MEASURING THE ALPHABETIC PRINCIPLE:

MAPPING BEHAVIORS ONTO THEORY

by

KELLY M. LAUGLE

A DISSERTATION

Presented to the Department of Special Education
and Clinical Sciences
and the Graduate School of the University of Oregon
in partial fulfillment of the requirements
for the degree of
Doctor of Philosophy

September 2009

University of Oregon Graduate School

Confirmation of Approval and Acceptance of Dissertation prepared by:

Kelly Laugle

Title:

"Measuring the Alphabetic Principle: Mapping Behaviors onto Theory"

This dissertation has been accepted and approved in partial fulfillment of the requirements for the Doctor of Philosophy degree in the Department of Special Education and Clinical Sciences by:

Roland Good, Chairperson, Special Education and Clinical Sciences
Kenneth Merrell, Member, Special Education and Clinical Sciences
Leanne Ketterlin Geller, Member, Educational Methodology, Policy, and Leadership
Jean Stockard, Outside Member, Planning Public Policy & Mgmt

and Richard Linton, Vice President for Research and Graduate Studies/Dean of the Graduate School for the University of Oregon.

September 5, 2009

Original approval signatures are on file with the Graduate School and the University of Oregon Libraries.

An Abstract of the Dissertation of

Kelly Laugle for the degree of Doctor of Philosophy

in the Department of Special Education and Clinical Sciences

to be taken September 2009

Title: MEASURING THE ALPHABETIC PRINCIPLE: MAPPING BEHAVIORS ONTO THEORY

Approved: _____
Roland H. Good III, Ph.D.

Research suggests that development of the alphabetic principle is a critical factor in learning to recognize words and becoming a successful reader. The alphabetic principle encompasses both the understanding that relationships exist between letters and sounds and the application of these relationships to reading words. This study investigated the degree to which different measures of the alphabetic principle were predictive of later reading development. These measures were examined in the context of Ehri's phase theory of sight word development to investigate how different behaviors associated with the alphabetic principle fit within a developmental framework.

Two cohorts of students (109 kindergarteners, 212 first graders) participated in this study from spring of 2007 until late fall of 2008 (58 second graders, 121 third graders). The predictive powers of single and combined measures of the alphabetic

principle were analyzed using sequential regression. Results indicated that each measure explained significant between-student variation in performance on measures of word reading fluency, oral reading fluency (ORF), vocabulary, and reading comprehension. A measure of letter-sounds embedded in nonsense words appeared to have more utility for the prediction of reading outcomes than a measure of letter-sounds presented in isolation. Additionally, including a measure of nonsense words with a measure of letter-sounds embedded in nonsense words increased the predictive power of the model over and above the predictive power of letter sounds alone.

Growth on ORF served as an additional criterion for the purpose of investigating the methodology of measuring growth. Two conceptualizations of growth were explored: raw score change over time and individual rates of growth over time (slope). Correlations and sequential regression were used to evaluate the relationship between raw score change and measures of the alphabetic principle. Hierarchical Linear Modeling (HLM) was used to model individual slopes on Lexile measures of ORF (LORF). In general, raw score change appeared largely unrelated to measures of the alphabetic principle. HLM analyses revealed that individual differences in slope on LORF were minimal and not very reliable, making the prediction of these differences difficult. Recommendations for future research and implications for practice are discussed.

CURRICULUM VITAE

NAME OF AUTHOR: Kelly M. Laugle

GRADUATE AND UNDERGRADUATE SCHOOLS ATTENDED:

University of Oregon, Eugene, Oregon
North Carolina State University, Raleigh, North Carolina
Gonzaga University, Spokane, Washington

DEGREES AWARDED:

Doctor of Philosophy, School Psychology, 2009, University of Oregon
Master of Science, Psychology, 2006, North Carolina State University
Bachelor of Arts, Psychology and Art, 2001, Gonzaga University

AREA OF SPECIAL INTEREST:

Academic Assessment, Instruction, and Consultation

PROFESSIONAL EXPERIENCE:

School Psychologist Intern, North Clackamas School District, 2008-2009
Milwaukie, Oregon

Teach For America Corps Member, Vance County Public Schools, 2001-2003
Henderson, North Carolina

GRANTS, AWARDS AND HONORS:

Dynamic Measurement Group, Dissertation Research Award, University of Oregon, 2008

David Moursund Technology Scholarship, University of Oregon, 2008

Psychology Emeritus Award for Research, North Carolina State University, 2005

ACKNOWLEDGEMENTS

I would like to acknowledge the dedicated educators at Greater Albany Public Schools for their active support of and involvement in this project, especially Valerie Mullen and Wayne Goates who helped facilitate data collection in each phase of this study. I would like to thank my adviser and committee chair, Dr. Roland Good, for inspiring and challenging me to learn everything I can about improving early literacy outcomes for all children. I wish to express sincere appreciation to my committee members for their guidance and assistance in the preparation of this manuscript. Special thanks to Jean Stockard for joining my committee this spring. I would also like to recognize the DIBELS Student Research Team for their commitment to this project and for their support throughout my time at the University of Oregon. This investigation was funded in part by the DIBELS Student Support Award in the School Psychology program and the Dynamic Measurement Group Dissertation Research Award in the College of Education at the University of Oregon.

To Jack,

Sisu.

TABLE OF CONTENTS

Chapter Page

Chapter Page

LIST OF FIGURES

LIST OF TABLES

Table Page

CHAPTER I

STATEMENT OF THE PROBLEM

Teachers of reading have a tremendous responsibility to ensure that their students develop into successful readers. Reading is more than a skill; it is an avenue for empowerment in our society. We have the moral obligation to support teachers, in what ever way possible, to ensure that all students develop into empowered citizens. However, recent findings from the National Center for Educational Statistics suggest that we are not providing sufficient support to teachers of reading. Approximately 33% of fourth grade students in the United States are considered proficient readers based on the National Assessment of Educational Progress for 2007 (National Center for Education Statistics, 2007). Although this finding represents a 2% increase in the number of proficient readers from 2005 to 2007, improving literacy remains a significant challenge across the nation. One initiative aimed at improving literacy among students in kindergarten through third grade is Reading First, a component of the No Child Left Behind Act (2001). A central tenet of Reading First is the reliance on research to guide selection of reading curricula and reading assessment tools (P.L. 107-110, Title I, Part B, Subpart 1, 2002).

Research is a critical avenue of support to teachers. Research on reading helps teachers understand how reading skills develop, allows teachers to translate this understanding of reading development into effective practices to support all students, and leads to the development of scientifically-validated instructional programs and assessment tools for teachers to use. Although research is clearly not the only avenue of

support we should provide to our teachers, it is an avenue that is necessary to ensuring that all students develop into proficient readers. This study contributes to the knowledge base on reading research by exploring ways to measure the alphabetic principle (an important component of learning to read), mapping these measures onto a theory of word reading development, and examining the power of these measures for predicting later reading development. This study also addresses a methodological issue that is currently unresolved in the field of reading research: how to model growth in oral reading fluency. Continued research in the area of reading assessment allows teachers to use the best tools available for identifying students in need of additional instructional support.

The first section of this chapter will describe skills identified through research as essential to becoming a successful reader. The next section will focus more specifically on one of these skills, the alphabetic principle, and discuss the issues associated with measuring this construct. The final section will outline the purpose of this study and establish a set of questions to guide the research.

Essential Skills for All Readers

In response to a Congressional mandate, the National Reading Panel (NRP, 2000) reviewed research related to reading instruction in kindergarten through third grade and identified essential skill areas related to overall reading development. These skill areas are phonemic awareness, phonics, fluency, vocabulary, and comprehension. Phonemic awareness is the understanding that spoken words are comprised of individual speech sounds called phonemes. Phonics refers to the development of the alphabetic principle, or

the understanding that there are systematic and predictable relationships between letters and sounds, and that these relationships can be applied to reading words. Fluency is the ability to read a text quickly, accurately, and with proper expression. Vocabulary refers to the ability to understand words to acquire meaning from text, and comprehension refers to the purposeful and active interaction of the reader with the text. While each of the five skill areas should be addressed from kindergarten through third grade, the relative emphasis placed on each skill area should shift across these grades (Simmons & Kame'enui, 1999). This study focused on the development of phonics for students at the end of kindergarten and at the end of first grade because research suggests that this skill area is an important focus of instruction for these grade levels (NRP, 2000). *Phonics* is often referred to as an approach to reading instruction. The term *alphabetic principle* will be used instead of *phonics* to distinguish the construct as a skill to be measured, rather than an instructional approach.

The Alphabetic Principle

The alphabetic principle is a construct that refers to both cognitive processes (i.e. understanding that relationships exist between letters and sounds) and early-reading behaviors (i.e. application of these relationships to reading words). The first part of this section will provide a thorough discussion of the alphabetic principle, including the cognitive processes associated with its development, and describe the alphabetic principle's influence on later reading development. The second part of this section will focus on measuring the development of the alphabetic principle.

Understanding the Alphabetic Principle

Development of the alphabetic principle relies upon phonological processing, or "the use of phonological information (i.e., the sounds of one's language) in processing written and oral language" (Wagner & Torgesen, 1987, p. 192). Research on the phonological processes that influence the development of the alphabetic principle and research on the alphabetic principle's influence on later reading development provide a basis for understanding both the importance and complexities of assessing the development of the alphabetic principle.

Phonological Processing and the Alphabetic Principle

The phonological processes most closely tied to reading are *phonological awareness*, *phonological memory*, and *phonological naming* (Wagner et al., 1997). Figure 1 displays the three phonological processes involved in learning to read and the related skills that stem from these processes. *Phonological awareness* refers to the understanding that oral language is comprised of discrete sounds. Students develop phonological awareness by learning to isolate and manipulate progressively smaller units of speech, from syllables, to onsets and rimes, and eventually to phonemes, the smallest units of sound in language (Adams, 1990). Once students are able to isolate and manipulate individual phonemes within words, they are considered to have established *phonemic awareness*, a critical component of *phonological awareness*. *Phonological memory* refers to the coding of individual letters or parts of printed words into sound-based representations that are temporarily stored in working memory. When these sound-

based representations are held in working memory, they can be blended together and recoded as a whole word. *Phonological naming* refers to the rapid retrieval of phonological information stored in long-term memory. Efficient recall of letter-sound correspondences from long-term memory facilitates the blending of these sounds into words in working memory.

