CHAPTER 1

INTRODUCTION

Universal preliteracy screening is an integral component of comprehensive academic programs for young children. Screening is important because it is used to identify students who are at risk for developing reading problems so that intervention efforts can start early and potentially prevent academic failure (Good, Gruba, & Kaminski, 2002). Screening and prevention efforts are particularly important in early elementary school because when students fall behind in reading during primary grades, many do not catch up to their peers, putting them at risk for a host of long-term academic problems (Juel, 1988; Stanovich, 1986). Moreover, the 2004 reauthorization of the Individuals with Disabilities Education Improvement Act (P.L. 108-446) allows for an approach to the identification of learning disabilities called Responsiveness to Intervention (RTI), which relies on effective universal screening to help identify students who may require additional instructional support or special education to meet academic standards (Fuchs, Fuchs, & Zumeta, 2008). For these reasons, screening should be considered a fundamental component of early literacy programs, and there is a need for measures that accurately identify students who require supplemental support so educators can intervene and hopefully prevent later difficulty. Given that screening is intended for all students, it is also important that measures be efficient and easy for teachers to administer.

At the same time, there are many challenges associated with assessment of young children (Vloedengraven & Verhoven, 2007), which can make development of accurate, technically adequate screening tools difficult. In this section, we provide information on early literacy screening, with emphasis on sublexical skills, particularly phonemic awareness. Next, we discuss the role of fluency in phoneme segmentation screening, a measure often used to assess phonemic awareness, highlighting one popular fluency-based screening tool that has been used in over 40 states (Manzo, 2005), the Phoneme Segmentation Fluency (PSF) subtest of the Dynamic Indicators of Basic Early Literacy Skills (DIBELS). We also describe challenges related to PSF screening and potential problems with current recommended scoring procedures. Finally, we discuss the promise of dynamic assessment (DA) to enhance validity and diagnostic accuracy of phonemic awareness screening. Based this discussion, we provide a rationale for the present study.

Why Consider Phonemic Awareness?

Phonemic awareness, a component of phonological awareness, is awareness that language comprises sounds, including words, syllables, and phonemes (Adams, 1990; Ehri, Nunes, Willows, Schuster, Yaghoub-Zadeh, & Shanahan, 2001). The phoneme is the smallest unit of sound in spoken language. Thus, phonemic awareness refers specifically to the knowledge and ability to manipulate individual sounds in spoken words (Adams; Ehri et al., 2001; Liberman, Shankweiler, Fischer, & Carter, 1974; Yopp, 1988). Phonemic awareness is an important precursor of students' understanding of the alphabetic principle because aside from simply recognizing letters, children must also understand that letters are associated with specific sounds before successful decoding

skill can develop (Adams). And, phonological and more specifically phonemic awareness have been shown to be important predictors of reading skill (NRP, 2000; Adams, 1990; Ehri et al., 2001).

Bradley and Bryant (1983) were the first researchers to demonstrate a causal relation between phonological knowledge and reading skill when they provided intensive instruction in sound categorization to 65 children, aged 4 to 5, and then monitored their reading and spelling development, along with 303 of their peers over the next 4 and 5 years. They found moderate significant relations between kindergarten sound categorization skill and reading and spelling skill in the students who were not intervened with more than 3 years later. They also found that students who received instruction in sound categorization significantly outperformed controls on measures of reading and spelling after 3 years.

The National Reading Panel's 2000 meta-analysis on the role of phonemic awareness instruction in reading acquisition corroborated Bradley and Bryant's (1983) earlier work (NRP, 2000; Ehri et al., 2001). Across studies, the panel found the average effect size of phonemic awareness training on reading and spelling outcomes immediately and over time ranged from .53 to .86. Further, they noted phonemic awareness was the single largest predictor of letter knowledge in young children, another important predictor of reading skill (NRP, 2001; Ehri et al., 2001).

Given the significance of phonemic awareness, it is often used as a predictor of reading development. Several skills are encompassed under the umbrella of phonemic awareness, and researchers have attempted to identify tasks that most accurately represent the range of skills. Authors of the NRP meta-analysis (2000; also see Ehri et al., 2001)

noted six distinct phonemic awareness skills: isolation, identification, categorization, blending, segmentation, and deletion. Sound isolation refers to the ability to recognize individual sounds in words, such as that the first sound in *cat* is /c/. Phoneme identification requires students to correctly identify matching sounds in different words. For example, when asked what sound is the same in the words, *dog* and *dice*, the child should respond /d/. Categorization refers to a child's ability to identify the odd sound in a sequence of words. For example, if presented with the words, *boat, bug, cow*, and *bat*, a student should be able to identify *cow* as the word that does not belong. Blending is the ability to combine component sounds into a word. For example, when presented with the sounds /f/ /o/ /g/, a child able to blend would know the word is *fog*. Segmenting is the opposite of blending and is typically considered a more difficult skill. When presented with a word, a child who can segment divides the word into its component sounds or phonemes. For example, if presented with the word *coat*, the child would successfully isolate the sounds, /c/ /oa/ /t/. Finally, phoneme deletion refers to the ability to remove a phoneme from a word to make a new word. For example, when presented with the word *string* and told to say it without the /er/ sound, the child would say *sting* (Ehri et al., 2001). This sequence of tasks is typically considered progressively difficult, and deficiencies in these skills have been noted as the primary cause of word reading difficulties (Ehri & McCormick, 1998; Pratt & Brady, 1988).

Given this range of tasks, researchers have attempted to identify which best characterizes phonemic awareness skill for the purposes of screening and efficient prediction of later reading difficulty. O'Connor and Jenkins (1999) tested isolation and phoneme blending tasks and found both were too easy (creating insufficient range) to

provide accurate prediction of later reading skill. Phoneme segmentation yielded stronger validity coefficients, although authors noted a tendency for the measure to overpredict problem readers in kindergarten. Several studies corroborate and expand on O'Connor and Jenkins' (1999) findings. Working with 135 students in preschool through first grade, Liberman et al. (1974) showed that children could segment syllables in words before they could segment sounds in words and that facility with both skills improved with age. Fox and Routh (1975) supported these findings with 50 children, aged 3 to 7, noting significant effects for age on children's ability to segment sentences into words, words into syllables, and syllables into phonemes. Perhaps not surprisingly, phoneme segmentation proved the most difficult task and was most sensitive to growth and change over time. Similarly, Vloedengraven and Verhoven (2007) found that phonological and phonemic awareness assessments were sensitive to growth over time, although they became less predictive of future reading success by late first grade. And Helfgott (1976), who worked with 135 kindergarteners, found segmenting tasks were more difficult than blending tasks.

Unlike other phonological awareness assessments that incorporated fluency of task performance, Chafouleas and Martens (2002) investigated the technical adequacy of accuracy-based phonological awareness measures with two cohorts of 107 kindergarten and first graders. They found that phoneme segmentation was the most sensitive to growth over time, compared to measures of rhyme-providing, sound-providing, blending, and phoneme deletion. This pattern was particularly evident in kindergarten students. Taken together, these findings suggest that different phonemic awareness tasks are more appropriate for the purpose of predicting later reading difficulty as a function of age. If an

assessment is too difficult, floor effects emerge. Conversely, an assessment that is too easy produces a ceiling effect. In both cases, this truncation of range in the predictor variable results in more modest correlations with the criterion variable than would be observed if scores varied across a larger span. This undermines the quality of prediction (Crocker & Algina, 1986). For phoneme segmentation, which is the focus of the present study, several authors have noted floor effects in early kindergarten (Catts, Petscher, Schatschneider, Bridges, & Mendoza , 2009; Elliott, Lee, & Tollefson, 2001; Morris, Bloodgood, & Perney, 2003) and ceiling effects by late first grade (Goffreda et al., 2009; Kaminski & Good, 1996; Vloedengraven & Verhoven, 2007). This suggests the time interval within which it may be most useful for screening students for later development of reading problems.

What Is the Role of Fluency in Sublexical Screening?

Sublexical skill refers to knowledge of subword concepts such as letter names, letter sounds, and phonological awareness (Ritchey & Speece, 2006). As noted, phonemic awareness, a component of phonological awareness, has been shown to be particularly important in the development of reading skill (Adams, 1990; Ehri et al., 2001; National Reading Panel, 2000). Thus, screening assessments that identify students who have phonemic awareness deficits may allow educators to intervene and prevent development of many persistent reading problems. And fluency, which is accurate and quick performance, may be an important component of reading screening assessment because it reveals automaticity. When students are fluent readers, they can devote greater cognitive energy to higher-order tasks such as comprehension (Lyon, 1996; LaBerge &

Samuels, 1974). Thus, measures that assess fluency may provide an indicator of mastery of a component reading skill, which permits students to perform more difficult reading tasks. The importance of fluency to the performance of reading tasks has been established in the research literature (Fuchs, Fuchs, Hosp, & Jenkins, 2001; Shinn, Good, Knutson, Tilly, & Collins, 1992).

