
Comparative fit index (CFI)	.944
Tucker-Lewis index (TFI)	.939
Residual mean square error of approximation (RMSEA)	
Estimate	.065
90% C.I.	.061–.068
Probability RMSEA < .05	.01
Standardized root mean square residual (SRMR)	.051

Note.— $df = 486$.
* $p < .01$.

represent the structure of real words. The CORE Phonics Survey has done a very good job with their pseudo-words. The only questionable pseudo-words would be: 'loe' and 'rew.'"

Content validity. The content of the CORE-PS assessment items were compared with the phonics instruction objectives found in the Reading Foundational Skills-Phoneme-Grapheme Correspondences (RFS-PGC) section of the Common Core State Standards (NGA & CCSSO, 2010). The content assessed in the subtests of the CORE-PS was found to agree 97% of the time with the CCSS RFS-PGC. This analysis revealed only two areas out of 69 possible categories where the CORE-PS, 2nd edition failed to agree with the CCSS RFG-PGC: selected silent consonant letter combinations and doublets.

Construct validity. The CORE-PS is designed with two major subsections: (1) alphabetic knowledge and (2) reading and decoding knowledge. To examine the hypothesized two-component factor structure, alphabetic knowledge and reading and decoding knowledge, we conducted a CFA using M*plus* v.7 for Mac. The four subtests, A–D, which each contained only a single scored item, were each modeled onto the alphabet knowledge and skills section of the CORE-PS using a single-level CFA. The eight subtests, E–L, each containing between 3 and 8 scored items, were initially modeled onto each subtest (E–L) as factors and then onto a second-order latent construct factor—reading and decoding. The results of five multiple goodness-of-fit tests and standardized factor loadings are found in Tables 5 and 6.

It is now generally accepted that researchers should not report a single goodness-of-fit estimate to determine adequate model fit (Bentler, 1990; Hooper, Coughland, & Mullen, 2008; Kline, 2010; Loehlin, 2004; Schumaker & Lomax, 2010; Thompson, 1998). In order to avoid this known problem, we report five goodness-of-fit indices: chi-square, the Bentler (1990) comparative fit index (CFI), Tucker-Lewis index (TLI), the root mean square residual (RMSEA), and the standardized root mean square residual (SRMR).

The chi-square (χ^2) test of model fit in this study indicated poor model fit ($p < .01$). Kline (2010) claims that a failed χ^2 test of model fit should always be reported. There exists, however, continuing controversy around whether the χ^2 test of model fit leads to erroneous conclusions because as sample size increases, particularly in sample sizes that exceed 200, this statistic tends to indicate a significant probability level (Hooper et al., 2008; Loehlin, 2004; Schumaker & Lomax, 2010). It appears that the most prudent approach for determining model fit is the one we have taken here:

Table 6. Standardization Factor Loadings

	Two-Tailed			
	Estimate	*SE*	Estimate/*SE*	*p*-Value
Alphabetic Knowledge Section by:				
Item 1	.720	.031	22.926	.000
Item 2	.612	.034	17.748	.000
Item 3	.563	.036	15.439	.000
Item 4	.369	.055	6.771	.000
Item 5	.364	.063	5.782	.000
Subtest E by:				
Item 6	.827	.016	50.882	.000
Item 7	.877	.014	63.995	.000
Item 8	.844	.015	54.958	.000
Subtest F by:				
Item 9	.808	.017	47.858	.000
Item 10	.832	.016	53.502	.000
Item 11	.821	.016	50.674	.000
Subtest G by:				
Item 12	.850	.013	65.838	.000
Item 13	.930	.008	116.685	.000
Item 14	.868	.012	73.501	.000
Subtest H by:				
Item 15	.889	.010	87.949	.000
Item 16	.928	.008	121.358	.000
Item 17	.873	.011	78.048	.000
Subtest I by:				
Item 18	.906	.009	106.009	.000
Item 19	.937	.006	144.255	.000
Item 20	.917	.008	117.692	.000
Subtest J by:				
Item 21	.900	.009	99.942	.000
Item 22	.903	.009	102.033	.000
Item 23	.879	.010	84.375	.000
Subtest K by:				
Item 24	.949	.006	158.378	.000
Item 25	.891	.010	91.070	.000
Item 26	.905	.009	102.662	.000
Subtest L by:				
Item 27	.889	.009	95.024	.000
Item 28	.808	.015	54.184	.000
Item 29	.875	.010	85.055	.000
Item 30	.913	.008	119.976	.000
Item 31	.885	.010	92.499	.000
Item 32	.922	.007	133.408	.000
Item 33	.902	.008	107.551	.000
Item 34	.897	.009	103.048	.000
Decode by:				
Subtest E	.748	.022	34.443	.000
Subtest F	.911	.012	75.457	.000
Subtest G	.928	.009	105.546	.000
Subtest H	.946	.007	131.131	.000
Subtest I	.948	.007	144.544	.000
Subtest J	.983	.005	196.460	.000
Subtest K	.931	.008	118.674	.000
Subtest L	.877	.011	78.998	.000
Reading Decoding Section:				
With Alphabetic Section	.625	.035	17.837	.000