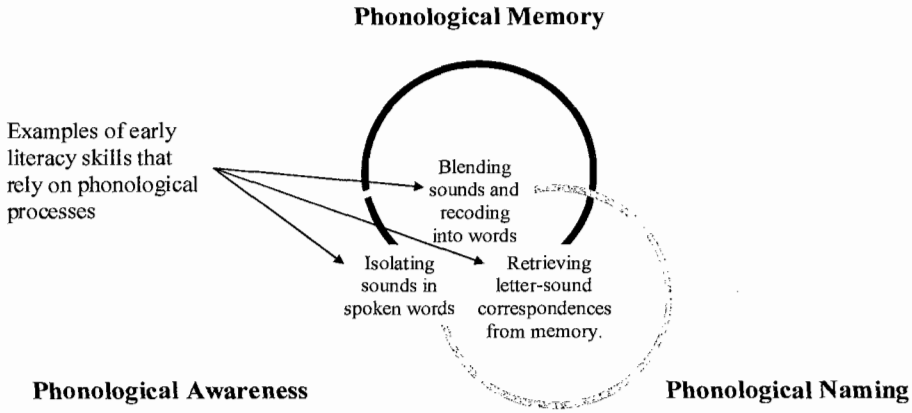


Figure 1. Phonological processes identified by Wagner et al. (1997) as critical to reading development and examples of early literacy skills that rely on these processes.

There is converging evidence that most reading difficulties stem from deficits in one or more areas of phonological processing that hinder the development of the alphabetic principle (Share & Stanovich, 1995; Torgesen & Burgess, 1998). For example, the National Reading Panel concluded that the establishment of phonemic awareness is causally related to the development of decoding skills and reading accuracy (NRP, 2000). Decoding, a critical behavior associated with the alphabetic principle, is the skill of

matching sounds to letters in words and blending those sounds together to form a whole word (Ehri & Roberts, 2006). Students who are unable to distinguish the individual sounds in spoken words are, consequently, unable to form connections between these sounds and their symbolic representations and to use those connections to decode words. Without firmly established letter-sound correspondences, the processes of phonological memory and phonological naming are limited by the reduced availability of accurate phonological information stored in long-term memory (Troia, 2004). Even when accurate phonological information is available in long-term memory, deficits in the rapid retrieval and subsequent manipulation of this information in working memory hinders the development of the alphabetic principle (Troia).

The Alphabetic Principle's Influence on Later Reading Development

Research suggests that development of the alphabetic principle is a critical factor in recognizing words and becoming a successful reader (Stanovich, 1986; Torgesen, 2002). When students fail to reach automaticity with applying the alphabetic principle to reading unfamiliar words, their reading remains slow and laborious (Ehri & Snowling, 2004). As a result, these students are exposed to less text than their peers, and the Matthew effect begins to take hold (i.e. the rich get richer and the poor get poorer) (Stanovich). In addition to reducing exposure to text, slow and laborious decoding of words occupies attentional resources which may prohibit students from effectively engaging in comprehending the text (Adams, 2001). Stanovich summarized the deleterious effects of poor decoding skills, "Thus, reading for meaning is hindered, unrewarding reading experiences multiply, and practice is avoided or merely tolerated

without real cognitive involvement. The downward spiral continues – and has further consequences" (p. 364).

Longitudinal research has provided evidence that early identification of, and intervention for, students who are struggling to develop skills associated with the alphabetic principle are critical to their success as readers (Francis, Shaywitz, Stuebing, Shaywitz, & Fletcher, 1996; Juel, 1988). Juel found, in a sample of 54 students, those students who were poor readers at the end of first grade had an 88% likelihood of remaining poor readers at the end of fourth grade. In addition, Francis and colleagues tracked the reading performance of 407 students from first through ninth grade and found that students who were struggling to learn how to read in first grade almost never caught up to their typically developing peers. Importantly, students' early decoding skills, assessed in first and second grade, accounted for 25% to 36% of the variability in comprehension scores at ninth grade.

Intervention research has also supported the effectiveness of early intervention for preventing and remedying deficits in the alphabetic principle (Torgesen, 2000). Early intervention that includes systematic and explicit instruction in phonemic awareness and the alphabetic principle for students identified as at-risk in kindergarten, first, and second grade, can result in all but a very small percentage of these students (2-6%) achieving commensurate with their typically-developing peers (i.e. above the 30th percentile on measures of word reading ability) (Torgesen, 2000). Systematic and explicit instruction in the alphabetic principle is beneficial because it helps students recognize the logic of

the alphabetic system and focus their attention on the critical relations between the alphabetic system and the sounds of our language (Adams, 2001).

Summary

Reading difficulties are likely to stem from phonological processing deficits that impair students' development of the alphabetic principle (Torgesen, 2002), and students who fail to develop the alphabetic principle during their initial years of schooling are likely to remain poor readers (Francis et al., 1996). Early intervention is critical for ensuring that students develop the alphabetic principle (Torgesen, 2002). While some students may develop the alphabetic principle with little to no formal instruction, other students may need systematic and explicit instruction (Stanovich, 1986). Regardless of the instructional path taken, educators must ensure that all students develop the alphabetic principle.

Measuring the Alphabetic Principle

Assessment tools that measure the alphabetic principle are critical for identifying students in need of additional instructional support. Scores derived from these tools can be compared to scores from a normative sample or to empirically-based criterion scores to identify students at-risk for later reading difficulties. Once identified, these students can begin receiving systematic and explicit instruction targeting the development of the alphabetic principle. Some of these assessment tools continue to play important roles during and after instruction by monitoring students' response to this instruction and by evaluating the overall outcomes of this instruction for preventing later reading difficulties.

The challenge of designing tools to assess development of the alphabetic principle is that cognitive processes are difficult to measure directly. Therefore, researchers must find observable behaviors that provide approximations for the development of these underlying processes and then find ways to measure these behaviors. Two behaviors that are typically measured are the production of letter-sounds from single or multiple letters or words in print (i.e. decoding) and the production of whole word units from words in print (i.e. recoding). Typically these words are nonsense words (e.g. ib, baf, shlee). Assessment tools that are currently in use measure these two types of behavior in slightly different ways. Multiple lines of evidence continue to be gathered to identify appropriate inferences and uses of the scores derived from these tools. This process reflects the science and art of test score validation.

Messick (1986) defined validity as "an overall evaluative judgment, founded on empirical evidence and theoretical rationales, of the *adequacy* and *appropriateness* of *inferences* and *actions* based on test scores" [emphasis in original] (p. 33). This definition, which is closely aligned with the definition of validity in the Standards for Educational and Psychological Testing (American Education Research Association (AERA), American Psychological Association (APA), & National Council on Measurement in Education (NCME), 1999), calls attention to both the inferences we draw from test scores and the ways is which we use the scores to make decisions. In the case of assessment tools that have been developed to measure the alphabetic principle, we *infer* that the behaviors being measured are in fact indicative of development of the alphabetic principle. We also *infer* that development of the alphabetic principle is

essential to learning how to read. Given these inferences, we may *use* the test scores to identify students in need of systematic and explicit instruction in the alphabetic principle. Various lines of validity evidence, both empirical and theoretical, have been gathered that address these inferences and the decisions that stem from them.

The next section will discuss a theoretical rationale for linking the development of the alphabetic principle to specific measures of behavior (providing an evidence base for the first inference). The following section will discuss how empirical evidence is gathered to link measures of the alphabetic principle to important reading outcomes (providing an evidence base for the second inference).

Theoretical Rationale for Linking the Development of the Alphabetic Principle to Measures of Behavior

Having a theoretical framework for understanding the relationship of specific behaviors to the development of the alphabetic principle allows the mapping of assessment tools onto a continuum of development based on the behaviors elicited by these tools. The following paragraphs will describe a theory for understanding the development of the alphabetic principle and how different measures of behavior can be mapped onto this theory.

Theory of early reading development. Theories of reading development help elucidate the phases students progress through on their route to becoming successful readers. Ehri (1999) created a model of the phases students typically experience as they develop sight vocabularies (i.e. words that are recognized automatically from memory). These phases involve progressively more sophisticated applications of the alphabetic

principle. While young students in the initial phase of sight word reading may recognize a few words by sight based on the shape of the letters or contextual cues (e.g. the golden arches for McDonalds), students in the more advanced phases rely primarily on the alphabetic principle for building their sight vocabularies. After frequent exposure to a word and successful decoding upon each exposure, the word (or chunks of the word) can be recognized automatically from memory without attentional resources being allocated to decoding (Ehri & Snowling, 2004).

Ehri's (1999) theory of sight word development. Ehri's (1999) phase theory of sight word reading includes four phases: *pre-alphabetic*, *partial alphabetic*, *full alphabetic*, and *consolidated alphabetic*. Young students in the *pre-alphabetic* phase of sight word development are not yet able to apply the alphabetic principle to identifying sounds within written words and typically rely on visual and contextual cues to guess at words (Ehri, 2005). Students in the *partial alphabetic* phase build connections between letters and sounds and begin to use those connections to attempt to read words. Students in this phase may produce only the first and final sounds in words, confuse similarly spelled words, and struggle to identify vowels and less frequently used letter-sounds (Ehri, 2005). These students may exhibit some signs of blending, but they are not yet able to fully decode words, and they may rely on partial phonetic cues to make predictions for unfamiliar words. Students in the *full alphabetic* phase develop the ability to decode words. Once students are able to form complete connections between all letter-sounds within a word, they can form connections to the corresponding pronunciation and meaning of the word. Once these connections between the word, its meaning, and its

pronunciation are well-established through repeated exposure, the word becomes part of the child's sight vocabulary (Ehri, 2005). Students in the *consolidated alphabetic* phase continue building their sight vocabularies for progressively more advanced types of words. This phase is characterized by chunking letter-sounds into larger units such as rimes, syllables, and morphemes that can be recognized automatically. Students will continue to use decoding as a strategy for attacking unfamiliar words, but they may also begin to identify more advanced words through the use of analogy (i.e., applying parts of known words like the rime *ight* in the word *night* to reading new words like *bright*) (Ehri, 2005).