The role of fluency in the assessment of sublexical skills is less clear, however. Ritchey and Speece (2006) defined sublexical fluency as "the speed and accuracy with which subword skills can be accessed and produced" (p. 302). According to Burke, Hagan-Burke, Kwok, and Parker (2009), it is typically measured through "the automatic retrieval of phonemes, letter names, and letter sounds and the fluent application of phonological and alphabetic knowledge" (p. 211). Several authors make a theoretical case for assessing sublexical skills through fluency-based assessments (Good, Simmons, & Kame'enui, 2001; Kaminski & Good, 1996; Schilling, Carlisle, Scott, & Zeng, 2007; Burke, Crowder, Hagan-Burkel, and Zou, 2009), but the empirical evidence is unclear. Research suggests that fluency-based sublexical measures are useful, but this work has emphasized letter and sound naming tasks (Elliott et al., 2001; Speece, Mills, Ritchey, & Hillman, 2003), not phoneme segmentation. Furthermore, Elliott et al. (2001) found phonemic awareness and rapid letter naming skill loaded onto separate factors which suggests that different approaches to measurement could be appropriate. Additionally, Chafouleas and Martens (2002) noted that most phonemic awareness tasks are taught as accuracy, not fluency-based skills. Also, work on accuracy-based measures of phoneme segmentation have not been directly compared to fluency-based measures, making uncertain which testing format is preferred in screening (Chalfouleas & Martens, 2002;

Elliott et al., 2001; Ritchey & Speece, 2006). Given the lack of conclusive evidence that phonemic fluency measures are superior to accuracy-based assessments, it is unclear whether fluency is a necessary component of this type of screening, making it a topic that warrants further investigation.

Dynamic Indicators of Basic Early Literacy Skills Phoneme Segmentation Fluency

The Dynamic Indicators of Basic Early Literacy Skills (DIBELS; Good & Kaminski, 2002; Kaminski & Good, 1996) were designed to help meet the demand for sound, efficient assessment of early reading skills. Developed at the University of Oregon in the 1980s and 1990s, these brief, standardized, fluency-based assessments are intended to allow for quick, efficient screening and progress monitoring of students in grades K-3 (Good & Kaminski; Kaminski & Good). DIBELS measures have been widely used to screen over 1,800,000 students (Samuels, 2007) in over 40 states (Manzo, 2005). Different measures of pre and early literacy skills are used at these grade levels to assess whether students may be at-risk for developing reading problems. One measure used to screen and monitor progress of students in kindergarten and first grade, DIBELS PSF, is a focus of the present study.

Although efficient and widely used, there are also several practical challenges associated with fluency-based phoneme segmentation screening. First, correct segmentations cannot be produced above a certain rate because doing so results in sounds not produced in isolation. Instead, sounds run together (and are thus no longer phonemes). And, timing the assessment requires examiners to make quick scoring decisions that can affect reliability. In addition, the speed with which the examiner

supplies words can affect a student's score. Also, a student's prior experience with phonemic awareness tasks may affect understanding of administration instructions, which could affect performance. Furthermore, evaluation of technical studies of DIBELS PSF showed that although it evidenced adequate reliability and validity, diagnostic accuracy was limited (Zumeta & Fuchs, 2009). In general, the measure overpredicted the number of students who would have low achievement on criterion measures of reading and phonological awareness. When Good and Kaminski's (2002) recommended cut-scores were used, correct classification rates ranged from 33% to 58% across criterion measures of reading fluency and phonological processing, and Area Under the Curve statistics fell in the fair to poor discrimination range (Hintze et al., 2001; Ryan, 2004; Tanner, 2006; Trucksess, 2009). This is problematic because diagnostic accuracy information is one of the most important considerations when determining a measure's utility as a screening tool (Jenkins, Hudson, & Johnson, 2007). Given the potential time and resource costs of inaccurate prediction, these findings suggest a need for research to improve technical characteristics of phoneme segmentation screening, most notably in the area of diagnostic accuracy.

One way to improve technical adequacy may be to implement stricter scoring criteria when evaluating student responses during phoneme segmentation screening. Good and Kaminski's (2002) current PSF scoring guidelines require examiners to count several responses that are not true phoneme segmentations as correct. This may inflate scores for some students. For example, they recommend schwa sounds be counted as correct. Elongations of sounds, even if they are not said in isolation are also considered correct. Thus, if a student says, *bbbbbbaaaaaattttt*, instead of /b/ /a/ /t/, segmentations are

counted correct even though phonemes were not segmented. Furthermore, if segmentations are incomplete, the student can earn partial credit. For example if a student says *c... ast* instead of /c/ /a/ /s/ /t/, the student would earn two rather than four points. Although it may be defensible to award a point for correctly isolating the /c/ sound, *ast* is not a phoneme and arguably should not earn credit. Good and Kaminski (2002) recommended a similar scoring rule for overlapping segmentations. If a student says *mi... it* instead of /m/ /i/ /t/, he or she earns two instead of three points. Yet again, technically, these are not phoneme segmentations. Given these issues, it may be useful to determine if stricter criteria that require students to produce isolated segmentations could enhance technical characteristics of phoneme segmentation screening by more clearly distinguishing students who know how to segment correctly from those who do not.

Dynamic Assessment (DA)

Prior research suggests addition of DA may improve the diagnostic utility of screening measures of phonemic awareness (O'Connor & Jenkins, 1999; Spector, 1992). DA is intended to identify students' learning potential by measuring their response to increasingly intensive levels of instructional support. Research in mathematics and reading suggests DA can be a useful predictive tool, and it may enhance diagnostic accuracy of other screening measures (D. Fuchs, Fuchs, Compton, Bouton, Caffrey, & Hill, 2007; L.S. Fuchs, Compton, Fuchs, Hollenbeck, Craddock, & Hamlett, 2008). More specifically, O'Connor and Jenkins (1999) investigated use of a DA of phoneme segmentation skills with 215 first graders and met with moderate success. They found that by using the number of trials needed for a child to learn to segment as a predictor,

they reduced the number of false positive and false negative classifications that occurred. The authors did not test the DA with kindergarten students, however, though they noted it as a potential avenue for future research.

Working with 38 kindergarteners, Spector (1992) also evaluated DA of phoneme segmentation skill. She reported moderate, significant concurrent and predictive validity coefficients, suggesting potential utility of a kindergarten DA of phoneme segmentation skill. She did not, however, report diagnostic accuracy. Given the suggestive albeit incomplete results reported by these authors, it appears further investigation of a kindergarten DA of phoneme segmentation skills is warranted. In particular, inclusion of a diagnostic accuracy analysis with a kindergarten sample could provide a useful addition to the current literature.

Purpose of Present Study

Due to patterns observed in the literature and noted problems with current PSF assessment, the present study was designed to investigate methods to enhance the usefulness of phoneme segmentation screening by evaluating the role of fluency, alternate scoring criteria, and use of DA in the identification of risk for word reading and phonological processing difficulties in emerging readers. Working with kindergarten students during the spring, we compared the technical features of DIBELS PSF to an accuracy-based measure of phoneme segmentation skill, which we refer to as Phoneme Segmentation Accuracy (PSA). In addition, we examined whether stricter scoring criteria affected technical characteristics of either PSF or PSA. Finally, we pilot tested a DA of phoneme segmentation skills with a subsample of students to determine if it enhanced

accurate identification of students with word reading or phonological processing difficulties. To guide this research, we asked the following research questions: How does the split-half reliability, concurrent validity, and concurrent diagnostic accuracy of PSF and PSA compare? Do stricter scoring rules affect reliability, concurrent validity, or concurrent diagnostic accuracy of PSF or PSA? Does DA used alone or in conjunction with PSF or PSA (using either DIBELS or strict scoring rules) improve diagnostic accuracy, compared to PSF alone?

CHAPTER II

METHOD

Participants

A sample of 93 students was recruited from 10 kindergarten classrooms in 3 schools in an ethnically and economically diverse urban school district. Six students moved prior to testing, reducing the final sample to 87 students. In addition, pilot data were collected from a subsample of 37 of the 87 students to evaluate a DA of phoneme segmentation skill. Students in the subsample, who were chosen based on scheduling convenience, came from 8 of the 10 classrooms across the 3 schools. The mean age for the complete sample was 6.2 years, with a standard deviation (*SD*) of 0.3. For the DA subsample, the mean age was 6.1 years (*SD* = 0.3). Additional student demographic information is reported in Table 1 for the complete sample and DA subsample.