to rely on multiple tests of model fit rather than on a single measure (Hooper et al., 2008; Kline, 2010; Schumaker & Lomax, 2010).

The confirmatory fit index indicated an acceptable goodness-of-fit estimate of .94. The criterion for a good model fit is when CFI values exceed .90 and approach .95 (Hooper et al., 2008; Schumaker & Lomax, 2010; Stevens, 1996). The Tucker-Lewis index (TLI) was .94, also indicating acceptable model fit (TLI $>$ .95). The root mean square error of approximation likewise was consulted as a determinant of model fit. The criterion for acceptable model fit to the data for RMSEA are values less than .08 (Hooper et al., 2008; Schumaker & Lomax, 2010). The RMSEA was calculated as .065, also indicating an acceptable model fit for the data. Finally, the standardized root mean square residual (SRMR) is the standardized difference between the observed covariance and the predicted covariance. A value of zero indicates a perfect fit. This measure tends to be smaller as sample size increases and as the number of parameters in the model increases. A value that is equal to or less than .05 is considered an acceptable fit. The SRMR for this model was .051, indicating an unacceptable fit (Schumaker & Lomax, 2010).

A rule of thumb for deciding which of the several goodness-of-fit statistics to report and how to choose cut-off values for declaring an acceptable model fit have been discussed repeatedly in the statistical literature (Hooper et al., 2008; Hu & Bentler, 1999; Kline, 2010; Schumaker & Lomax, 2010). When RMSEA values are close to .08 or below, SRMR values are less than.05, and CFI and TLI coefficients are greater than .90, a model evidences an acceptable fit (Hooper et al., 2008; Hu & Bentler, 1999; Schumaker & Lomax, 2010). Therefore, if our reported RMSEA (.065), CFI (.94), and TLI (.94) are viewed through this interpretive lens, these findings taken together argue for acceptable model fit of the CORE-PS, 2nd edition's hypothesized two-component factor structure: (1) alphabetic knowledge and skills and (2) reading and decoding of the CORE-PS. On the other hand, the rejection of the CORE-PS's model fit as tested by the χ^2 estimate and the SRMR argues for a degree of caution or humility in concluding "good" rather than acceptable model fit.

Consequential validity. To explore the consequential validity of the CORE-PS, 2nd edition, we used six structured interview questions with five K–3 teachers and one literacy coach randomly selected from two schools that participated in the study. The teachers and coach interviewed indicated that they had learned about the CORE-PS through district in-service programs and from coaches or state office of education sponsored Reading First summer seminars. All the teachers and coaches indicated that they knew about the CORE-PS, 2nd edition as one of two measures they used, the other being a running record school-based informal measure, to investigate or troubleshoot students' low oral reading fluency scores. When asked how they used the results, teachers answered that they used the CORE-PS scores and subtest items to pinpoint students' "phonics knowledge gaps" and target small intervention groups to "fill these gaps." Consequently, teachers and coaches use the CORE-PS to make instructional decisions for intervention group phonics lesson content and for grouping students who have similar instructional needs. Finally, when asked what would happen to them or their students if they were no longer permitted to use the CORE-PS, the K–3 teachers and the coach interviewed indicated that they would lose very useful data for knowing how to help students who struggle with decoding, fluency, and other related reading components.