Linking measures of the alphabetic principle to theory. Although Ehri's theory of sight word development is insufficient for understanding the entire process of reading development, the theory pinpoints subtle variations in behavior that are indicative of different phases of development of the alphabetic principle (e.g. identification of initial and final sounds within words, the blending of sounds together to form complete words, the partial decoding of words paired with the use of analogy, etc). Different assessments of the alphabetic principle that are currently in use appear to map onto different phases of Ehri's (1999) theory of sight word development beginning with the *partial alphabetic* phase. Figure 2 provides a visual representation of how different assessments might map onto progressively more advanced phases of development, and Table 1 outlines the hypothesized links between phases of development and measures of behavior.

Ehri's (1999) Phases of Sight Word Development

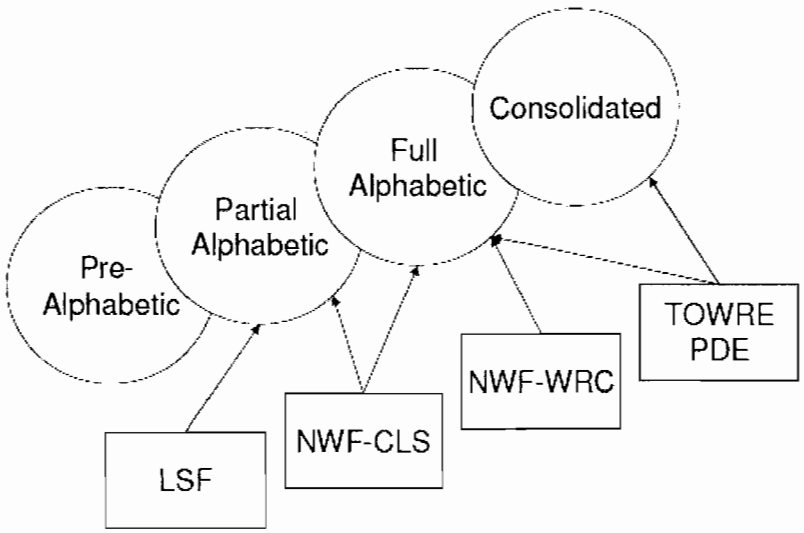


Figure 2. Mapping measures of the alphabetic principle onto Ehri's (1999) theory of sight word reading. LSF = Letter Sound Fluency. NWF = Nonsense Word Fluency. CLS = Correct Letter Sounds. WRC = Words Recoded Completely and Correctly. TOWRE PDE = Test of Word Reading Efficiency, Phonemic Decoding Efficiency subtest

As can been seen in Figure 2 and Table 1, a measure that requires students to produce individual letter-sounds that correspond to lower-case letters presented in isolation (e.g. Letter Sound Fluency (LSF) from AIMSweb Test of Early Literacy, Harcourt Educational Measurement, 2007) would appear to map onto the *partial alphabetic* phase because the measure requires only the production of individual letter sounds and not the decoding or recoding of whole words. A measure that presents decodable nonsense words that students can either read sound-by-sound or recode as whole words (e.g. Nonsense Word Fluency (NWF) from the Dynamic Indicators of Basic

Table 1

Hypothesized Links between Phases of Development and Measures of Behavior

Phase	Defining Characteristics	Behaviors Measured	Assessment
Partial Alphabetic	Building connections between letters and sounds. Beginning to use those connections to read words. Some blending but not able to fully decode yet.	Production of letter-sounds from letters presented in isolation OR Production of letter-sounds from nonsense words	LSF NWF (CLS score)
Full Alphabetic	Able to form complete connections between all letter-sounds within a word Form connections to the pronunciation and meaning of words	Recoding of letter-sounds into whole words (presented in the context of easily decodable nonsense words)	NWF (WRC score) PDE (nonsense words 1-14)
Consolidated	Building sight vocabularies for more advanced word types Able to chunk letter-sounds into larger units such as rimes, syllables, and morphemes Able to use advanced word attack strategies	Recoding of letter-sounds into whole words (presented in the context of more complex nonsense words requiring advanced decoding skills and knowledge of spelling patterns)	PDE (nonsense words 15-63)

Note. Phases are based on Ehir's (1999) theory of sight word development. LSF = Letter Sound Fluency NWF = Nonsense Word Fluency. CLS = Correct Letter Sounds. WRC = Words Recoded Completely and Correctly. PDE = Phonemic Decoding Efficiency.

Early Literacy Skills (DIBELS), Good & Kaminski, 2002) would appear to map onto the *partial* and *full* alphabetic phases because students could receive credit for each letter-sound produced and each nonsense word that was successfully recoded. A measure that presents both easily decodable nonsense words and more complex nonsense words (Phonemic Decoding Efficiency (PDE) from the Test of Word Reading Efficiency (TOWRE), Torgesen, Wagner, & Rashotte, 1999) would appear to map onto the *full* and *consolidated* phases because students would have to rely upon advanced knowledge of spelling patterns and more sophisticated decoding strategies.

When identifying measures indicative of the alphabetic principle, it is also important to consider similar measures which map onto Ehri's (1999) theory that may not be indicative of the alphabetic principle. For example, a measure that requires students to read a list of *real* words would appear to map onto the *full*, and *consolidated* alphabetic phases of Ehri's (1999) theory of word reading development, but scores from such a measure may not be indicative of development of the alphabetic principle. The distinction is based on the way in which the student may approach the task of *real* word reading differently from the way in which the student may approach the task of *nonsense* word reading.

On a *real* word reading task, students may automatically recognize the word from memory or they may apply one or more strategies for accessing words that may or may not involve some degree of decoding (Ehri & Snowling, 2004). Although decoding is likely to be the most reliable and frequently used strategy for determining unfamiliar words, a measure of real word reading provides no direct measure of whether students

employed this strategy and to what degree it was employed. Additionally, measures of *real* word reading often include irregular words that students are unlikely to be able to decode effectively and high frequency words that students are likely to recognize automatically. These factors compromise the validity of word reading measures as indicators of the alphabetic principle because the test user cannot assume that students relied on decoding strategies to perform the task. Measures of *nonsense* word reading, on the other hand, provide a direct assessment of students' development of the alphabetic principle because these measures *require* students to apply strategies of decoding to reading words that should not be familiar to students (Rack, Snowling, & Olson, 1992). Although the complex task of word reading may be an important skill to measure, it is better understood as a developmentally proximal criterion measure to measures of the alphabetic principle rather than as a direct measure of the alphabetic principle.

Linking Measures of the Alphabetic Principle to Important Reading Outcomes

Kame'enui, Good, and Harn (2005) stated that measuring the construct of reading "requires identifying the behavioral dimensions of reading with the most predictive power for determining later reading risk" (p. 70). The previous section described behavioral dimensions of the alphabetic principle, identified ways of measuring these behaviors, and described how these measures are hypothesized to map onto Ehri's (1999) phase theory of sight word development. Evaluating the utility of these measures for identifying students at-risk for later reading difficulties requires investigating the strength of their predictive relations for determining later reading development. Predictive relations can also provide evidence for how each of these measures fit within a

developmental framework (i.e. Ehri's (1999) phase theory of sight word development). For example, if multiple measures are administered to the same sample of students, then measures that map onto early phases of alphabetic development are hypothesized to hold less predictive strength than measures that map onto later phases of alphabetic development because the early phases are developmentally distal to reading outcomes compared to the later phases.

Investigations of predictive-related evidence explore the relationship between a predictor measure and a criterion measure and provide one type of evidence that contributes to an overall evaluative judgment of the predictor measure's validity (Messick, 1986). Estimates of predictive strength for measures of the alphabetic principle hold value only when these estimates derive from the prediction of criterion measures that are generally considered to accurately and meaningfully represent later reading development. Three indicators of later reading development that appear in the research literature are word reading fluency, fluency and accuracy with connected text, and reading comprehension. Growth on reading skills over time is also an important criterion to consider but one that is not yet well-defined or adequately measured. The following sections will define each criterion and identify existing approaches to measurement.

Word reading fluency. Word reading involves the automatic recognition of familiar words and the application of various strategies for determining unfamiliar words (Ehri & Snowling, 2004). Fluency measures of word reading typically involve reading a list of real words that include words with regular and irregular spellings as well as high-frequency words that are likely to be recognized automatically. These measures provide

an indication of the breadth of a student's sight vocabulary and the efficiency with which the student can apply strategies for determining unfamiliar words.

Fluency and accuracy with connected text. The NRP (2000) defined reading fluency as "the ability to read a text quickly, accurately, and with proper expression" (Chapter 3, p. 5). The phrase *fluency and accuracy with connected text* operationally defines the measurement of fluency as including only the first two elements of the NRP's definition (i.e. quickness and accuracy but not proper expression). Oral reading fluency (ORF), defined as the rate and accuracy of reading connected text out loud, is the most common behavior associated with the construct of fluency (Good, Simmons, et al., 2001). ORF is a strong indicator of overall reading competence and typically becomes valid for assessing students once they reach the middle of first grade (Fuchs, Fuchs, Hosp, & Jenkins, 2001). In summarizing Fuchs et al.'s review of the research supporting the use of ORF as a valid measure of overall reading competence, Kame'enui and Simmons (2001) stated:

> In practice, a high number of words read correctly per minute, when placed in proper developmental perspective, indicate efficient word-level processing, a robust vocabulary knowledge base, and meaningful comprehension of the text. In contrast, a low ORF rate suggests inefficient word recognition skills, a lean or impoverished vocabulary, and faulty text comprehension skills. (p. 208)

ORF indirectly reveals a student's competence at both the prerequisite skills necessary for developing reading fluency and the comprehension of text.

Reading comprehension. The NRP (2000) defined comprehension as dependent on "active and thoughtful interaction between the text and the reader" (Chapter 4, p. 11) and emphasized that the construct of comprehension cannot be understood without

considering the development of vocabulary knowledge. Measures of reading comprehension are typically group-administered, multiple-choice tests that include the assessment of vocabulary knowledge, sentence comprehension, and passage comprehension. Although ORF is considered to be a strong indicator of reading comprehension, including a more direct measure of comprehension and vocabulary knowledge as an additional criterion for evaluating measures of the alphabetic principle builds confidence in the assumption that the predictive criterion measures meaningfully represent important components of reading proficiency.