With respect to achievement, the sample was roughly representative in terms of phonological processing skill, although there was some overrepresentation of students with low performance on word reading. When compared to a normative sample of scores for the Phonological Awareness Composite of the Comprehensive Test of Phonological Processing (CTOPP-PAC), 26.4% of students in the present sample scored in the bottom quartile, 52.9% scored in the middle two quartiles, and 20.7% scored in the top quartile. When compared to a representative sample of fall of first-grade Word Identification Fluency (WIF) scores (Zumeta, Compton, & Fuchs, 2010), 36.8% of students in the present sample scored in the bottom quartile, 44.8% scored in the middle two quartiles,

Table 1
Demographic Information

	Complete Sample (n = 87) % (n)		DA Subsample (n = 37) % (n)	
Gender				
Male	49.4	(43)	51.4	(19)
Ethnicity				
Caucasian	29.9	(26)	29.7	(11)
African American	55.2	(48)	62.2	(23)
Hispanic	11.5	(10)	8.1	(3)
Other	3.4	(3)		
Subsidized Lunch	59.8	(52)	59.5	(22)
Missing	6.9	(6)	16.2	(6)
IEP	6.0	(4)	8.1	(3)
Retained	2.3	(2)	2.7	(1)
ELL	8.0	(7)	10.8	(4)

and 18.4% scored in the top quartile. Distributions for the DA subsample were similar. When compared to the normative sample on the CTOPP-PPC, 18.9% scored in the bottom quartile, 62.2% scored in the middle two quartiles, and 18.9% scored in the top quartile. On WIF, 40.5% scored in the bottom quartile, 43.3% scored in the middle two quartiles, and 16.2% scored in the top quartile. Given that the present sample's end-of-kindergarten WIF scores were compared to representative data from fall of first grade, the overrepresentation of students in the lowest quartile is not surprising.

Screening Measures

Data were collected on phoneme segmentation skills using two lists, PSF (Good & Kaminski, 2002, 2007), and a modified version, from which the PSA score was derived. In addition, a DA of phoneme segmentation skill was administered to a subsample of 37 students, as described above.

PSF (Good & Kaminski, 2002, 2007). Using the standard DIBELS scripted administration procedure, the examiner verbally presents a list of words, and the child says the sounds in each word. Most lists contain words with two to four phonemes (although a few alternate forms have a word with five phonemes), and each list contains 24 words. The examiner supplies words for 1 min or discontinues testing if a student fails to produce any correct segmentations within the first five words presented. The score is the number of correct segmentations in 1 min. Kaminski and Good (1996) reported alternate form reliability as .88. Good et al. (2004) reported median concurrent validity with the Woodcock-Johnson Readiness cluster as .54 during spring of kindergarten.

Two scores were derived from the PSF assessment, one using Good and Kaminski's (2002) scoring rules (PSF) and one using strict rules (PSF-S). Examiners followed Good and Kaminski's rules to record and score responses during test administration. Then, the author used audio files of the test sessions to rescore assessments using the strict criteria. If audio files were unavailable (due to a dead recorder battery, background noise, etc.), measures were rescored based on the examiner's written notation of student responses. Under the strict scoring rules, responses were counted as correct only if phoneme segmentations were produced in isolation. For example, if a student said *c… ast* instead of /c/ /a/ /s/ /t/, he/she earned one point for the correct isolation of the /c/ sound, but no points for *ast*. If a student said *mi… it* instead of /m/ /i/ /t/, he or she did not earn any points because no phonemes were isolated. Similarly, if a student said *bbbbaaaatttt* instead of /b /a/ /t/, no points were awarded. Schwa sounds were counted incorrect the first time, but not on subsequent responses. This rule prevented the student from being repeatedly penalized for the same pronunciation error. Consistent with traditional PSF scoring guidelines (Good & Kaminski, 2002), accent or dialect differences were not considered errors, as was the case with all data collected in the present study. Scores were independently entered into two databases, and discrepancies were identified and rectified to ensure accurate data were recorded for analysis. This procedure was repeated for all measures in the study.

To assess accuracy of administration, all assessment sessions were audio recorded and 20% were randomly selected across examiners and coded for accuracy using an itemized checklist to determine which points in the testing protocol were correctly addressed. Resulting percentages were calculated and average accuracy of administration

for the PSF list was 97.6%. The first author and a trained research assistant who was not involved with data collection performed the coding and calculations.

PSA. PSA uses the same 24 words, presented in the same order as in the PSF list, but the assessment is untimed. The administration script also differs from the PSF script because it emphasizes that the student should (a) say each sound by itself and (b) not try to go so fast that he/she makes mistakes. Examiners administer all words on the PSA list unless the student fails to produce any correct segmentations in the first five words, at which point the test is discontinued. The PSA score is the total number of correct segmentations produced over the entire list. As with PSF, two scores were derived, one using Good and Kaminski's (2002) scoring rules (PSA) and the other using the strict rules that were outlined above (PSA-S). Using previously described procedures, average accuracy of administration was calculated as 95.6%. Split-half reliability and concurrent validity were evaluated during the study and are reported in the Results section.

DA. The DA is an individually administered assessment that includes a pretest, five levels of instruction, and five level tests requiring students to segment words into component phonemes. The same five words comprise the pretest and each of the level tests, but they are presented in random order on each test. Words with two to four sounds were selected from existing DIBELS PSF lists to create tests and ensure similar task difficulty to DIBELS. During the DA, the examiner orally presents words using a scripted procedure and records student responses. The maximum score on each test is 20 correct segmentations, and the mastery criterion was set at 18 correct segmentations. (Strict scoring rules are used so that only segmented phonemes are counted correct.) Thus, if a student earns a score of 18 or better on the pretest, the assessment is

discontinued. If mastery is not achieved, the examiner provides instruction by following a script to model how to segment words into phonemes. Then, the examiner administers the Level 1 test (a parallel version of the pretest). If the student earns a score of 18 correct segmentations or better, testing is discontinued. If the student does not demonstrate mastery, another, more intensive round of scripted instruction is delivered. This procedure is repeated for up to five levels of instruction, with each level providing increased scaffolding to help the student learn to segment phonemes correctly. The examiner discontinues the assessment once the student achieves mastery or after giving the fifth level of instruction (and corresponding test).

The DA score is derived from the number of levels of instruction a student requires to achieve mastery of the skill. A student who demonstrates mastery at pretest earns six points; a student who masters after one level of instruction receives five points; a student earns four points for mastering after two levels of instruction, and so on. Students who do not demonstrate mastery after five levels of instruction receive no points. Technical characteristics were evaluated during the study and are reported in the Results section. Average accuracy of administration was 100%.

Criterion Measures

Data were collected on four criterion measures of early literacy skills concurrently with the PSF, PSA, and DA. Three were measures of phonological processing; one was an assessment of word reading.

Comprehensive Test of Phonological Processing (CTOPP). CTOPP (Wagner, Torgesen, & Rashotte, 1999) assesses components of phonological processing. Students

were assessed using the Elision, Blending Words, and Sound Matching subtests, which comprise the Phonological Awareness Composite (CTOPP-PAC) for children aged 5 to 6. The Elision consists of 20 items and measures ability to delete sounds in words. The Blending subtest contains 20 items where the student must combine syllables or individual phonemes to make real words. Sound Matching task comprises 20 items and measures the ability to identify words that contain like first and last sounds within a multiple choice format. For each test, the score is the number of correctly answered items. As reported by Wagner et al. (1999), test-retest reliability for the CTOPP-PAC is .79 for 5 to 7 year-olds; predictive validity for the CTOPP-PAC score with the Woodcock Reading Mastery-Revised Decoding Composite is .71; for Elision; .74, for Blending, .61; for Sound Matching, .49; test-retest reliability is .88 for Elision and Blending and .83 for Sound Matching. Average accuracy of administration was 100% for Blending, 97.5% for Elision, and 98.7% for Sound Matching.

Word Reading with Word Identification Fluency (WIF; Fuchs, Fuchs, & Compton, 2004). Students have 1 min to read a list of 50 high-frequency words randomly sampled from 100 high-frequency words from the Dolch preprimer, primer, and first-grade lists. If a student hesitates on a word for 3 sec, the examiner supplies the word. The score is the number of words read correctly. Alternate form reliability at first grade ranges from .95-.97, and concurrent validity with the Woodcock Reading Mastery Word Identification subtest is .91 (Zumeta et al., 2010). Average accuracy of administration was 100%. Although not typically used until first grade, WIF was chosen for its strong technical characteristics. Also, were collected in April and May of kindergarten, when students' WIF performance is likely to be similar to fall of first grade.