Criterion validity. As indicated in the consequential validity data, educators often use the CORE-PS, 2nd edition to troubleshoot underlying problems with students' oral reading fluency scores. We calculated a Pearson's *r* correlation between 129 CORE-PS, 2nd edition total scores and primary grade students' total wcpm scores on the DIBELS NWF (kindergarten) and ORF (grades 1–3) scores. Results indicated adequate correlations by grade levels, above .60 (Cronbach, 1990), all of which were significant at the $p < .01$ level: K = .66, 1st = .78, 2nd = .86, 3rd = .67, and total (K–3) of .84. Given the CORE-PS's two-factor structure as indicated earlier, we also calculated correlations between CORE-PS, 2nd edition subtest 1 (alphabetic knowledge) scores and subtest 2 (reading decoding) scores and students' wcpm scores on the DIBELS NWF (kindergarten) and ORF (grades 1–3) scores. Results showed that subtest 1 (alphabetic knowledge) evidenced generally weak criterion-related validity as indicated in the following grade-level coefficients: K = .37, 1st = −.02, 2nd = .13, 3rd = .37, and total (K–3) = .42. Subtest 2 (reading decoding), on the other hand, evidenced acceptable criterion-related validity coefficients by grade level: K = .69, 1st = .79, 2nd = .86, 3rd = .67, and total (K–3) = .86.

Discussion

The purpose of this study was to explore the reliability and validity of the CORE-PS, 2nd edition. We begin with a discussion of the reliability analyses, followed by a discussion of the validity analysis.

Reliability Analyses

Research question 1 focused on exploring evidence for the reliability of the CORE-PS, 2nd edition. The CORE-PS, 2nd edition scores remained stable over short periods of time between testing across grade levels. This result was expected in that CORE-PS scores should remain stable over short periods of time since word recognition develops incrementally over several years of elementary school (Apel, 2011; Ehri, 2009; Templeton & Bear, 1992).

Estimates of internal consistency showed a high degree of interitem correlation within subtests E–L, but not on subtests A–D. The less than acceptable alpha associated with CORE-PS subtests A–D raises serious questions about the use of subtests A–D. An inspection of the means and standard deviations in Table 3 demonstrated that the students tested had already mastered the alphabetic knowledge component leading to ceiling effects. With little variability, interitem correlations for subscales A–D were suppressed. Further, the use of a single scored item for each subtest A–D also suppressed potential variation leading to less than acceptable reliability.

In view of these findings, CORE-PS users should exercise caution when interpreting scores obtained from administrations of the CORE-PS subtests A–D, especially if the students assessed have already mastered the phonics knowledge measured by these subtests. We recommend that subtests A–D may not be an appropriate measure from midyear kindergarten or for students who have already mastered the content. Furthermore, we recommend that the CORE-PS alphabetic knowledge section, subtests A–D, be revised to include the use of more than a single item score per subtest. Given recent research on letter-name recognition, a variety of item sets could be designed to probe letter-name knowledge using student names, the letter-name

pronunciation effect, letter frequency, etc. (Jones & Reutzel, 2012; Justice, Pence, Bowles, & Wiggins, 2006; Piasta & Wagner, 2010).

The *G* study asked the question, to what degree do raters and rating occasions contribute to score variance on the CORE-PS, 2nd edition? Results from the *G* study and the Cohen's kappa statistics demonstrated that when different raters scored the CORE-PS, the results showed very strong agreement. Thus, when the CORE-PS is administered and scored by two different trained raters (teachers), the resultant scores will be nearly identical. This finding also demonstrates the fact that the relatively brief amount of training needed to prepare teachers, coaches, and interventionists to reliably administer and score the CORE-PS, 2nd edition is sufficient.

Nearly 50% of the variance in student scores on Section 1, alphabetic knowledge, of the CORE-PS was due to true variance in students' knowledge of the alphabetic principle. There was a clear restriction in range (ceiling effect) for the scores in the alphabetic knowledge section with this sample population. Low score variability often leads to an underestimation of the variance. Had the sample of scores evidenced greater variability, say by including younger children in the sample who were still developing their knowledge of the alphabetic principle or the inclusion of more items per subscale, the amount of variance accounted for may have increased, leading to improved estimates of internal consistency on subtests A–D (Apel, 2011; Ehri, 1987; Templeton & Bear, 1992). Because of the constrained nature of alphabetic knowledge to be learned by young children (Paris, 2005) and its early development, this finding argues for administering the alphabetic knowledge section of the CORE-PS to a younger sample population such as preschool and very early kindergarten children where alphabet knowledge would be expected to vary more dramatically than with a midyear kindergarten to third-grade population. Given the developmental nature of word recognition in young children, the findings from Section 2, reading and decoding knowledge, of the CORE-PS seemed to reflect true differences in students' development of orthographic knowledge (Apel, 2011; Ehri, 1987; Templeton & Bear, 1992).