Growth. Growth in essential skills associated with reading is recognized as necessary and critical to becoming a successful reader; however, growth, as a construct to be measured, is not well established as either a predictor of later outcomes or as a criterion for evaluating initial status on measures of prerequisite skills. Baker et al. (2008) provided evidence that measuring growth on ORF contributes unique variance to the prediction of a measure of comprehensive reading performance above and beyond initial status on ORF. This research suggests that growth on ORF may be a valid and meaningful predictor of overall reading competence. It stands to reason that growth on ORF may also serve as a meaningful criterion for evaluating measures of early literacy. Initial status on measures of the alphabetic principle at the end of kindergarten and first grade is hypothesized to influence the rate at which students build fluency and accuracy with reading connected text.

Summary

Different lines of evidence continue to be gathered to identify appropriate *inferences* and *uses* of scores derived from assessment tools that measure the alphabetic principle. Mapping behaviors measured by these tools onto Ehri's theory of sight word development provides a theoretical rationale for inferring that scores from these measures reflect development of the alphabetic principle and allows for investigation of different facets of the alphabetic principle as a measurement construct. Examining the power of these measures for predicting later reading development provides empirical evidence to compare to the theoretical rationale and to evaluate the tools' relevance for identifying students at-risk of later reading difficulties.

Purpose of the Study

Continued research is needed to evaluate the quality of measures of the alphabetic principle so that teachers can use the best tools to identify students in need of additional instruction. A tool with strong predictive power for determining later reading development is a tool that can be used to accurately identify students who are most at-risk for experiencing later reading difficulties. These students often fail to develop the alphabetic principle, an essential skill for becoming a proficient reader (NRP, 2000). This study investigated the degree to which single and combined measures of the alphabetic principle were predictive of later reading development for students who were screened for reading difficulties at the end of kindergarten and at the end of first grade. Performance on these measures was examined in the context of Ehri's (1999) phase

theory of sight word development to build understanding of how different behaviors associated with the alphabetic principle fit within a developmental framework.

Predictor Measures

Behaviors indicative of the alphabetic principle that were examined in this study include (a) production of letter-sounds from letters presented in isolation, (b) production of letter-sounds from nonsense words, and (c) recoding of nonsense words. Measures of these behaviors included three timed, fluency-based measures: (a) LSF (Harcourt Educational Measurement, 2007), (b) NWF (Good & Kaminski, 2002), and (c) PDE, (Torgesen, Wagner, & Rashotte, 1999). Two scoring options were explored for the NWF measure: the number of correct letter-sounds produced in one minute (CLS) and the number of nonsense words that were recoded completely and correctly in that same minute (WRC).

Criterion Measures

The strength of predictive relations for measures of the alphabetic principle was determined by exploring their relationship to criterion measures of later reading development (word reading fluency, oral reading fluency, reading comprehension, and growth on oral reading fluency). Figure 3 provides a graphic representation of the predictor-criterion relationships that were explored in this study. Operational definitions for each criterion are provided in the following paragraphs.

Word reading fluency was defined as quickness and accuracy in reading a list of words out loud. Two measures of word reading fluency were administered in this study:

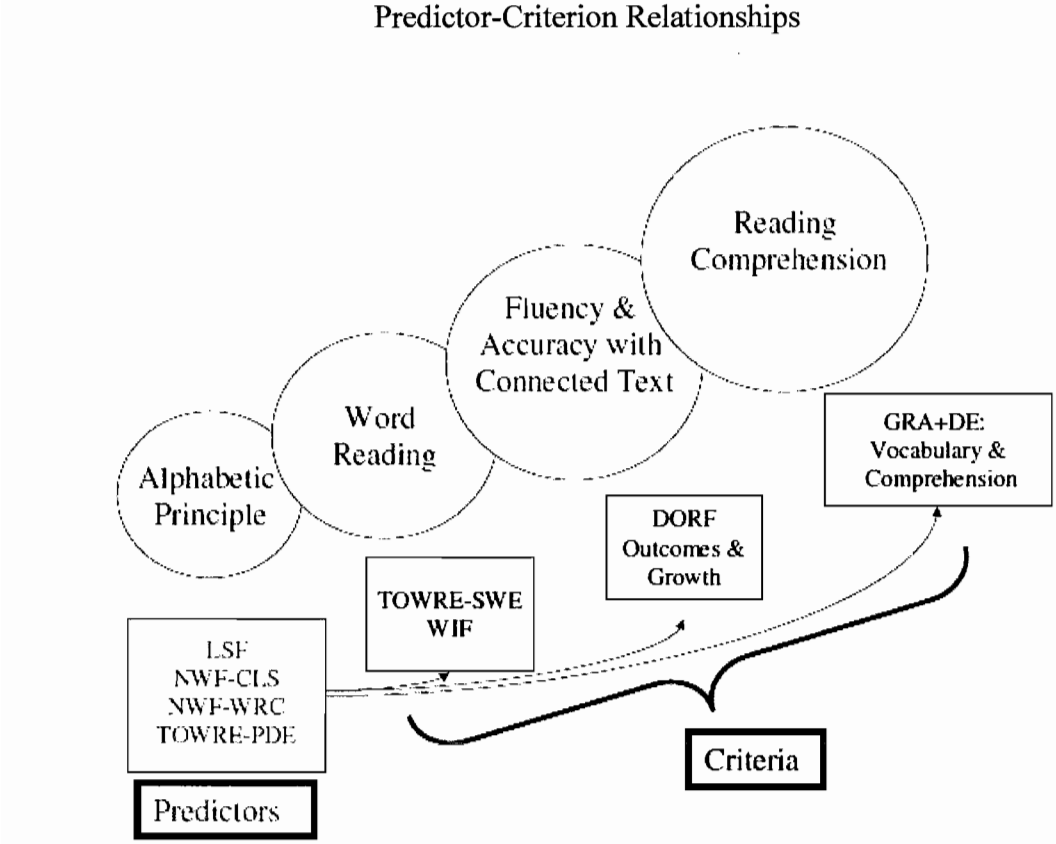


Figure 3. Graphic representation of predictor-criterion relationships. LSF = Letter Sound Fluency. NWF = Nonsense Word Fluency. CLS = Correct Letter Sounds. WRC = Words Recoded Correctly. TOWRE PDE = Test of Word Reading Efficiency, Phonemic Decoding Efficiency subtest. SWE = Sight Word Efficiency subtest. WIF = Word Identification Fluency. DORF = DIBELS Oral Reading Fluency. GRA+DE = Group Reading Assessment and Diagnostic Evaluation.

Sight Word Efficiency, a subtest of the TOWRE (Torgesen et al., 1999), and Word Identification Fluency (Deno, Mirkin, & Chiang, 1982), a curriculum based measure.

Fluency and accuracy with connected text was defined as quickness and accuracy in reading connected text out loud. The DIBELS ORF measure (Good, Kaminski, & Dill, 2002) was used to assess fluency and accuracy with connected text from the middle of first grade to the beginning of third grade.

Reading comprehension was defined as accuracy in completing multiple-choice questions pertaining to vocabulary knowledge, sentence comprehension, and passage comprehension. The Group Reading Assessment and Diagnostic Evaluation (GRA+DE, American Guidance Service, 2001) was used to assess these skill areas in second and third grade.

Growth was defined in two ways: (a) raw score change from one measurement occasion to the next on ORF and (b) rate of progress (slope) on ORF across multiple measurement occasions. This study explored the relationship between both definitions of ORF growth and measures of the alphabetic principle as a means of examining the methodology of measuring ORF growth.

Research Questions

Four research questions were investigated in this study. The first three questions address the prediction of scores on outcome measures of word reading fluency, fluency and accuracy with connected text, and reading comprehension at specific points in time.

The fourth research question addresses the prediction of growth on oral reading fluency over time.

Research Question One

How much between-student variation in word reading fluency, oral reading fluency, and reading comprehension is explained by initial performance on each measure of the alphabetic principle?

Research Question Two:

Does including a measure of letter-sounds in nonsense words (i.e. CLS) add significantly to the between-student variation explained in word reading fluency, oral reading fluency, and reading comprehension beyond a measure of letter-sounds in isolation (i.e. LSF)?

Research Question Three:

Does including a measure of nonsense word recoding (i.e. WRC, PDE) add significantly to the between-student variation explained in word reading fluency, oral reading fluency, and reading comprehension beyond a measure of letter-sounds (i.e. LSF, CLS)?

Research Question Four:

When predicting change/growth on repeated measures of oral reading fluency over time: (a) How much between-student variation is explained by initial performance on each measure of the alphabetic principle? (b) Does combining a measure of letter-sounds in nonsense words with a measure of letter-sounds in isolation add significantly to

the variance explained? (c) Does combining a measure of nonsense word recoding with a measure of letter-sounds add significantly to the variance explained?

CHAPTER II

LITERATURE REVIEW

The first section will review research on brief, fluency measures of behaviors indicative of the alphabetic principle (i.e. producing letter-sounds from letters in isolation, producing letter-sounds from nonsense words, and recoding nonsense words). The second section will review research addressing the modeling of oral reading fluency (ORF) growth in statistical analyses.

Review of Research on Measuring the Alphabetic Principle

This section will review existing lines of validity evidence for scores derived from assessment tools designed to measure the alphabetic principle for the purpose of identifying students at-risk for later reading difficulties. Most studies focused on concurrent and predictive-related evidence for determining relationships with existing criterion measures. Concurrent correlation coefficients provide evidence for the strength of relationships between scores from different tools that are administered at the same point in time. Predictive correlation coefficients provide evidence for the degree to which scores from one measure are related to scores from another measure that is developmentally and/or temporally distal from the first measure. A unique study conducted by Harn, Stoolmiller, and Chard (2008) will also be reviewed. This study explored the utility of deriving multiple scores from a single measure of the alphabetic principle (NWF) for improving predictive strength.