Procedure

Kindergarten students were assessed during individual testing sessions by seven master's and doctoral level research assistants during the spring. Testing was delayed for approximately 80% of the sample due a natural disaster that resulted in a week of school closures. Administration of the PSF and PSA was counterbalanced to control for order effects. To minimize memory for words, one list was administered at the beginning and the other list was given at the end of the session.

A subsample of 37 students was also tested to obtain pilot data for the DA, which was administered at the end of the testing battery so instruction did not affect students' scores on the PSF, PSA, or other measures of phonological processing. The assessment took approximately 2-10 min to administer, depending on the point at which students demonstrated mastery. The author administered the DA to all students in the pilot subsample and a research assistant not involved with data collection for the project listened to 20% of these sessions to evaluate accuracy of administration (described above).

All examiners received test administration training, in which test administration was modeled, scoring procedures were explained, and questions were answered. In addition, research assistants' administration accuracy was assessed during individual practice sessions prior to testing through checklists that indexed the percentage of test administration procedures correctly addressed. In cases where errors occurred, they were corrected and rechecked before the examiner began testing students. Across examiners, the average percentage of points accurately addressed prior to error correction during these sessions was 95.8, with a range of 91.4 to 97.1.

CHAPTER III

RESULTS

As noted, four phoneme segmentation scores were derived: (a) PSA with Good and Kaminski's scoring rules (PSA), (b) PSF with Good and Kaminski's rules (PSF), (c) PSA with strict scoring (PSA-S), and (d) PSF with strict scoring (PSF-S). Means and *SD*s for these scores, the DA, and criterion measures are reported in Table 2 for the complete sample and the DA subsample.

Reliability

We assessed split-half reliability of the phoneme segmentation measures (PSA, PSF, PSA-S, and PSF-S) by correlating scores from odd-numbered rows of items with scores from even-numbered rows of items. Each row contained items from the first and second columns of words so that words from the first and second halves of the assessment were equally represented in even and odd scores. Correlations were .96 for PSA, .95 for PSF, .93 for PSA-S, and .91 for PSF-S. Reliability was not evaluated for the DA.

Concurrent Validity

Concurrent validity coefficients for PSA, PSF, PSF-S, PSA-S, and DA against the CTOPP-PAC and WIF are reported in Table 3. All correlations were statistically significant. With respect to the complete sample, concurrent validity coefficients ranged

Table 2
Descriptive Statistics

	Complete Sample (n = 87)		DA Subsample (n = 37)	
	Mean	*(SD)*	Mean	*(SD)*
Predictive Measures				
PSA	51.25	(20.11)	53.41	(20.08)
PSA-S	39.30	(19.21)	43.70	(19.44)
PSF	32.82	(14.32)	33.70	(13.14)
PSF-S	23.69	(13.47)	25.43	(12.69)
DA	--	--	2.92	(2.55)
Criterion Measures				
CTOPP-PAC (SS)	29.69	(6.23)	30.19	(5.28)
WIF	23.83	(20.53)	22.68	(21.94)

PSA is Phoneme Segmentation Accuracy; PSF is Phoneme Segmentation Fluency; PSA-S is Phoneme Segmentation Accuracy with strict scoring rules; PSF-S is Phoneme Segmentation Fluency with strict scoring rules; DA is Dynamic Assessment; CTOPP-PAC (SS) is the Standard Score of the Phonological Awareness Composite of the Comprehensive Test of Phonological Processing; WIF is Word Identification Fluency.

Table 3

Concurrent Validity for Phoneme Segmentation Predictors

	CTOPP-PAC	WIF
Complete Sample (n = 87)		
PSA	.59	.38
PSF	.54	.38
PSA-S	.54	.34
PSF-S	.51	.34
DA Subsample (n = 37)		
PSA	.61	.46
PSF	.49	.43
PSA-S	.65	.44
PSF-S	.59	.43
DA	.63	.59

$p < .01$ for all validity coefficients.

from .51 to .59 when the CTOPP-PAC was the criterion; from .34 to .38 when WIF was the criterion. For the DA subsample, coefficients ranged from .49 to .65 against the CTOPP-PAC criterion; from .43 to .59 against WIF. These validity coefficients were compared using Walker and Lev's (1953) formula to determine if differences were statistically significant. As Table 4 shows, *t* test comparisons revealed no significant differences in the magnitudes of validity coefficients.

Commonality Analysis

Next, commonality analysis was conducted by block-entering predictors into regression models to determine unique and shared variance explained by PSA, PSF, PSA-S, and PSF-S. Results for the complete sample are reported in Table 5. When PSA and PSF were used to predict CTOPP-PAC performance, PSA uniquely explained an additional 7.0% variance, which was statistically significant. By contrast, PSF did not significantly explain any additional variance. When PSA-S and PSF-S were used to predict CTOPP-PAC performance, PSA-S accounted for a significant additional 4.8% variance, whereas PSF-S did not account for any significant additional variance. When PSA and PSA-S were used to predict CTOPP-PAC performance, PSA accounted for an additional 5.9% unique variance, which was significant; the contribution of PSA-S was not significant. The same pattern was evident when PSF and PSF-S were used to predict CTOPP-PAC scores. PSF explained a significant 3.4% additional unique variance; the contribution of PSF-S was not significant. With respect to models predicting WIF performance, none of the phoneme segmentation predictors (PSA, PSF, PSA-S, or PSF-F) were uniquely predictive. Across analyses, the model that included PSF and PSA

Table 4

Tests of the Differences in Concurrent Validity Coefficients

	Outcomes	
	CTOPP-PAC	WIF
	t value	*t* value
Complete Sample (*n* = 87)		
PSA v. PSF	0.98	0.00
PSA v. PSA-S	1.34	0.93
PSA v. PSF-S	1.24	0.54
PSF v. PSA-S	-0.81	0.56
PSF v. PSF-S	0.87	0.84
PSA-S v. PSF-S	0.53	0.00
DA Subsample (*n* = 37)		
PSA v. PSF	1.25	0.28
PSA v. PSA-S	-0.97	0.42
PSA v. PSF-S	0.20	0.26
PSA v. DA	-0.20	-1.17
PSF v. PSA-S	-1.13	0.00
PSF v. PSF-S	-1.53	0.00
PSF v. DA	-1.14	-1.45
PSA-S v. PSF-S	0.65	-0.09
PSA-S v. DA	0.21	-1.24
PSF-S v. DA	-0.33	-1.21

No differences were significant. See Appendix for correlations used in Walker and Lev's (1953) *t* test comparison equation.

Table 5

*Overall, Common, and Unique Variance Explained by PSA, PSF, PSA-S, and PSF-S for the Complete Sample (*n = *87)*

	Outcome	
	CTOPP-PAC R^2%	WIF R^2%
Model 1		
PSA Only	34.9***	14.6**
PSF Only	28.6***	14.6**
PSA + PSF Overall	35.6	16.0
PSA + PSF Common	28.5	13.2
PSA Unique	7.0**	1.4
PSF Unique	0.1	1.4
Model 2		
PSA-S Only	29.1***	11.8**
PSF-S Only	25.7***	11.7**
PSA-S + PSF-S Overall	30.5	13.0
PSA-S + PSF-S Common	24.3	10.5
PSA-S Unique	4.8*	1.3
PSF-S Unique	1.4	1.2
Model 3		
PSA Only	29.1***	14.6***
PSA-S Only	34.9***	11.8***
PSA + PSA-S Overall	34.9	14.6
PSA + PSA-S Common	29.0	11.8
PSA Unique	5.9**	2.8
PSA-S Unique	0.0	0.0
Model 4		
PSF Only	28.6***	14.6***
PSF-S Only	25.7***	10.6***
PSF + PSF-S Overall	29.1	14.6
PSF + PSF-S Common	25.3	11.7
PSF Unique	3.4*	2.9
PSF-S Unique	0.4	0.0

* $p < .05$; ** $p < .01$; *** $p < .001$.

explained the largest amount of overall variance (against both CTOPP-PAC and WIF criterion measures).

In addition, a commonality analysis was conducted for the DA subsample, and results are reported in Table 6. Given that these are pilot results derived from a small sample, *p*-values less than .10 were flagged as potentially significant. Across models and criterion measures, the DA uniquely explained significant additional variance. When CTOPP-PAC was the criterion, the DA significantly explained 5.7% to 18.0% additional variance. Other phonemic awareness predictors explained 4.6% to 10.5% significant additional variance. When WIF was the criterion, DA's significant contributions were even larger, ranging from 13.6% to 18.5% additional unique variance. Across these models, none of the other phoneme segmentation predictors (PSA, PSF, PSA-S, or PSF-S) contributed significant unique variance when WIF was the criterion. Notably, the DA contributed the most additional unique variance in models where the other predictor was a fluency-based (PSF or PSF-S), not an accuracy-based (PSA or PSA-S) measure.