The single facet in the *D* study that contributed the most to error reduction and increased score reliability was testing occasion. Thus increasing the number of testing occasions from one to two increased the reliability of the CORE-PS for making absolute decisions about student performance more than increasing the number of raters. Final phi coefficients, obtained from the *D* study, indicated that with two raters and two testing occasions overall error variance decreased and reliability of CORE-PS alphabetic knowledge section scores increased to acceptable levels.

From this finding, it is clear that, if administered, Section 1, alphabetic knowledge, of the CORE-PS should be given on two occasions with two raters to attain adequate levels of reliability. Although administering the alphabetic knowledge section on two occasions with two raters increased generalizability to within acceptable levels, we recommend that scores obtained from administration of subtests A–D in particular be interpreted with caution for students in late kindergarten and beyond who may have already mastered the phonics knowledge assessed by these subtests and because of the extremely limited number of items leading to a restriction in range. We stop short of recommending that the alphabetic knowledge section of the CORE-PS be dropped or ignored by practitioners or researchers, but would suggest that future editions of the CORE-PS attend to the concerns expressed here to improve this instrument's subtest A–D reliability. Adding another testing occasion or rater also

increased the coefficients for the CORE-PS Section 2, reading and decoding knowledge, but not enough to warrant adding another testing occasion or rater to obtain highly reliable scores.

Validity Analyses

The face validity of the CORE-PS was explored by obtaining expert opinion about its content and structure. Taken together, the reported evaluation of the CORE-PS by two nationally recognized phonics instruction and assessment experts suggests that the CORE-PS adequately assessed expected areas of orthographic knowledge and orthographic pattern recognition needed by students to develop into successful early readers (Apel, 2011; Cunningham et al., 2011; Ehri, 1987, 2009; Templeton & Bear, 1992).

The content validity of the CORE-PS was explored by assessing the degree of overlap between CORE-PS items and the Reading Foundational Skills-Phoneme-Grapheme Correspondences section of the Common Core State Standards (NGA & CCSSO, 2010). This analysis demonstrated a high degree of match between the content of the CORE-PS, 2nd edition and national standards. Consequently, practitioners will likely find that the CORE-PS, 2nd edition provides useful data for making informed decisions about targeted, evidence-based, and differentiated decoding instruction and student placement into intervention groups for such decoding instruction.

The third validity research question focused on evidence for the construct validity of the CORE-PS. The authors of the CORE-PS constructed a phonics assessment composed of two discrete sections or components: (1) alphabetic knowledge and (2) reading and decoding knowledge. Results of the CFA reported in this study generally supported a two-factor structure with three of five goodness-of-fit test statistics meeting or exceeding recommended levels. Again, we note the failed chi-square and SRMR tests of model fit in this study and urge caution in concluding that the construct validity of the current CORE-PS, 2nd edition as currently designed is anything more than acceptable. On the other hand, it is one that might be improved with the revisions previously suggested. The second section of the CORE-PS, 2nd edition, reading and decoding, evidences very strong factor loadings, indicating that teachers and researchers can use the subtests (E–L) to determine where students' reading and decoding knowledge is well developed and where it has not yet developed to a level of mastery. The consequential validity interviews indicated that among teachers who use the CORE-PS, 2nd edition, the use has been primarily to troubleshoot low oral reading fluency scores to provide targeted phonics instruction in order to increase students' oral reading fluency. The concurrent validity estimate between the total CORE-PS scores and DIBELS ORF scores demonstrated that between 44% and 71% of the variance in oral reading fluency may be accounted for by students' phonics knowledge on the CORE-PS. On the other hand, the analysis of CORE-PS subtest scores with DIBELS presented a different picture, with extremely low criterion validity coefficients with subtest 1, alphabetic knowledge, and DIBELS wcpm scores. Taken together, these findings seem to support the continued use of the CORE-PS for the purposes teachers and coaches were using it, with the cautions already expressed surrounding use and interpretation of subtest 1.