Letter-Sound Fluency

Studies that investigated the concurrent and predictive-related evidence of scores from Letter-Sound Fluency (LSF) will be reviewed. This review will include two studies of LSF in first grade and three studies of LSF in kindergarten. Table 2 presents the reliability and validity coefficients reported by each study and the criterion measures that were selected for comparison.

Speece and Case (2001) designed a 1-minute, measure to be used in first grade for screening students at-risk for reading disabilities. The probes were developed by randomly arranging the 26 lower-case letters of the alphabet in two columns on a page. If a student completed the probe before time had expired, the student was directed to begin again at the top of the page. Two probes were administered, and the score was the average number of letter sounds produced in one minute across the two consecutive trials. Reliability and validity estimates are presented in Table 2. Speece and Case's sample may not reflect the typical distribution of first grade performance because the sample consisted of 74 students who were identified as at-risk (i.e. mean performance on LSF below the 25th percentile in their classrooms), and 64 students who represented a purposive sample (i.e. 5 students selected from each class based on mean LSF performance; 2 students at the median and 1 each at the 30th, 75th and 90th percentiles).

Using the same measure of LSF and the same sampling method as Speece and Case (2001), Speece and Ritchey (2005) investigated the predictive strength of LSF for determining ORF growth and outcomes for 276 first grade students (140 at-risk students). As seen in Table 2, concurrent and predictive correlations for the combined sample (at-

Table 2

Reliability and Validity Evidences for Measures of Letter-Sound Fluency

Study	Sample/Time of Year	Reliability Coefficient	Reliability Type	Validity Coefficient	Criterion Measure	Validity Type
Elliot, Lee, & Tollefson, 2001	75 kindergarten students / Spring	.82	Interrater	.58	WJ-R ACH: BR	Concurrent
		.83	Test Retest	.72	WJ-R ACH: Skills	
		.82	Alternate Form	.62	WJ-R ACH: LWID	
Ritchey & Speece, 2006	92 kindergarten students / Winter	_	_	.50	WRMT: WI	Concurrent
				.65	WRMT: WI	Predictive (Win. to Spr.)
				.65	TOWS-4	Predictive (Win. to Spr.)
Speece & Case, 2001	138 1st grade students (74 at-risk) / Fall	.93	Alternate Form	.66	WJ-R: BRC	Predictive (Fall to Spr)
Speece & Ritchey, 2005	276 1st grade students (140 at-risk) / Fall & Winter	_	_	.03 *ns*[a]	WRE	Concurrent (Fall)
				.24	WJ-R: BRC	Predictive (Fall to Spr.)

Table 2 (continued)

Study	Sample/Time of Year	Reliability Coefficient	Reliability Type	Validity Coefficient	Criterion Measure	Validity Type
				.25	ORF	Concurrent (Win.)
				.03 *ns*	ORF	Predictive (Fall to Win)
Stage, Sheppard, Davidson, & Browning, 2001	59 kindergarten students / Spring	98% agreement	Interrater	.72, .77, .73, .71	ORF-	Predictive (Spr K to Oct, Jan., Mar., May 1st)

Note. Dashes indicate reliability was not evaluated by the study. WJ-R ACH = Woodcock Johnson Tests of Achievement-Revised; BR = Broad Reading Score. Skills = Basic Skills Cluster Score. LWID = Letter Word Identification (Woodcock & Johnson, 1989). WRMT = Woodcock Reading Mastery Test- Revised. WI = Word Identification subtest (Woodcock, 1988). TOWS-4 = Test of Written Spelling-Fourth Edition (Larsen, Hamill, & Moates, 1999).WJ-R: = Woodcock-Johnson Psychoeducational Battery-Revised. BRC = Basic Reading Cluster Score (Woodcock & Johnson, 1989).WRE = Word Reading Efficiency subtest of the prepublication TOWRE (Torgesen, Wagner, & Rashotte, 1999). ORF = Oral Reading Fluency.

[a]Correlation for at-risk sample only.

All correlations significant at $p < .05$ unless noted with *ns* (non-significant correlation).

risk and purposive) were low (.24, .25) or not significant. Interestingly, correlation with the same criterion measure (Basic Reading Cluster, Woodcock Johnson Psychoeducational Battery-Revised, Woodcock and Johnson, 1989) across the same time span (fall to spring of first grade) was .66 in Speece and Case's study and .24 in Speece and Ritchey's study. Descriptive statistics indicate that Speece and Ritchey's sample had a greater percentage of students identified as at-risk (57% compared to 54%), and the at-risk sample in Speece and Ritchey's study had lower mean performance on LSF and ORF than the sample from Speece and Case's study. Standard deviations for mean performance indicate floor effects on LSF for both studies, but floor effects appear more pronounced in Speece and Ritchey's sample which could explain the difference in correlation coefficients.

Although the relation between LSF scores and later performance was low in Speece and Ritchey's study, they found risk status, based on fall LSF performance (mean score below the 25th percentile in class), to be a significant predictor of ORF performance at the end of the year and slope of progress on ORF over the year. The remaining analyses focused specifically on the predictive strength of LSF for the at-risk sample. After accounting for initial status on ORF in January of first grade, LSF in January explained 1.3% additional unique variance in ORF performance at the end of first grade and 3.1% additional unique variance in ORF from January to May. However, LSF in winter of first grade was not a significant predictor of end-of-year ORF performance or growth in second grade after accounting for ORF performance in January of first grade and ORF growth during first grade. These findings suggest that LSF may be a meaningful

predictor of reading performance within first grade for students identified as at-risk, but the unique predictive strength of LSF is unlikely to extend beyond first grade when ORF is also administered.

The predictive strength of LSF in kindergarten was also investigated. Ritchey and Speece (2006) administered the same LSF measure developed by Speece and Case (2001) to 92 students in the winter of kindergarten. A measure of word reading (Word Identification) was administered in the winter and spring of kindergarten, and a measure of spelling (Test of Written Spelling-Fourth Edition) was administered in the spring of kindergarten. Ritchey and Speece found that LSF in the winter of kindergarten contributed 5.6% unique variance to word reading in the spring after accounting for word reading in the winter. Additionally, LSF in the winter contributed 42.3% of the variance to spelling scores in the spring. These findings suggest that LSF provides unique information in the winter of kindergarten for the prediction of word reading and spelling at the end of kindergarten.

Stage, Sheppard, Davidson, and Browning (2001) used hierarchical linear modeling and hierarchical multiple regression to evaluate the predictive strength of LSF in comparison to Letter Naming Fluency (LNF), a similar measure that requires students to produce the name of the letter instead of its sound. Both measures were administered to a sample of 59 students at the end of kindergarten. When each measure was entered into a growth curve model as an individual predictor, both LNF and LSF contributed to the prediction of ORF growth during first grade and ORF outcomes at the end of first grade. When entered simultaneously, both measures contributed to the prediction of ORF

outcomes, but only LNF contributed unique and significant variance to ORF growth. Additionally, when the first ORF measurement, in the fall of first grade, was entered as a predictor into a hierarchical multiple regression analysis, only LNF made unique contributions to the prediction of ORF growth in first grade. These results suggest that, at the end of kindergarten, LNF may serve as a more powerful predictor of first grade ORF than LSF. Stage et al. investigated these results further by examining the score distributions for the predictors and outcome measure. Stage et al. noted that both LSF at the end of kindergarten and ORF at the beginning of first grade were significantly positively skewed, where as LNF at the end of kindergarten was more normally distributed. Twelve percent of the students scored a zero on LSF, and 25% of students scored a zero on ORF. These results suggest that LSF and ORF may be impacted by floor effects at these points in time that prevent the accurate identification of students most at-risk for reading difficulties.

Elliot, Lee, and Tollefson (2001) evaluated the reliability and concurrent-related evidence of Sound Naming Fluency, a measure developed for a modified version of the DIBELS, for a sample of 75 students at the end of kindergarten. This measure included both upper and lower case letters presented in random order in eleven rows of ten letters. Reliability coefficients ranged from .82 to .83, and validity coefficients ranged from .58 to .72 with the Basic Reading subtests from the Woodcock Johnson Tests of Achievement, Revised (Woodcock & Johnson, 1989) suggesting initial support for the use of SNF as a screener at the end of kindergarten.

In summary, LSF (and a similar measure, SNF) administered in kindergarten and first grade has evidence supporting both its reliability and validity for use as a measure of the alphabetic principle. Reliability coefficients range from .82 to .93 across studies suggesting that the measure is reliable enough to be used as a screener and potentially for individual decision making (Salvia & Ysseldyke, 2003). Validity coefficients for LSF in kindergarten range from .50 to .77 with measures of word identification, nonsense word reading, spelling, and oral reading fluency (Elliot et al., 2001; Ritchey & Speece, 2006; Stage et al., 2001). Validity coefficients for first grade were less consistent across studies with one study reporting .24 correlation with the Basic Reading Cluster score from the WJ-R in the spring of first grade (Speece & Ritchey, 2005) and another study reporting .66 correlation with the same cluster score across the same time span (Speece & Case, 2001) suggesting that additional evidence is needed to support the use of LSF as an effective screening tool in first grade. Research suggests that LSF in the fall of both kindergarten and first grade explains a unique and significant percent of the variance in reading outcomes at the end of the same year (Ritchey & Speece, 2006; Speece & Ritchey, 2005), but this unique predictive strength is not likely to extend to second grade when measures of ORF are also used as predictors (Speece & Ritchey, 2005). Additionally, the unique predictive strength of LSF from fall to spring of the same grade level may be overshadowed by other measures that are commonly administered in kindergarten and first grade such as Letter Naming Fluency (Stage et al., 2001).

Nonsense Word Fluency

Research addressing the concurrent and predictive-related evidence of the DIBELS Nonsense Word Fluency (NWF) measure will be reviewed in the following paragraphs. The letter-sound scoring procedure for NWF, which gives credit for each correct letter-sound (CLS) that the student identifies, regardless of whether the student reads sound-by-sound, partially decodes, or recodes a word, is unique to this specific measure, and most research-to-date has focused exclusively on this scoring method for NWF. Reliability and validity coefficients from the DIBELS Technical Adequacy Report (Good et al., 2004) and for each study that is reviewed are summarized in Table 3.