Diagnostic Accuracy

Finally, diagnostic accuracy was assessed for predictors of the CTOPP-PAC and WIF. Diagnostic classifications fall into four categories that drive evaluations of an assessment's overall diagnostic accuracy (Swets, 1988, 1992). True positive classifications (also known as sensitivity) occur when a measure accurately detects when a disorder is present. True negative classifications (also known as specificity) are the correct determination that a disorder is not present. False positive classifications occur

Table 6

Overall, Common, and Unique Variance Explained by PSA, PSF, PSA-S, PSF-S and DA for the DA Subsample (n = *37)*

	Outcome	
	CTOPP-PAC R^2%	WIF R^2%
Model 1		
PSA Only	37.4***	21.1***
DA Only	39.6***	35.2***
PSA + DA Overall	45.9	35.8
PSA + DA Common	32.8	20.6
PSA Unique	4.6*	0.6
DA Unique	8.5**	14.6***
Model 2		
PSF Only	24.2***	18.6***
DA Only	39.6***	35.2***
PSF + DA Overall	42.4	36.5
PSF + DA Common	21.7	18.2
PSF Unique	2.7	1.3
DA Unique	18.0***	17.0***
Model 3		
PSA-S Only	41.7***	19.2***
DA Only	39.6***	35.2***
PSA-S + DA Overall	47.6	35.2
PSA-S + DA Common	33.7	19.1
PSA-S Unique	8.0**	0.1
DA Unique	5.9*	16.0***
Model 4		
PSF-S Only	34.6***	18.4***
DA Only	39.6***	35.2***
PSF-S + DA Overall	48.4	36.9
PSF-S + DA Common	25.8	16.7
PSF-S Unique	8.8**	1.7
DA Unique	13.8***	18.5***
Model 5		
PSA + PSF Only	37.7***	22.9**
DA Only	39.6***	35.2***
PSA + PSF + DA Overall	46.0	36.5
PSA + PSF + DA Common	31.4	21.6
PSA + PSF Unique	6.3	1.3
DA Unique	8.3**	13.6**
Model 6		
PSA-S + PSF-S Only	44.4***	21.6**
DA Only	39.6***	35.2***
PSA-S + PSF-S + DA Overall	50.1	37.3
PSA-S + PSF-S + DA Common	33.9	19.4
PSA-S + PSF-S Unique	10.5**	2.2
DA Unique	5.7*	15.7***

*$p < .10$; **$p < .05$; *** $p < .01$.

when a screening tool incorrectly identifies a disorder as present; false negatives occur when a measure fails to detect a disorder that is present (Swets, 1988, 1992). For the purposes of RTI and other efforts to identify early reading risk, the goal of screening is to maximize identification of true positives while limiting false positives. To evaluate diagnostic accuracy, we used logistic regression to contrast competing screening models' ability to accurately predict difficulty status on phonological awareness and word reading. Sensitivity was set at .90 (or as close as possible), and the resulting effect on specificity across models was observed to determine if accuracy or fluency-based screens, strict scoring criteria, or use of DA affected correct classification of students. Models were contrasted by examining sensitivity, specificity, the correct classification percentage (or overall hit rate), reduction in false positive classifications, and the receiver operator characteristic (ROC) curve. ROC curves graphically depict true positive rates against false positive rates across a range of cut-points and are analyzed by examining the area under the curve (AUC). AUC is a measure of discrimination that can contrast the predictive accuracy of logistic regression models (Swets, 1988, 1992). Put another way, AUC is the proportion of randomly chosen pairs of students for whom screens correctly classify the presence or absence of phonological awareness or word reading deficits. Larger AUC values mean classifications were less likely due to chance. AUC values below .70 indicate a poor prediction model; .70 to .80 fair; .80 to .90 good; and above .90 excellent (Swets, 1992). ROC analysis also provides confidence intervals for AUC statistics, and lack of overlap between confidence intervals indicates a significant difference in the predictive accuracy of contrasted models. Chi-square tests that compare

contrasted models are less strict and can also be used to provide evidence of the presence of a significant difference between tested models.

Results of the diagnostic accuracy analysis for the complete sample are reported in Table 7. For the purpose of dichotomizing risk status, the cut-score was set at the 25th percentile on both criterion measures. When CTOPP-PAC was the criterion and sensitivity was set at .87, diagnostic accuracy statistics for PSF yielded specificity of .30, a correct classification percentage of 44.8%, a false positive rate of 50.6%, and an AUC of .74. Across indicators, diagnostic accuracy statistics favored the PSA. When sensitivity was set at .87, PSA had the highest specificity (.65), classification accuracy (71.3%), and AUC value (.82), and the lowest false positive rate (25.3%), compared to PSF, PSA-S, and PSF-S. Confidence intervals overlapped, which showed that models were not significantly different from one another. However, chi-square comparisons showed that AUC differences for PSA v. PSF approached significance (p = .058) and were significant for PSA v. PSA-S (p = .037). In addition, classification accuracy was 26.5 percentage points higher for PSA than for PSF, and PSA had half the number of false positive classifications, compared to PSF.

When WIF was the criterion, diagnostic accuracy statistics were poor across models. With sensitivity set at .91, specificity for PSF was .13, the correct classification percentage was 42.5%, the false positive rate was 54.0%, and AUC was .66, suggesting poor discrimination. Diagnostic accuracy statistics for the other predictors were similar, with all AUC statistics .70 or below, overlapping confidence intervals, and chi-square comparisons that yielded nonsignificant differences for most comparisons. An exception was noted for the PSA v. PSA-S (p = .005) chi- square comparison; it also approached

Table 7

Diagnostic Accuracy Analysis: Complete Sample (n = 87)

	Sens.	Spec.	TN	FN	TP	FP	% Correct	AUC *(SE)*	Model X^2	Confidence Interval
CTOPP-PAC Criterion										
PSA**	.87	.65	42	3	20	22	71.3	.82 (.05)	22.25***	.72-.92
PSF**	.87	.30	19	3	20	45	44.8	.74 (.07)	17.29***	.61-.87
PSA-S**	.87	.53	34	3	20	30	62.1	.77 (.06)	17.53***	.66-.87
PSF-S**	.91	.55	35	2	21	29	64.4	.77 (.06)	14.82***	.66-.88
WIF Criterion										
PSA**	.91	.17	9	3	30	45	44.8	.70 (.06)	11.49***	.58-.82
PSF**	.91	.13	7	3	30	47	42.5	.66 (.06)	8.84**	.53-.78
PSA-S*	.91	.15	8	3	30	46	43.7	.63 (.07)	5.74*	.50-.76
PSF-S	.91	.11	6	3	30	48	41.4	.63 (.07)	3.75	.50-.75

*$p < .05$; **$p < .01$; ***$p < .001$. Criterion measure cut scores were set at the 25th percentile. TN is the number of True Negative identifications, FN is the number of False Negatives, TP is the number of True Positives, and FP is the number of False Positives.

significance for the PSA v. PSF-S (p = .078) comparison.

Results of the diagnostic accuracy analysis for the DA subsample are reported in Table 8. When CTOPP-PAC was the criterion and sensitivity was set at .86, specificity ranged from .60 to .80 across models. Classification accuracy ranged from 64.9% (DA alone) to 81.1% (PSF-S alone) and false positive classification rates ranged from 16.2% to 32.4%. AUC statistics ranged from .77 to .88, suggesting fair to good discrimination for tested models. Although DA was a significant predictor when it was the only phoneme segmentation predictor included in the model, it was not significant when included with the other predictors (PSA, PSF, PSA-S, or PSF-S). All AUC confidence intervals overlapped, and chi-square tests revealed nonsignificant differences between models. (The chi-square comparison between DA v. PSA AUC statistics approached significance, however: p = .084.)