In general, the CORE-PS, 2nd edition is a reasonably valid assessment of students' phonics knowledge as adjudicated by experts, after a comparison to national phonics instructional standards and from results of a confirmatory factor analyses. Teacher and coach answers to interview questions and estimates of criterion validity also support this conclusion, with the cautions previously noted.

Limitations

This exploratory study of the CORE-PS, 2nd edition was limited in several ways. First, participants were not selected randomly from the total target population; therefore, findings cannot be generalized to all schools and student CORE-PS scores. The schools selected were accessible to the researchers and were selected because they were using the CORE-PS, 2nd edition (CORE, 2008) instead of the previous CORE-PS, 1st edition (CORE, 1999). Although schools were not randomly selected, consideration was given to selecting schools that evidenced similar variability in student SES and measures of adequate yearly progress.

Second, the number of expert opinions that were solicited for this study was limited to those of only two experts on phonics instruction. Although the feedback from these two experts was valuable, it would have strengthened the evidence for face validity of the CORE-PS, 2nd edition if additional expert opinions had been obtained.

Implications for Future Research and Instruction

The extensive use the CORE-PS, 2nd edition in schools and universities as a criterion-referenced, mastery measurement of phonics knowledge argued strongly for an exploratory study examining its validity and reliability. Future research should employ a sample of younger students—pre-K and early kindergarten—to further explore the reasons behind the less than acceptable item intercorrelations and unacceptable criterion-related validity coefficients obtained for subtests A–D.

The phonics knowledge content assessed by the CORE-PS, 2nd edition closely paralleled phonics objectives found in the nationally disseminated Reading Foundational Skills-Phoneme-Grapheme Correspondences recommendations of the Common Core State Standards (NGA & CCSSO, 2010). Future editions of the CORE-PS should be modified to include assessment of phonics elements identified as missing from the current edition in this study. As indicated in informal discussions with the two members of the rater pair who assessed interrater reliability, a written pronunciation guide for scoring pseudo-words would assist teachers and researchers and we suspect would all but eliminate rater variance in scoring the reading and decoding section.

The potential of the CORE-PS, 2nd edition as a developmental indicator of student progress through word recognition stages could be further refined by conducting a longitudinal study of young children's simultaneous reading and spelling development. Such a study would increase user confidence that CORE-PS subtest scores, especially subtests E–L, could be used to provide finely grained information about students' orthographic knowledge growth across the developmental stages of word recognition in reading (Ehri, 2009; NIFL, 2008; Piasta & Wagner, 2010).

In summary, the CORE-PS, 2nd edition is a reasonably reliable instrument as per arguments A_4–A_6 for assessing students' phonics knowledge (Kane, 1992, 2006), with the exception of subtests A–D in the alphabetic knowledge section. Referring back to Kane's (1992, 2006) argument structure, the CORE-PS, 2nd edition was also generally supported by expert opinion and aligned well with the Reading Foundations phonics concepts specified in the Common Core Standards. The validity evidence indicated a wide variety of support using arguments A_7–A_{10} for the general construction, fit, use, and ability of the CORE-PS scores for the uses teachers and coaches indicated, with the cautions and limitations previously noted (Kane, 1992, 2006). Several distinctive features of the CORE-PS, 2nd edition, especially in the reading and decoding section (subtests E–L), allow teachers and researchers access to theoretically grounded assessment of young students' development of orthographic knowledge and word recognition as per argument A_1 (Kane, 1992, 2006). For example, the CORE-PS, 2nd edition, subtest 2, assesses two components of orthographic knowledge: (1) specific mental representations of written words (real words) represented in long-term memory, for example, sight words, amalgams, and mental graphic representations (MGRs); and (2) individuals' knowledge of spelling patterns and rules that operate with a given language's orthographic system using real and pseudo-word recognition tasks (Apel, 2011; Cunningham et al., 2011). By analyzing student subtest 2 items and scores using real words versus pseudo-words on the CORE-PS, teachers and researchers can examine to what degree young students are developing these dual components of orthographic knowledge. Furthermore, the two sections of the CORE-PS, 2nd edition, alphabetic knowledge (subtests A–D) and reading and decoding knowledge (subtests E–L), are arranged in such a way as to facilitate analysis by researchers and practitioners alike of young students' progress through stages of increasingly complex orthographic pattern recognition (Ehri, 1987, 1992, 1995, 2005, 2009; Templeton & Bear, 1992) as per arguments A_2 and A_3 (Kane, 1992, 2006). Finally, by carefully examining students' performance on the CORE-PS, 2nd edition subtests A–L, researchers and practitioners can determine the appropriate content of phonics instruction needed and the assignment of students into small phonics intervention groups to receive targeted phonics instruction that is closely associated with improvements in decoding automaticity as measured by assessments of oral reading fluency.