Speece, Mills, Ritchey and Hillman (2003) evaluated the degree to which NWF uniquely contributed to the prediction of reading outcomes after accounting for other significant predictors (i.e. performance on measures of phonological awareness (PA) and letter-name knowledge(LNK)) for a sample of 40 students who were assessed at the end of kindergarten and again at the end of first grade. The PA measure assessed students' ability to combine orally presented syllables, onset-rimes, and phonemes and to produce a word while omitting a selected phoneme. The LNK measure assessed students' accuracy in naming 10 upper-case letters. At the end of kindergarten, NWF explained 34.9% of the variance in performance on the Letter-Word Identification (LWID) subtest of the WJ-R after accounting for performance on PA and LNK which explained 50% of the variance. Of the three measures administered at the end of kindergarten (LNF, NWF, and PA), PA was the only measure to contribute significant variance to LWID at the end of first grade. For the Word Attack subtest of the WJ-R, which was also administered at the end of first

Table 3

Reliability and Validity Evidences for DIBELS Nonsense Word Fluency (Correct Letter-Sounds Score)

Study	Sample/Time of Year	Reliability Coefficient	Reliability Type	Validity Coefficient	Criterion Measure	Validity Type
Chard, Stoolmiller, Harn, Wanzek, Vaughn, Linan-Thompson et al., 2008	668 1st grade students / Spring	_	_	.66, .65, .59	ORF	Concurrent & Predictive (Spr. 1st to Spr. 2nd & Spr. 3rd)
				.29, .33	SAT-10 Comp, Vocabulary	Predictive (Spr. 1st to Spr. 3rd)
Good, Kaminski, Shinn, Bratten, Shinn, Laimon et al., 2004	53-148 1st grade students / Winter	.83 (n = 148)	Median 1-month alternate form	.51 (n = 123)	WJ-R: RC	Concurrent
				.81 (n = 146)	ORF	Predictive (Win. to Spr.)
				.68 (n = 53)	ORF	Predictive (Win 1st to Spr. 2nd)
				.66 (n = 107)	WJ-R: TRC	Predictive (Win 1st to Spr. 2nd)
Good, Baker, & Peyton, 2009	358,032 1st grade students / Fall & Winter	_	_	.74, .69	ORF	Predictive (Fall to Win., Spr)
				.73	ORF	Concurrent (Win)

Table 3 (continued)

Study	Sample/Time of Year	Reliability Coefficient	Reliability Type	Validity Coefficient	Criterion Measure	Validity Type
				.73	ORF	Predictive (Win to Spr.)
Riedel, 2007	1,518 1st grade students / Fall, Winter & Spring	_	_	.45, .45	GRA+DE	Predictive (Fall,Win to Spr)
				.46	GRA+DE	Concurrent (Spr.)
				.39. .38, .37	TerraNova	Predictive (Fall, Win, Spr 1st to Spr 2nd)
Speece, Mills, Ritchey, & Hillman, 2003	40 students assessed in spring of kindergarten and again in spring of 1st grade	_	_	.91, .71	WJ-R: LWID	Concurrent (Spr. K, Spr 1st)
				.59	WJ-R: LWID	Predictive (Spr. K to Spr 1st)
				.75	WJ-R: WA	Concurrent (Spr 1st)
				.59	WJ-R: WA	Predictive (Spr K to Spr 1st)

Table 3 (continued)

Study	Sample/Time of Year	Reliability Coefficient	Reliability Type	Validity Coefficient	Criterion Measure	Validity Type
				.74	ORF	Concurrent (Spr 1st)
				.71	ORF	Predictive (Spr. K to Spr. 1st)

Note. Dashes indicate reliability was not evaluated by the study. ORF = Oral Reading Fluency. SAT-10 = Stanford Achievement Test , 10th Edition (Harcourt Educational Measurement, 2002).WJ-R: = Woodcock-Johnson Psychoeducational Battery-Revised; RC = Readiness Cluster; TRC = Total Reading Cluster; LWID = Letter Word Identification; WA = Word Attack (Woodcock & Johnson, 1989) WRMT = Woodcock Reading Mastery Test- Revised; WI = Word Identification; WA = Word Attack (Woodcock, 1988). CRAB = Comprehensive Reading Assessment Battery (Fuchs, Fuchs, & Hamlett, 1989). GRA+DE = Group Reading Assessment and Diagnostic Evaluation (American Guidance Service, 2001). TerraNova = Reading subtest of the TerraNova, Second Edition (CTB/McGraw-Hill, 2003).

All correlations significant at $p < .05$.

grade, kindergarten PA contributed 70% of the variance and kindergarten NWF contributed an additional 5.8% of the variance in performance. For ORF in first grade, NWF in kindergarten contributed an additional 11.8% of the variance beyond the 56% accounted for by PA. These finding suggest that NWF, administered in the spring of kindergarten and the spring of first grade, provides unique and significant contributions to the prediction of reading outcomes beyond that provided by measures of phonological awareness and letter name knowledge.

Good, Baker, and Peyton (2009) critically examined the predictive strength of NWF administered at the beginning and middle of first grade for determining ORF outcomes at the middle and end of first grade. Based on the DIBELS Data System sample of 358,032 first grade students from 44 states in the U.S. and Canada, correlations between NWF and ORF ranged from .69 to .74. Initial risk status on NWF in the fall (at-risk = scores from 0-12, some risk = scores from 13-23, low-risk = scores above 23) was found to be an extremely strong predictor of first grade ORF outcomes, explaining 44% of the variance in ORF performance at the end of first grade. Given risk category, initial score on NWF explained 6% additional variance in ORF scores.

Concerned with over reliance on ORF as a criterion for evaluating the DIBELS measures, Riedel (2007) used measures of reading comprehension to examine the predictive validity of the DIBELS measures, including NWF, administered in the fall, winter, and spring of first grade. Validity coefficients for NWF are presented in Table 3 and range from .45-.46 for a comprehension measure administered at the end of first

grade and .37-.39 with a different comprehension measured administered at the end of second grade.

Chard et al. (2008) provided additional evidence for the predictive strength of NWF for determining performance on both ORF and measures of reading comprehension. Chard et al. tracked the performance of 668 kindergarten and first grade students through third grade to identify the critical student variables that predicted later reading outcomes within the context of schools implementing multi-tiered, school-wide reading interventions. These students were identified as needing strategic or intensive reading interventions based on the DIBELS decision rules (Good, Simmons, Kame'enui, Kaminski, & Wallin, 2002) and were assessed on a number of early reading measures in the spring of first grade. Validity coefficients for NWF with ORF in first through third grade and SAT-10 at the end of third grade are presented in Table 3. The Word Attack and Word Identification subtests of the WRMT-R, which are untimed tests that require students to recode nonsense words (Word Attack) and read real words that increase in difficulty (Word Identification), were also correlated with ORF and SAT-10. Correlation coefficients with ORF were similar across the three tests (NWF: .59-.66; WA: .53-.58; WI: .54-.62). Correlation coefficients for the Comprehension and Vocabulary subtests of the SAT-10 were somewhat stronger for the WA and WI subtests than for NWF (NWF: .29, .33; WA: .43, .48; WI: .41, .45). These data suggest that NWF at the end of first grade holds comparable predictive relations to published, standardized measures of early reading for predicting ORF across three grade levels, but NWF may hold slightly less predictive strength for determining reading comprehension at the end of third grade.

In summary, the DIBELS NWF measure (CLS score) has evidence supporting both its reliability and validity for use as a screening tool in first grade. Median alternate form reliability is .83 suggesting that a single administration of NWF is appropriate for use as a screener (Salvia & Ysseldyke, 2003). Reliability is estimated to improve to .94, a level acceptable for individual decision making, when the median score is taken from three administrations of NWF (Good et al., 2004). Estimates of predictive strength are moderate to strong for criteria that are developmentally and/or temporally proximal to NWF (i.e. measures of word reading (.59-.91), decoding (.59-.75), and ORF in first and second grade (.65-.81)) and moderate to low for criteria that are developmentally and temporally distal (comprehensive assessments of reading (.29-.46) and ORF in third grade (.59)). Research suggests that NWF contributes unique and significant variance to the prediction of reading outcomes beyond measures of phonological awareness and letter name knowledge (Speece et al; 2003), and NWF holds comparable predictive strength to published, standardized measures of decoding like the TOWRE-PDE and Word Attack (Chard et al., 2008). Additional research is needed to evaluate the predictive strength of NWF administered in kindergarten.

Nonsense Word Recoding Fluency

The following paragraphs will review research on nonsense word measures that require students to read whole words (i.e. recode) to receive credit. Research on the use of nonsense word reading measures to identify students with reading difficulties will be reviewed. Additionally, two measures of nonsense word reading and the Words Recoded

Completely and Correctly (WRC) scoring option on the DIBELS NWF measure will be described.

Rack, Snowling and Olson (1992) conducted a review of research on dyslexia and found that, in a majority of studies employing measures of nonsense word reading as outcomes, students with reading difficulties who were matched based on reading-level with younger peers showed significant deficits in nonsense word reading. These deficits ranged from 9 to 43 percentage points below the performance of the reading-level-matched group with a median deficit of 19 percentage points. Yet these students performed comparably to their reading-level-matched peers on other tasks of reading that were related to orthography instead of phonology. More recently, Torgesen, Rashotte, and Alexander (2001) found that measures of nonsense word reading provided unique contributions to the prediction of oral reading fluency (as measured by the Gray Oral Reading Test – Revised) beyond the variance explained by measure of real word reading for students experiencing reading difficulties who participated in remediation or prevention-based reading interventions. Rack et al.'s (1992) review and Torgesen et al.'s (2001) research provide evidence that measures of nonsense word reading contribute unique and important information for the identification of students with reading difficulties.