The significance of the DA predictor changed when WIF was the criterion of interest. As Table 8 shows, DA was a significant predictor of WIF when it was the only predictor in the model, and remained significant even when other phoneme segmentation predictors were included. Importantly, DA also yielded enhanced specificity and classification accuracy and reduced false positive classifications. When DA was included in models and sensitivity was set at .87, specificity was between .59 and .68; this compares to .09-.41 when other phoneme segmentation predictors were included alone. Furthermore, classification accuracy ranged from 70.3% to 75.7% when DA was included in prediction models, compared to 40.5% to 59.5% when other phoneme segmentation predictors were included alone. False positive classifications ranged from 18.9% to 24.3% when DA was included in models, compared to 35.1% to 54.1% when other predictors were used alone. AUC statistics ranged from .76 to .78 for models with DA; from .64 to .73 for models without DA. All AUC confidence intervals overlapped, and chi-

Table 8

Diagnostic Accuracy Analysis: DA Subsample (n = 37)

	Sens.	Spec.	TN	FN	TP	FP	% Correct	AUC *(SE)*	Model X^2	Confidence Interval
CTOPP-PAC Criterion										
DA**	.86	.60	18	1	6	12	64.9	.77 (.08)	5.38**	.62-.93
PSA**	.86	.77	23	1	6	7	78.4	.87 (.07)	9.21***	.70-1.0
PSF**	.86	.70	21	1	6	9	73.0	.84 (.08)	8.92***	.68-.99
PSA-S**	.86	.63	19	1	6	11	67.6	.78 (.10)	7.48***	.59-.97
PSF-S**	.86	.80	24	1	6	6	81.1	.88 (.06)	10.60***	.77-.99
DA + PSA**	.86	.73	22	1	6	8	75.7	.86 (.07)	9.42***	.71-1.0
DA + PSF*	.86	.73	22	1	6	8	75.7	.84 (.07)	9.25**	.70-.99
DA + PSA-S	.86	.60	21	1	6	9	64.9	.82 (.08)	7.71**	.65-.98
DA + PSF-S*	.86	.80	24	1	6	6	81.1	.88 (.06)	10.67**	.77-.99
WIF Criterion										
DA**	.87	.59	13	2	13	9	70.3	.76 (.09)	9.71***	.60-.93
PSA**	.87	.41	9	2	13	13	59.5	.73 (.09)	6.28**	.56-.91
PSF**	.87	.14	3	2	13	19	43.2	.69 (.10)	5.11**	.50-.88
PSA-S*	.87	.23	5	2	13	17	48.6	.64 (.10)	3.43*	.45-.84
PSF-S	.87	.09	2	2	13	20	40.5	.66 (.10)	1.93	.47-.86
DA* + PSA	.87	.64	14	2	13	8	73.0	.78 (.09)	10.04***	.61-.95
DA** + PSF	.87	.64	14	2	13	8	73.0	.78 (.09)	10.09***	.61-.95
DA** + PSA-S	.87	.68	15	2	13	7	75.7	.77 (.09)	9.90***	.60-.93
DA*** + PSF-S	.87	.59	13	2	13	9	70.3	.77 (.09)	9.81***	.61-.93

*$p < .10$; **$p < .05$; *** $p < .01$; ****$p < .001$. Criterion measure cut scores were set at the 25th percentile. TN is the number of True Negative identifications, FN is the number of False Negatives, TP is the number of True Positives, and FP is the number of False Positives.

square comparisons were nonsignificant, though the DA v. PSA-S comparison approached significance, p =.080.

CHAPTER IV

DISCUSSION

Commonality and diagnostic accuracy analyses indicated some advantage for PSA and DA over PSF, although reliability for the various indices appears comparable. In terms of split-half reliability, coefficients were similarly high regardless of whether accuracy or fluency was indexed and for the more lenient and strict scoring criteria, with figures ranging between .91 (for PSF-S) to .96 (PSA). These results are consistent with findings from prior work on PSF. Kaminski and Good (1996) reported split-half reliability of .99; others have reported different forms of reliability, including test-retest coefficients between .69 and .85 (Catts et al., 2009; Elliott et al., 2001) and alternate form coefficients between .84 to .92 (Elliott et al., 2001; Good et al., 2004; Kaminksi & Good, 1996). Schedule and timing constraints prohibited us from including a test-retest or alternate form reliability analysis in the present study, but given the age of several of these other studies, such an evaluation of phoneme segmentation screeners would be a useful component of future research. Also, the relatively small sample and brevity of level tests made it impossible to evaluate reliability for the DA. Thus, future research on DA of phoneme segmentation skill should examine test-retest or alternate form reliability.

With respect to concurrent validity, coefficients for all screening measures, including DA, were statistically significant. The magnitude of the validity coefficient for PSF against the CTOPP-PAC (.54) was consistent with prior studies in which phonological processing was the criterion (Hintze et al., 2003; Nelson, 2008). Further, the PSA validity coefficient (also .54) was consistent with results of Elliott et al. (2001) assessing an accuracy-based phoneme segmentation

task. When WIF, the word reading measure, was the criterion, concurrent validity coefficients for the fluency and accuracy phoneme segmentation screener were considerably lower than when phonemic awareness was the criterion, but again the coefficients for the fluency and accuracy-based versions of the screeners were identical at .38. This is consistent with previous research in which assessments of decoding or sight word reading were the criterion measures (Elliott et al., 2001; Kamii & Manning, 2005; Rouse & Fantuzzo, 2006). With respect to the DA, validity coefficients (.63 against CTOPP-PAC criterion; .59 against WIF) were consistent with the values Spector (1992) reported in the fall administration of her DA (.51 to .60 against spring measures of phonological processing and word reading).

Our inferential tests of differences between these various validity coefficients, however, revealed no significant differences, indicating that none of the screening tools (PSA, PSF, PSA-S, or PSF-S) explained more total variance in criterion outcomes than the others. As with the reliability results, therefore, validity data do not provide a basis for favoring accuracy over fluency, or strict over traditional DIBELS scoring rules. Yet, across these reliability and validity analyses, results also do not suggest the superiority of a fluency-based phoneme segmentation task. A similar pattern was observed for the DA in comparison to PSA, PSF, PSA-S, and PSF-S.

At the same time, despite the lack of significant differences in *total* variance explained, commonality analysis provided evidence of the importance of accuracy- over fluency-based phoneme segmentation screening, use of traditional DIBELS scoring rules, and the potential utility of the DA. When both PSF and PSA were used to predict CTOPP-PAC performance, PSA contributed significant additional variance beyond PSF alone, whereas PSF failed to explain any additional unique variance. A similar pattern was evident when PSA-S and PSF-S were used to predict CTOPP-PAC. PSA-S explained significant additional variance, but PSF-S did not. Thus,

these findings provide evidence of the importance of accuracy-based phoneme segmentation screening used in addition to fluency-based phoneme segmentation screening. In terms of scoring, when conventional DIBELS and strict rules were compared, using CTOPP-PAC as the criterion, DIBELS rules explained additional unique variance across accuracy and fluency-based assessments (PSF and PSA v. PSF-S and PSA-S, respectively); strict scoring rules did not.

Overall, the model that included PSF and PSA together accounted for the largest amount of total variance against the CTOPP-PAC criterion. Taken together, these results suggest the addition of PSA to PSF may enhance the prediction of phoneme segmentation screening when phonological awareness is the criterion of interest and that across accuracy and fluency-based assessments, DIBELS scoring rules are preferred to the strict scoring rules. At the same time, when WIF was the criterion, none of the phoneme segmentation tasks (PSA, PSF, PSA-S, or PSF-S) explained significant additional variance. Also, the total variance explained across models was small. Thus, when used alone, none of tested the measures appeared to be a particularly useful correlate of word reading in late kindergarten.

This pattern of findings changed dramatically, however, when the DA was considered. Across models and criterion measures, DA added additional unique variance beyond other phoneme segmentation predictors. Unique contributions were particularly notable when WIF was the criterion, with DA explaining an additional 13.6 to 18.5 percentage points of significant additional variance across models; PSA, PSF, PSA-S, and PSF-S did not explain any significant additional variance. When CTOPP- PAC was the criterion, DA also uniquely explained 5.7 to 18.0 percentage points of additional variance across models. Similar, although smaller patterns were observed for PSA, PSA-S, and PSF-S. However, PSF did not contribute any additional unique variance, and DA increased the proportion of variance explained the most when it was

added to the PSF model. This is notable because PSF is used by thousands of schools across the country as an early literacy screening tool (Manzo, 2005). Our results occurred even for models that included both PSF and PSA (the strongest overall model when the complete sample was analyzed). When CTOPP was the criterion, inclusion of the DA uniquely explained an additional 8.3% variance. And, when WIF was the criterion, inclusion of the DA uniquely explained an additional 13.6% variance. Inclusion of PSA and PSF together did not significantly increase the proportion of variance explained over DA alone against either criterion (CTOPP-PAC or WIF).

Taken together, these findings provide preliminary evidence that DA may be worthwhile addition to a screening battery, especially when practitioners are interested in word reading as the criterion. Of course, future research that predicts performance over time and incorporates a larger sample is needed. If results for DA hold in these future studies, then DA might be used productively to identify and intervene early with students who are likely to have reading problems. This would help address one of the primary purposes of screening, particularly for schools implementing RTI.