References

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

Anderson, R. C., Hiebert, E. H., Scott, J. A., & Wilkinson, I. A. G. (1985). *Becoming a nation of readers: The report of the Commission on Reading*. Washington, DC: National Institute of Education.

Apel, K. (2010). Kindergarten children's initial spoken and written word learning in a storybook context. *Scientific Studies in Reading*, **14**(5), 440–463.

Apel, K. (2011). What is orthographic knowledge? *Language, Speech, and Hearing Services in Schools*, **42**(4), 592–603.

Banerjee, M., Capozzoli, M., McSweeney, L., & Sinha, D. (1999). Beyond kappa: A review of interrater agreement measures. *Canadian Journal of Statistics*, **27**(1), 3–23.

Bear, D. R., Invernizzi, M., Templeton, S., & Johnston, F. (2008). Primary and Elementary Qualitative Spelling Inventories. In D. R. Bear, M. Invernizzi, S. Templeton, & F. Johnston, *Words*

Their Way: Word study for phonics, vocabulary, and spelling instruction. (4th ed.). Boston: Pearson Education.
Bentler, P. M. (1990). Comparative fit indexes in structural models. *Psychological Bulletin*, **10**(2), 238–246.
Bond, G. L., & Dykstra, R. (1967). The cooperative research program on first-grade reading instruction. *Reading Research Quarterly*, **2**(4), 5–142.
Brennan, R. L. (1983). *Elements of generalizability theory*. Iowa City: American College Testing Program.
Camilli, G., Vargas, S., & Yurecko, M. (2003). Teaching children to read: The fragile link between science and federal policy. *Education Policy Analysis Archives*, **11**(15), 1–52.
Chall J. (1967). *Learning to read: The great debate*. New York: McGraw-Hill.
Chall, J. S. (1983). *Stages of reading development*. New York: Harcourt Brace College.
Cohen, J. (1988). *Statistical power analysis of the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.
Consortium on Reading Excellence. (1999). *Assessing reading: Multiple measures for kindergarten through eighth grade* (pp. 63–80). Novato, CA: Arena Press.
Consortium on Reading Excellence. (2008). Core phonics surveys. In B. Honing, L. Diamond, & R. Nathan (Eds.), *Assessing reading: Multiple measures for kindergarten through eighth grade* (2nd ed., pp. 63–80). Novato, CA: Arena.
Cronbach, L. J. (1988). Five perspectives on validity argument. In H. Wainer & H. Braun (Eds.), *Test validity* (pp. 3–17). Hillsdale, NJ: Erlbaum.
Cronbach, L. 1990. *Essentials of psychological testing*. New York: Harper & Row.
Cummings, K. D., Dewey, E. N., Latimer, R. J., & Good, R. H. (2011). Pathways to word reading and decoding: The roles of automaticity and accuracy. *School Psychology Review*, **40**(2), 284–295.
Cunningham, E., Nathan, R. G., & Raher, K. S. (2011). Orthographic processing in models of word recognition. In M. Kamil, P. Pearson, E. Moje, & P. Afflerbach (Eds.), *The handbook of reading research* (Vol. **4**, pp. 259–284). New York: Routledge.
Cunningham, P. (1990). The Names Test: A quick assessment of decoding ability. *Reading Teacher*, **44**(2), 124–129.
Deneey, T. A. (2010). One-minute fluency measures: Mixed messages in assessment and instruction. *Reading Teacher*, **63**(6), 440–450.
Deno, S. L. (1985). Curriculum-based measurement: The emerging alternative. *Exceptional Children*, **52**, 219–232.
Ehri, L. C. (1987). Learning to read and spell words. *Journal of Reading Behavior*, **19**(1), 5–31.
Ehri, L. C. (1992). Re-conceptualizing the development of sight word reading and its relationship to recoding. In P. B. Gough, L. C. Ehri, & R. Treiman (Eds.), *Reading acquisition* (pp. 107–143). Hillsdale, NJ: Erlbaum.
Ehri, L. C. (1995). Phases of development in learning to read words by sight. *Journal of Research in Reading*, **18**(1), 116–125.
Ehri, L. C. (2005). Development of sight word reading: Phases and findings. In M. Snowling & C. Hulme (Eds.), *The science of reading: A handbook* (pp. 135–154). Oxford: Blackwell.
Ehri, L. C. (2009). Learning to read in English: Teaching phonics to beginning readers from diverse backgrounds. In L. M. Morrow, R. Rueda, & D. Lapp (Eds.), *Handbook of research on literacy and diversity* (pp. 292–319). New York: Guilford.
Ehri, L. C., Nunes, S., Stahl, S., & Willows, D. (2001). Systematic phonics instruction helps students learn to read: Evidence from the National Reading Panel's meta-analysis. *Review of Educational Research*, **71**(3), 393–447.
Good, R. H., & Kaminski, R. A. (Eds.). (2002). *Dynamic indicators of basic early literacy skills* (6th ed.). Eugene, OR: Institute for Development of Educational Achievement.
Gough, P. B., & Hillinger, M. L. (1980). Learning to read: An unnatural act. *Bulletin of the Orton Society*, **30**, 180–196.
Grimm, L. G., & Yarnold, P. R. (2000). *Reading and understanding more multivariate statistics*. Washington, DC: American Psychological Association.
Hooper, D., Coughland, C., & Mullen, M. R. (2008). Structural equation modeling: Guidelines for determining model fit. *Electronic Journal of Business Research Methods*, **6**(1), 53–60.