There are two published, standardized tests of nonsense word reading that frequent the research literature: the Phonemic Decoding Efficiency (PDE) subtest of the TOWRE and the Word Attack (WA) subtest of the WRMT-R (also included in the Woodcock Johnson Psychoeducational Battery and Tests of Achievement). Both subtests

begin with easily decodable two-letter words, and the difficulty of the words gradually increases requiring students to engage in progressively more sophisticated word analysis and to apply known spelling patterns to the accurate decoding of words. PDE is a 45-second test, and WA is not timed; however students are given five seconds on WA to identify a word before the administrator prompts the student to go on to the next word. Validity coefficients for the two subtests ranges from .85 to .91 (Torgesen et al., 1999). Given the focus on fluency-based measures of the alphabetic principle and PDE's strong correlation with WA, only PDE was investigated in this study. Table 4 provides a summary of the reliability and validity evidences for PDE taken from the TOWRE Examiner's Manual (Torgesen et al., 1999). Reliability coefficients for students ages 6 to 9 range from .90 to .97 providing support for the use of PDE for individual decision making (Salvia & Ysseldyke, 2003).

The Words Recoded Completely (WRC) scoring option on the NWF measure may hold similar predictive strength to PDE and other nonsense word measures for determining later reading outcomes, but research is needed to explore its predictive strength relative to other measures of nonsense word recoding (i.e. PDE) and its unique predictive strength beyond the variance in performance that can be explained by the Correct Letter Sounds score. To receive credit for the WRC scoring option, students must read the nonsense word as a whole word. Students may read the word sound-by-sound prior to recoding and still receive credit for the WRC score. Having a measure that provides information on both the number of letter-sounds correctly produced and the number of words correctly recoded may provide useful instructional information on the

Table 4

Reliability and Validity Evidences for the Phonemic Decoding Efficiency Subtest of the Test of Word Reading Efficiency (Torgesen, Wagner, & Rashotte, 1999)

Sample	Reliability Coefficient	Reliability Type	Validity Coefficient	Criterion Measure	Validity Type
1500 individuals from 30 states, ages 6-24	.86-.97[a]	Alternate form	_	_	_
72 individuals, ages 6-24	.82-.97[b]	2-week test-retest	_	_	_
145 1st grade students	_	_	.85	WRMT: WA	Concurrent (Spr 1st)
125 1st-3rd grade students at risk for reading failure	_	_	.89-.91	WRMT: WA	Concurrent

Note. Dashes indicate reliability was not evaluated by the study. WRMT = Woodcock Reading Mastery Test- Revised; WA = Word Attack (Woodcock, 1988).

[a]Reliability coefficients for 6 and 7 year olds in the sample ($n = 295$) ranged from .95 to .97.

[b]Reliability coefficient for 6 to 9 year olds in the sample ($n = 29$) was .90.

All correlations significant at $p < .05$.

degree to which students have transitioned from the *partial* to the *full* or *consolidated* phases of alphabetic development.

Deriving Multiple Scores from a Single Measure

Harn, Stoolmiller, and Chard (2008) developed a set of scoring procedures to derive four unique scores from NWF beyond the traditional CLS score. Each score reflects a slightly different approach to reading the nonsense words, and Harn et al. hypothesized that the dominant approach employed by the student would indicate which of Ehri's phases of alphabetic development the student had reached. The different approaches that were identified are (a) reading sound-by-sound (*sound-only*), (b) reading sound-by-sound and then recoding (*recode*), (c) partially blending sounds (*partial blends*), and (d) reading the word as a unit without any attempt at sounding out or partially blending (*unit*). Proportion scores were derived from the occurrence of each of these approaches to nonsense word reading. For example, if, upon presentation of 10 nonsense words, a student did not read any of the words sound-by-sound, partially blended four of the words, and read the remaining six words as units, then the student's proportion-scores across the four behaviors would be 0, 0, .40, and .60 respectively.

Harn et al. (2008) evaluated the predictive-related evidence of the four proportion-scores (derived from completed NWF protocols for 109 students in winter of first grade) for determining ORF performance in winter and spring. The *unit* strategy was positively correlated with ORF in winter (.57) and spring (.66). The remaining strategies were negatively correlated with ORF in winter and spring, except for *partial blends* with spring ORF, (*partial blends*: -.06 (winter), .06 (spring); *recode*: -.35, -.43; *sound only*: -

.40, -.49). Across several models, the *unit* strategy consistently contributed to the prediction of ORF performance. The *partial blend* strategy contributed predictive strength for spring ORF after controlling for winter ORF, and fall and winter NWF, but the *partial blend* strategy did not appear to provide any additional predictive strength beyond the *unit* strategy. Both the *unit* and the *partial blend* strategies indicate that the student is exhibiting some degree of unitizing (recognizing chunks and whole words automatically) and the distinction between the degree of unitization may be useful for teachers using the measure to inform their instruction, however the evidence suggests that measuring just the *unit* strategy is sufficient for predicting later performance on ORF. Neither the *recode* nor the *sound-only* strategy contributed uniquely to the prediction of spring ORF after controlling for winter ORF, fall and winter NWF, and their interaction. It appears likely that the timed nature of NWF would account for differences in student performance that may have resulted from recoding or reading sound-by-sound because students who predominantly use these strategies are likely to attempt fewer words during the one minute test than students who predominantly use more efficient strategies like unitizing and partial blending. Again, coding the use of *recode* and *sound-only* strategies may provide useful instructional information, as indicated by Harn et al., but doing so does not appear to add to the predictive strength of NWF in winter for ORF in spring.

Additionally, Harn et al. (2008) reported that 53% of the students used multiple strategies during the one-minute measure which calls into question the assumption that dominant strategy usage can be reliably measured in one minute. Typically, students who are at benchmark in fall of first grade will attempt only 9 or 10 words on NWF. If over

half of the students used two, three, or four different strategies to attack 9 or more words over the course of one minute, then this sampling of behavior may not provide a reliable indication of a student's dominant approach to reading nonsense words.

In summary, Harn et al. (2008) provided initial support for the measurement of subtle distinctions in behaviors associated with the alphabetic principle, and the evidence suggests that at least one of these distinctions (unitizing or not unitizing) can improve upon the power of a measure of the alphabetic principle for predicting later reading performance, if it can be measured reliably. However, Harn et al. examined only one tool for measuring the alphabetic principle (NWF) which restricted the measurement of behaviors to the context of two and three-letter nonsense words. Expanding the analysis to include letters presented in isolation and more complex nonsense words may provide additional valuable information for mapping behaviors onto Ehri's phases of alphabetic development.

Summary of Research on Measuring the Alphabetic Principle

LSF, NWF (CLS score), and PDE have concurrent and predictive-related evidence supporting their use as measures of the alphabetic principle and indicators of later reading development. Research suggests LSF may hold greater predictive strength in kindergarten than in first grade but may not provide unique information for predicting later reading outcomes when other measures are also administered (i.e. LNF) (Elliot et al., 2001; Ritchey & Speece, 2006; Speece & Case, 2001; Speece & Ritchey, 2005; Stage et al., 2001). Additional research is needed to explore the predictive strength of LSF in kindergarten relative to its predictive strength in first grade and its unique contributions

to predicting later reading outcomes beyond other measures of the alphabetic principle. Additional research is also needed to explore the validity of NWF for use as a screener in kindergarten; currently only one study has examined NWF in kindergarten (Speece et al., 2003). Research on the Words Recoded Correctly (WRC) scoring option is not yet available, but Harn et al. (2008) provided initial support for deriving multiple scores from the NWF measure. Research is needed to explore the degree to which the WRC score improves upon the predictive strength of the CLS score and the degree to which the WRC score holds comparable predictive strength to scores from PDE. Finally, only one study to date (Harn et al., 2008) has mapped the measurement of the alphabetic principle onto a theoretical framework of word reading development. Additional research is needed to explore how different measures of the alphabetic principle map onto Ehri's (1999) theory of sight word development. Each of these issues was addressed by this study.

Review of Research on Measuring Oral Reading Fluency Growth

Recent research has employed growth on ORF as both a predictor of later reading performance (Baker et al., 2008; Speece & Ritchey, 2005) and as a criterion for evaluating the predictive strength of measures of early reading skills (Briggs, Good, & Rogers, 2007; Chard et al., 2008; Speece & Ritchey, 2005; Stage et al., 2001). This section will review the ways in which each of these studies modeled ORF growth in statistical analyses.

Stage, Sheppard, Davidson, and Browning (2001) employed Hierarchical Linear Modeling (HLM) to model individual growth curves based on four measurement

occasions for ORF from October to May of first grade. Given that all measurement occasions occurred within the same grade-level, Stage et al. did not have to address the issue of nonlinearity between grade-levels. Stage et al. did not describe the linearity of the growth patterns within the grade-level, and it is assumed that the researchers applied a linear growth model to the data set.

Speece and Ritchey (2005) also used HLM growth curve analysis to model ORF growth from January to May of first grade and from November to May of second grade. ORF was administered weekly or monthly across 20 weeks of school for the first grade cohort and weekly or monthly across 36 weeks of school for the second grade cohort. Similar to Stage et al. (2001), Speece and Ritchey (2005) only explored growth within grade levels. Speece and Ritchey (2005) examined the linearity of the data for both cohorts of students. A quadratic model proved to provide the best fit of the data for the first grade cohort, while a linear model was sufficient for fitting data from the second grade cohort.

Chard et al. (2008) collected ORF data across first through third grades and created composite scores of ORF in the spring of each grade level. Unlike the previous studies reviewed, Chard et al. explored ORF growth across grade levels but not ORF growth within grade levels. Composite scores were derived from scores on DIBELS ORF passages at each grade level and a second measure of ORF constructed for growth modeling, where passage difficulty was held constant at the first grade level. Growth across the three composite ORF scores indicated a slight deceleration from the end of second grade to the end of third grade, as compared to growth from the end of first grade

to the end of second grade. Chard et al. fit a standard linear growth model to the three data points and conducted model fit tests. The tests indicated poor model fit, and Chard et al. concluded that the poor fit was due to curvilinear individual growth trajectories. By freely estimating the factor loading for the third grade ORF assessment, Chard et al. were able to improve the model's fit to the data, but results were identical to the linear growth model.