We also conducted diagnostic accuracy analyses to gain insight into the practical utility of these phoneme segmentation screening tools. Using logistic regression, risk cut-offs were set at the 25th percentile for both the CTOPP-PAC and WIF criterion. Across models, sensitivity was held at .90 (or as close as possible) and resulting specificity was observed. When CTOPP-PAC was the criterion, PSF had the worst specificity, resulting in the largest number of false positives. By comparison, PSA had the best specificity and correct classification percentage, with a false positive rate that was less than half that of PSF and an AUC value (.82) indicating good predictive accuracy (Swets, 1992). Although AUC confidence intervals overlapped, chi-square comparisons did approach significance when PSA was compared to PSF. This overall enhanced

classification accuracy provides further evidence that PSA may be a useful addition to, if not a replacement for, PSF for screening kindergarten students, at least when CTOPP was the criterion.

When WIF was the criterion, however, differences between screening tools were less distinct, and quality of prediction was poor across models, with AUC statistics .70 or below. When sensitivity was set at .91, specificity was low, as was classification accuracy and the false positive rate. Therefore, none of the brief screening tools (PSA, PSF, PSA-S, or PSF-S) provided acceptable diagnostic accuracy when used alone with WIF as the criterion. At the same time, the addition of DA to logistic models did enhance diagnostic accuracy when WIF was the criterion. This was the case across models. When added to PSF, DA reduced false positives by more than half. The AUC was in the fair range for DA, whereas it was poor for PSF, although the models did not differ significantly. Further, when DA and PSF were combined, results were comparable to sole reliance on DA. Moreover, PSF was not a significant predictor, suggesting it does not enhance accurate identification of word reading risk. Again, these findings indicate DA may be a useful addition to a kindergarten literacy screening battery, particularly when word reading is the outcome of interest.

By contrast, when CTOPP-PAC was the criterion, DA performed less well. When DA was the only predictor in the model, classification accuracy was 65%, with a false positive rate of 32%, higher than models in which other phoneme segmentation predictors were used alone. The AUC value was fair, but not significantly different from other models. Further, DA was not significant when added to models that already included PSA, PSF, PSA-S, or PSF-S. Thus, DA may not enhance classification of phonological processing deficits beyond what is achieved with more efficient phoneme segmentation screening tools.

DA may improve classification of word reading difficulty because it helps control for task novelty and ambiguity as reasons for poor performance. That is, at screening, students may not sort accurately into two groups: a not-at-risk group who can segment sounds with accuracy who are likely to profit from other literacy instruction and an at-risk group who cannot segment sounds with accuracy and have true phonological processing, attention, or cognitive deficits that inhibit reading development. Rather, a third group may exist. These children may simply not understand the instructions for the phoneme segmentation task or may be insufficiently familiar with the task to perform it competently. With standard, static screening, these students are categorized with the at-risk groups of children. Yet, a small amount of intervention, as provided in the DA, may reveal the capacity to profit from instruction. Thus, these students' initial performance on such tasks, as with standard, static screening, may belie their true ability to identify and manipulate sounds. This is how DA may serve to reduce the number of false positive classifications. This phenomenon may not hold when CTOPP-PAC is the criterion, however, because screening tasks (PSF, PSA, PSA-S, PSF-S, and DA) and CTOPP-PAC are representative of the same domain, phonological processing.

Findings have preliminary implications for practice, even as they raise questions for future study. First, it would be worthwhile to learn if testing alternate phoneme segmentation screening measures (PSA, PSA-S, PSF-S) earlier in kindergarten (i.e., winter or early spring, but not fall due to previously observed floor effects (Catts et al., 2009; Elliott et al., 2001; Morris et al., 2003) for PSF) would enhance validity coefficients due to a broader range in student performance at those earlier times of the year. In addition, predicting outcomes over time is an essential test for considering the value of these screeners. Also, it may be worthwhile to assess

validity and diagnostic accuracy against more comprehensive measures of reading to determine the extent to which these screening tools predict performance on high-stakes tests.

With respect to the DA, these pilot data need to be replicated with a larger sample, with evaluation of reliability, predictive validity, and predictive diagnostic accuracy. In addition, it may be worthwhile to incorporate other phonological processing skills such as initial sound identification or blending as different levels of instruction within the DA. Such an addition could be particularly useful if the study were conducted early in the fall of kindergarten when initial sound fluency, not PSF, is recommended for screening within DIBELS (Good & Kaminski, 2002; Kaminski & Good, 1996). Finally, it could be worthwhile to investigate the feasibility of using DA in a second stage of screening to verify the status of students identified as at-risk by other, more efficient, universal screening tools (i.e. DIBELS PSF). In a related way, it may also be useful to learn whether DA scores can be used to identify students who are likely to be unresponsive to secondary (i.e., Tier 2) intervention within an RTI system and should therefore proceed directly to a more intensive level of the prevention system.

As these recommendations for future research indicate, several methodological issues represent important limitations to the present study. First, the sample was small, perhaps underpowering the study and making it hard to detect true differences. This is particularly problematic for analyses involving DA, where the sample size was 37. Relatedly, regression models may have been underspecified, particularly with respect to DA analyses. Thus, future research with larger samples should consider the role of other covariates such as attention and language skills in the evaluation of DA. In addition, testing occurred late in the school year. Also, DA was administered by a single examiner (also the author), which may have inadvertently affected scores, despite the documented strong accuracy with which testing occurred.

With these limitations in mind, the data presented here must be considered only suggestive. Even so, they do raise questions about the previously assumed but untested assumption that fluency is a necessary component of valid phoneme segmentation screening assessment. Although PSA and DA explained comparable amounts of *overall* variance, they did significantly increase the amount of explained variance when added to PSF, even as they enhanced classification accuracy in many instances. By contrast, PSF did not increase the amount of explained variance in any of the models. Also, application of DIBELS scoring rules generally increased explained variance and enhanced some indicators of diagnostic accuracy over strict rules, which may indicate that word segmentation, not phoneme segmentation, is the necessary skill to evaluate. This hypothesis warrants further investigation. Finally, although DA results are preliminary, they suggest that dynamic testing of phoneme segmentation skill may enhance diagnostic accuracy at kindergarten.

APPENDIX

Correlations used in Walker and Lev's Formula Calculations Reported in Table 3 for the Complete Sample (n =87)

	PSA	PSF	PSA-S	PSF-S
PSA	1.00	.83	.91	.73
PSF		1.00	.75	.89
PSA-S			1.00	.81
PSF-S				1.00

All correlations are significant at p< .01 level.

Correlations used in Walker and Lev's Formula Calculations Reported in Table 3 for the DA Subsample (n =37)

	PSA	PSF	PSA-S	PSF-S	DA
PSA	1.0	.75	.95	.71	.68
PSF		1.0	.68	.90	.57
PSA-S			1.0	.74	.71
PSF-S				1.0	.54
DA					1.0

All correlations are significant at p< .01 level.

REFERENCES

Adams, M.J. (1990). *Beginning to Read: Thinking and Learning about Print.* Boston: MIT Press.

Bradley, L., & Bryant, P.E. (1983). Categorizing sounds and learning to read—A causal connection. *Nature,* 301, 419-421.

Burke, M.D., Crowder, W., Hagan-Burke, S., & Zou, Y. (2009). A comparison of two path models for predicting reading fluency. *Remedial and Special Education,* 30, 84-95.

Burke, M.D., Hagan-Burke, S., Kwok, O., & Parker, R. (2009). Predictive validity of early literacy indicators from the middle of kindergarten to second grade. *Journal of Special Education,* 42, 209-236.

Catts, H.W., Petscher, Y., Schatschneider, C., Bridges, M.S., & Mendoza, K. (2009). Floor effects associated with universal screening and their impact on the early identification of reading disabilities. *Journal of Learning Disabilities,* 42, 163-176.

Chafouleas, S.M., & Martens, B. K. (2002). Accuracy-based phonological awareness tasks: Are they reliable, efficient, and sensitive to growth? *School Psychology Quarterly,* 17, 128-147.

Crocker, L., & Algina, J. (1986). *Introduction to Classical and Modern Test Theory.* Belmont, CA: Wadsworth.

Ehri, L.C., & McCormick, S. (1998). Phases of word learning: Implications for instruction with delayed and disabled readers. *Reading & Writing Quarterly,* 14, 135-163.

Ehri, L.C., Nunes, S.R., Willows, D.M., Schuster, B.V., Yagoub-Zadeh, Z., & Shanahan, T. (2001). Phonemic awareness instruction helps children learn to read: Evidence from the National Reading Panel's meta-analysis. *Reading Research Quarterly,* 36, 250-287.

Elliott, J.K., Lee, S.W., & Tollefson, N. (2001). A reliability and validity study of the dynamic indicators of basic early literacy skills—Modified. *School Psychology Review,* 30, 33-49.