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling*, **6**(1), 1–55.

Hudson, R. F., Torgesen, J. K., Lane, H. B., & Turner, S. J. (2012). Relations among reading skills and sub-skills and text-level reading proficiency in developing readers. *Reading Writing Quarterly*, **25**, 483–507.

Jones, C. D., & Reutzel, D. R. (2012). Enhanced alphabet knowledge instruction: Exploring a change of frequency, focus, and distributed cycles of review. *Reading Psychology: An International Quarterly*, **33**(5), 448–464.

Justice, L. M., Pence, K., Bowles, R. B., & Wiggins, A. (2006). An investigation of four hypotheses concerning the order by which 4-year-old children learn the alphabet letters. *Early Childhood Research Quarterly*, **21**, 374–389.

Kane, M. T. (1992). An argument-based approach to validity. *Psychological Bulletin*, **112**(3), 527–535.

Kane, M. T. (2006). Validation. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 17–64). Westport, CT: American Council on Education/Praeger.

Kline, R. B. (2010). *Principles and practices of structural equation modeling* (3rd ed.). New York: Guilford.

Kuhn, M. R., Schwanenflugel, P. J., & Meisinger, E. B. (2010). Aligning theory and assessment of reading fluency: Automaticity, prosody, and definitions of fluency. *Reading Research Quarterly*, **45**(2), 230–251.

Loehlin, J. C. (2004). *Latent variable models: An introduction to factor, path, and structural equation analysis* (4th ed.). Mahwah, NJ: Erlbaum.

Mather, N., Sammons, J., & Schwartz, J. (2006). Adaptations of the names test: Easy-to-use phonics assessments. *Reading Teacher*, **60**(2), 114–122.

McBride, J. R., Ysseldyke, J., Milone, M., & Stickney, E. (2010). Technical adequacy and cost benefit of four measures of early literacy. *Canadian Journal of School Psychology*, **25**, 189–204.

McKenna, M. C., & Stahl, K. A. D. (2009). *Assessment for reading instruction* (2nd ed.). New York: Guilford.

Mislevy, R. J. (2004). Can there be reliability without "reliability"? *Journal of Educational and Behavioral Statistics*, **29**(2), 241–244.

Morris, D., Trathen, W., Frye, E., Kucan, L., Ward, D., Schlagal, R., & Hendrix, M. (2012). The role of reading rate in the informal assessment of reading ability. *Literacy Research and Instruction*, **52**(1), 52–64.