Fox, B., & Routh, D.K. (1975). Analyzing spoken language into words, syllables, and phonemes: A developmental study. *Journal of Psycholinguistic Research,* 4, 331-342.

Fuchs, D., Fuchs, L.S., Compton, D.L., Bouton, B., Carffrey, E., & Hill, L. (2007). Dynamic assessment as responsiveness to intervention: A scripted protocol to identify young at-risk readers. *Teaching Exceptional Children, 39,* 58-63.

Fuchs, L. S., Fuchs, D., & Compton, D.L. (2004). Monitoring early reading development in first grade: Word identification fluency versus nonsense word fluency. *Exceptional Children, 71,* 7-21.

Fuchs, L.S., Fuchs, D., & Zumeta, R.O. (2008). Responsiveness to Intervention: An alternative strategy for the prevention and identification of learning disabilities. In E.L. Grigorenko (Ed.). *Educating individuals with disabilities: IDEIA 2004 and beyond.* New York: Springer.

Fuchs, L.S., Fuchs, D., Hosp, M.K., & Jenkins, J.R. (2001). Oral reading fluency as an indicator of reading competence: A theoretical, empirical, and historical analysis. *Scientific Studies of Reading, 5*, 239-256.

Fuchs, L.S., Compton, D.L., Fuchs, D., Hollenbeck, K.N., Craddock, C., & Hamlett, C.L. (2008). Dynamic assessment of algebraic learning in predicting third graders' development of mathematical problem solving. *Journal of Educational Psychology, 100*, 829-850.

Goffreda, C.T., DiPerna, J.C., & Pedersen, J.A. (2009). Preventive screening for early readers: Predictive validity of the dynamic indicators of basic early literacy skills (DIBELS). *Psychology in the Schools, 46,* 539-552.

Good, R.H., & Kaminski, R.A. (Eds.). (2002). *Dynamic Indicators of Basic Early Literacy Skills* (6th ed.). Eugene, OR: Institute for the Development of Educational Achievement. Available: http://dibels.uoregon.edu/.

Good, R.H., Gruba, J. & Kaminski, R.A. (2002) Best practices in using Dynamic Indicators of Basic Early Literacy Skills (DIBELS) in an outcomes-driven model. In A. Thomas & J. Grimes. *Best Practices in School Psychology* (4th ed.). Bethesda, MD: National Association of School Psychologists.

Good, R.H., Kaminski, R.A., & Smith, S. (2007). Phoneme Segmentation Fluency. In *Dynamic Indicators of Basic Early Literacy Skills* (6th ed.). Eugene, OR: Institute for the Development of Educational Achievement. Available: http://dibels.uoregon.edu/.

Good, R. H., Simmons, D. C., & Kame'enui, E. J. (2001). The importance of decision-making utility of a continuum of fluency-based indicators of foundational reading skills for third-grade high-stakes outcomes. *Scientific Studies of Reading, 5,* 257-288.

Good, R.H., Kaminski, R.A., Shinn, M., Bratten, J., Shinn, M., Laimon, D., Smith, S., & Flindt, N. (2004). *Technical Adequacy of DIBELS: Results of Early Childhood Research Institute on Measuring Growth and Development* (Technical Report, No.7) Eugene, OR: University of Oregon.

Helfgott, J.A. (1976). Phonemic segmentation and blending skills of kindergarten children: Implications for beginning reading acquisition. *Contemporary Educational Psychology,* 1, 157-169.

Hintze, J.M., Ryan, A.L., & Stoner, G. (2003). Concurrent validity and diagnostic accuracy of the dynamic indicators of basic early literacy skills and the comprehensive test of phonological processing. *School Psychology Review,* 32, 541-556.

Individual with Disabilities Education Improvement Act, Public Law 108-446 (2004).

Jenkins, J.R., Hudson, R.F., & Johnson, E.S. (2007). Screening at-risk readers in a response to intervention framework. *School Psychology Review, 582-600.*

Juel, C. (1988). Learning to read and write: A longitudinal study of 54 children from first through fourth grades. *Journal of Educational Psychology, 80,* 437-447.

Kaminski, R.A., & Good, R.H. (1996). Toward a technology for assessing basic early literacy skills. *School Psychology Review, 25*, 215-227.

LaBerge, D., & Samuels, SJ. (1974). Toward a theory of automatic information processing in reading. *Cognitive Psychology, 6*, 293-323.

Liberman, I.Y., Shankweiler, D., Fischer, F.W., & Carter, B. (1974). Explicit syllable and phoneme segmentation in the young child. *Journal of Experimental Child Psychology,* 18, 201-212.

Lyon, G.R. (1996). Learning disabilities. *The Future of Children, 6* (1), 54-76.

Manzo, K.K. (2005). National clout of DIBELS test draws scrutiny. *Education Week, 25,* 11-12.

Morris, D., Bloodgood, J., & Perney, J. (2003). Kindergarten predictors of first- and second-grade reading achievement. *The Elementary School Journal,* 104, 93-103.

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 of the subgroups.* Bethesda, MD: National Institute of Child Health and Human Development. Retrieved August 9, 2009, from http://www.nichd.nih.gov/publications/nrp/report/cfm.

O'Connor, R.E., & Jenkins, J.R. (1999). Prediction of reading disabilities in kindergarten and first grade. *Scientific Studies of Reading, 3,* 159-197.

Pratt, A.C., & Brady, S. (1988). Relation of phonological awareness to reading disability in children and adults. *Journal of Educational Psychology, 80*, 319-323.

Ritchey, K.D., & Speece, D.L. (2006). From letter names to word reading: The nascent role of sublexical fluency. *Contemporary Educational Psychology, 31*, 301-327.

Ryan, A.L. (2004). *Diagnostic accuracy of the Dynamic Indicators of Basic Early Literacy Skills in the prediction of first grade oral reading fluency.* Unpublished doctoral dissertation, University of Massachusetts, Amherst.

Samuels, S.J. (2007). The DIBELS test: Is speed of barking at print what we mean by reading fluency? *Reading Research Quarterly, 42,* 546-567.

Schilling, S.G., Carlisle, J.F., Scott, S.E., & Zeng, J. (2007). Are fluency measures accurate predictors of achievement? *Elementary School Journal, 107*, 429-448.

Shinn, M. R., Good, R. H., Knutson, N., Tilly, W. D., & Collins, V. L. (1992). Curriculum-based measurement reading fluency: A confirmatory analysis of its relation to reading. *School Psychology Review, 21*, 459-479.

Spector, J. (1992). Predicting progress in beginning reading: Dynamic assessment of phonemic awareness. *Journal of Educational Psychology, 84,* 353-363.

Speece, D.L., Mills, C., Ritchey, K.D., & Hillman, E. (2003). Initial evidence that letter fluency tasks are valid indicators of early reading skill. *The Journal of Special Education, 36,* 223-233.

Stanovich, K. (1986). Matthew effects in reading: Some consequences of individual differences in the acquisition of literacy. *Reading Research Quarterly, 21,* 360-407.

Swets, J.A. (1988). Measuring the accuracy of diagnostic systems. *Science, 240,* 1285-1293.

Swets, J.A. (1992). The science of choosing the right decision threshold in high-stakes diagnostics. *American Psychologist, 47,* 522-532.

Tanner, P.L. (2006). *Concurrent validity and diagnostic accuracy of curriculum based assessment: Comparing the DIBELS to the CTOPP.* Unpublished doctoral dissertation, Northern Arizona University, Flagstaff.

Trucksess, D. D. (2009). *An investigation of the predictive validity of kindergarten Dynamic Indicators of Basic Early Literacy Skills (DIBELS) to the Third Grade Pennsylvania System of School Assessment.* Unpublished doctoral dissertation, Bryn Mawr College, Bryn Mawr, PA.

Vloedengraven, J.M., & Verhoven, L. (2007). Screening of phonological awareness in the early elementary grades: An IRT approach. *Annals of Dyslexia,* 57, 33-50.

Wagner, R., Torgesen, J., & Rashotte, C. (1999). *Comprehensive Test of Phonological Processing.* San Antonio, TX: Pearson.

Walker, H. M. & Lev, J. (1953). *Statistical inference.* New York: Holt & Co.

Yopp, H.K. (1988). Validity and reliability of phonemic awareness tests. *Reading Research Quarterly, 23,* 159-177.

Zumeta, R.O., & Fuchs, L.S. (2009). *Using phoneme segmentation fluency screening to identifyreading risk: An evaluation of strengths, weaknesses, and avenues for future research.*Unpublished manuscript, Vanderbilt University, Nashville, TN.

Zumeta, R.O., Compton, D.L., & Fuchs, L.S. (2010). *Using Word Identification Fluency to assess first-grade reading development: A comparison of two word-sampling approaches.* Manuscript submitted for publication.