Murray, M. S., Munger, K. A., & Clonan, S. M. (2012). Assessment as a strategy to increase oral reading fluency. *Intervention in School and Clinic*, **47**(3), 144–151.

Muthén, L. K., & Muthén, B. O. (2011). *Mplus: Statistical analysis with latent variables, v. 7*. Los Angeles: Author.

National Association for the Education of Young Children. (2003). *NAEYC joint position statement: Early childhood curriculum, assessment and program evaluation; building an effective, accountable system in programs for children birth through age 8*. Retrieved January 23, 2012, from http://www.naeyc.org/files/naeyc/file/positions/CAPEexpand.pdf

National Governors Association Center for Best Practices, Council of Chief State School Officers. (2010). *Common Core State Standards*. Retrieved October 27, 2011, from http://www.corestandards.org/

National Institute of Child Health and Human Development. (2000). *Teaching children to read: An evidence-based assessment of the scientific research literature on reading and its implications for reading instruction* (NIH Publication No. 00-4769O). Washington, DC: U.S. Government Printing Office.

National Institute for Literacy. (2008). *Developing early literacy: Report of the National Early Literacy Panel*. Jessup, MD: ED Publications.

Neuendorf, K. A. (2002). *The content analysis guidebook*. New York: Guilford.

No Child Left Behind Act of 2001, PL 107-110, 115 Stat. 1425, 20 U.S.C. § 6301 *et seq*.

Norman, K. A., & Calfee, R. C. (2004). Tile test: A hands-on approach for assessing phonics in the early grades. *Reading Teacher*, **58**(1), 42–52.

NVivo qualitative data analysis software (Version 7) [Computer software]. (2006). QSR International Pty Ltd.

Paris, S. G. (2005). Reinterpreting the development of reading skills. *Reading Research Quarterly*, **40**(2), 184–202.

Piasta, S. B., & Wagner, R. K. (2010). Developing early literacy skills: A meta-analysis of alphabet learning and instruction. *Reading Research Quarterly*, **45**(1), 8–38.

Pressley, M. (2006). *Reading instruction that works: The case for balanced teaching*. New York: Guilford.

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

Reutzel, D. R., & Cooter, R. B. (2012). *Teaching children to read: The teacher makes the difference* (6th ed.). Boston: Pearson Education.

Reynolds, C. R., Livingston, R. B., & Willson, V. (2009). *Measurement and assessment in education* (2nd ed.). Columbus, OH: Merrill/Prentice-Hall.

Schumaker, R. E., & Lomax, R. G. (2010). *A beginner's guide to structural equation modeling* (3rd ed.). New York: Routledge.

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

Snow, C. E., & Van Hemel, S. B. (Eds.). (2008). *Early childhood assessment: Why, what, and how*. Washington, DC: National Academy Press.

SPSS Inc. (2011). *SPSS Base 20.0 for MAC user's guide*. Chicago: Author.

Stahl, K. A. D., & McKenna, M. C. (2013). *Reading assessment in an RTI framework*. New York: Guilford.

Stahl, S. A., Duffy-Hester, A. M., & Dougherty-Stahl, K. A. (2006). Everything you wanted to know about phonics (but were afraid to ask). In K. Stahl & M. McKenna (Eds.), *Reading research at work: Foundations of effective practice* (pp. 126–154). New York: Guilford.

Sterbinsky, A. (2007). *Words Their Way spelling inventories: Reliability and validity analyses*. Memphis, TN: Center for Research in Educational Policy. Retrieved May 10, 2012, from http://timberridge.typepad.com/WTWReport.pdf

Stevens, J. (1996). *Applied multivariate statistics for the social sciences* (3rd ed.). Mahwah, NJ: Erlbaum.

Stuebing, K. K., Barth, A. E., Cirino, P. T., Francis, D. J., & Fletcher, J. M. (2008). A response to recent re-analyses of the National Reading Panel Report: Effects of systematic phonics instruction are practically significant. *Journal of Educational Psychology*, **100**(1), 123–134.

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

Thompson, B. (1998). Statistical significance testing and effect size reporting: Portrait of possible future. *Research in the Schools*, **5**(2), 33–38.

Woodcock, R. W. (2011). *Woodcock Reading Mastery Tests* (3rd ed.). San Antonio, TX: PsychCorp.