Baker et al. (2008) examined ORF growth from the middle of first grade to the end of second grade for one cohort of students, and ORF growth from the beginning of second grade to the end of third grade for a second cohort of students. Having multiple data points within and across grade levels allowed Baker et al. to fit an accelerated longitudinal growth model to the data. Anticipating nonlinear growth between and within grade levels, Baker et al. performed two adjustments. First, they added two observation-level effects to adjust for a change in level of performance at the beginning of second and third grade. Second, Baker et al. adjusted within year growth by adding .20 to the middle of second grade observation to specify a 20% increase in growth in the first half of second grade and by subtracting .4 from the end of third grade to specify a 40% deceleration from the middle to the end of third grade. Baker et al. performed a series of model fit analyses and determined that "the best-fitting growth model included parameters for time and level of adjustments for Grades 2 and 3. These effects were allowed to vary for individual students and to correlate with each other" (p. 28).

Briggs et al. (2007) also modeled ORF growth within and across grades. Briggs et al. reported nonlinear ORF growth from the middle of first grade to the end of third grade

for a sample of 48,043 students whose ORF data had been entered into the DIBELS Data System. Briggs et al. presented three different approaches to modeling non-linear growth curves using HLM. These approaches include fitting a polynomial regression model, fitting a piecewise regression model, and linearizing the time scale in order to achieve a linear model. Given the complex nature of the nonlinear growth within and across multiple grades, Briggs et al. chose to linearize the time scale. The intercept was set at the end of third grade, and seven measurement occasions from the middle of first grade to the end of third grade were initially coded from -7 to 0 Then the position of each time point was adjusted to create a linearized time scale. The final coding of time across each measurement occasion was -7.54, -5.50, -6.02, -2.96, -1.58, -3.11, -1.33, and 0.04 from the middle of first grade to the end of third grade.

Summary of Research on Measuring Oral Reading Fluency Growth

Research has taken several approaches to modeling ORF growth within and across grade-levels. Currently, only two studies have modeled both within and across grade-level growth for the same sample of students (Baker et al., 2008; Briggs et al., 2007). Baker et al. and Briggs et al. suggested that model adjustments may be needed to account for acceleration and deceleration of growth within and across grades due to the developmental nature of fluency growth over time, summer vacation, and changes in passage difficulty from grade to grade. Currently, no study has investigated approaches to standardizing the ORF scores across grade levels to account for changes in passage difficulty. This study examined this approach and compared it to the use of raw score change as a broad approximation of growth in order to contribute new information to the

methodology of measuring growth in oral reading fluency. Exploring the methodology of measuring a given construct (i.e. growth) is an important step toward building evidence for the measure's use as a meaningful predictor of, or criterion for, later reading development.

CHAPTER III

METHOD

The primary purpose of this study was to explore the strength of predictive relations of different measures of the alphabetic principle for determining word reading fluency, oral reading fluency, and reading comprehension outcomes. This study used data from an ongoing research project conducted by the DIBELS Student Research Team under the advisement of Dr. Roland Good. The author of this study co-led the DIBELS Team project and coordinated each stage of data collection. The DIBELS Team used Pearson product moment correlations to investigate the concurrent and predictive-related evidence for different measures of the alphabetic principle. This study extended beyond the work of the DIBELS Team in two ways. First, a measure of reading comprehension was administered and included as an additional criterion for investigating the predictive-related evidence of different measures of the alphabetic principle. Second, regression analysis and multilevel modeling were employed to provide a more comprehensive investigation of predictor-criterion relationships by examining the predictive strength of single and combined measures of the alphabetic principle across multiple criterion measures.

Participants

Four schools within the Greater Albany school district in Albany, Oregon were recruited for participation. The district encompasses both urban and rural areas, and its

residents are predominantly Caucasian (91.6%) (Hispanic/Latino: 5.5%, Asian/Pacific Islander: 1.4%, Other: 1.5%). Schools were selected by the district's Curriculum Coordinator based on the number of students enrolled in kindergarten and first grade, the use of DIBELS to make instructional decisions, and the interest for conducting research at each school. Two of the schools received Title I funds. The percent of students in each school receiving free and reduced-price lunch in 2007-2008 ranged from 35% to 78%. All students in kindergarten and first grade in each of the four schools were eligible to participate. One of the schools elected to limit participation to first graders only due to scheduling conflicts with the dates of data collection and limited space available for testing students. The kindergarten sample included 109 students, 14 of whom were English Learners. The first grade sample included 212 students, 29 of whom were English Learners. An English Learner was defined as a student receiving English language services by his or her school in the spring of 2007.

Measures

Predictor measures included three measures of the alphabetic principle: Letter Sound Fluency (LSF), Nonsense Word Fluency (NWF) and Phonemic Decoding Efficiency (PDE). Criterion measures included measures of word reading fluency (Sight Word Efficiency (SWE) and Word Identification Fluency(WIF), oral reading fluency (DIBELS ORF), and reading comprehension (Group Reading Assessment and Diagnostic Evaluation (GRA+DE)).

Measures of the Alphabetic Principle

Measures used in this study were designed to measure the alphabetic principle in different ways. Measures were selected that would provide a broad sample of behaviors indicative of development of the alphabetic principle. Each measure selected for use in the study is described. Reliability and validity evidences are summarized in Tables 2-4 in Chapter II.

Letter Sound Fluency (LSF, AIMSweb Test of Early Literacy, Harcourt Educational Measurement, 2002)

LSF is a curriculum based measure that requires students to produce the sounds for letters-in-isolation. One hundred lower case letters are presented in random order in a ten-by-ten layout, and students must produce the most common sound for each letter. The score is the number of correct letter-sounds produced in one minute. Although the AIMSweb version of the LSF measure is the most widely used and easily accessible version, reliability and validity estimates for the AIMSweb version of LSF are not available. The format and administration of this version is similar to the format and administration of versions appearing in the research literature, and studies of these measures reported moderate to strong reliability coefficients (.82 to .93) and weak to strong validity coefficients (.20-.77) (Elliot et al., 2001, Speece & Case, 2001; Speece & Ritchey, 2005; Stage et al., 2001). See Table 2 from Chapter II.

Nonsense Word Fluency (NWF, DIBELS, Good & Kaminski, 2002)

NWF is a one-minute measure of nonsense word reading. Students are presented with a list of randomly ordered vowel-consonant and consonant-vowel-consonant

nonsense words (e.g. uk, puj). The words are all decodable, and the students may read the words sound-by-sound, with partial blends, or as whole words. Two scores are derived from this test: (1) total number of correct letter-sounds produced in one minute (CLS) and (2) total number of words recoded completely and correctly (WRC) in one minute. Students must produce the most common sound for each letter to receive credit. For the WRC score, the student must read the nonsense word as a whole word without elongating sounds or pausing between sounds. If a student sounds-out the word first and then recodes the word, the student still receives credit for recoding. Accurate recoding of nonsense words results in two or three points for the letter-sounds score (depending on whether the word is a two- or three-letter word) and one point for recoding. See Table 3 from Chapter II for reliability and validity evidences for the CLS score.

Phonemic Decoding Efficiency (PDE, Test of Word Reading Efficiency (TOWRE), Torgesen et al., 1999)

The TOWRE is a published, norm-referenced test that includes two 45-second subtests. For the PDE subtest, students read a list of nonsense words. The nonsense words increase in difficulty, and the score is the number of nonsense words read correctly in 45 seconds. Unlike NWF, no credit is given on PDE for the production of individual letter sounds or the partial decoding of a word. See Table 4 from Chapter II for a summary of the reliability and validity evidences reported in the TOWRE Examiner's Manual.

Criterion Measures

The word reading fluency measures and the first measurement occasion for ORF from the spring of 2007 (first grade cohort only) functioned as concurrent criterion

measures. The remaining ORF measurement occasions and the reading comprehension measures functioned as predictive criterion measures.

Sight Word Efficiency (SWE, TOWRE, Torgesen et al., 1999)

The SWE subtest of the TOWRE is a measure of word reading fluency. Students read a list of common regular and irregular words that increase in difficulty, and the score is the number of words read correctly in 45 seconds. Alternate form reliability for SWE was .97, and test-retest reliability was .96. SWE's concurrent correlation with the Word Identification subtest of the Woodcock Reading Mastery Test-Revised (WRMT) was .92 (Torgesen et al., 1999).

Word Identification Fluency (WIF, Curriculum Based Measurement, Deno, Mirkin, & Chiang, 1982)

WIF involves reading words from a list of 50 high-frequency words (two- to six-letters in length) presented in random order, and the score is the number of words read correctly in one minute. The words are randomly selected from 100 high frequency words on the Dolch preprimer, primer, and first-grade level lists. Alternate form/test-retest reliability across two weeks was .97 in one study (Fuchs, Compton, Fuchs, & Bryant, 2004) and .88 in another study (Fuchs, Fuchs, & Compton, 2004). Concurrent correlation with the WRMT-R Word Identification subtest was .77 in the fall of first grade and .82 in the spring of first grade. WIF correlated less strongly with the WRMT Word Attack subtest (.59 in the fall of first grade and .52 in the spring of first grade) (Fuchs, Fuchs, & Compton, 2004).

Oral Reading Fluency (DORF, DIBELS, Good, Kaminski, & Dill, 2002)

DIBELS ORF (DORF) is a General Outcome Measure of students' ability to accurately and fluently read connected text. Students read a passage aloud for one minute, and the score is the number of words read correctly. Omitted or substituted words and words where the student hesitates longer than three seconds are scored as errors. If a student self-corrects a word within three seconds, the word is scored as correct. The student is given three passages to read, and the final score recorded is the median correct words per minute from the three passages. Alternate-form reliability for administration of a single passage ranges from .89 to .96. Concurrent correlations with the Test of Reading Fluency (Children's Educational Services, 1987) range from .91 to .96 across alternate forms of second grade DORF passages (Reading First Assessment Committee, 2002). Buck and Torgesen (2003) found a correlation of .70 between DORF and the Florida Comprehensive Assessment Test in reading at the end of third grade. A similar correlation was found with the Oregon State Assessment in reading at the end of third grade ($r = .67, p < .001$) (Good, Simmons, and Kame'enui, 2001). Predictive validity of DORF from the spring of first grade to the spring of second grade is .82 (Good, Simmons, and Kame'enui, 2001).

GRA+DE (American Guidance Service, 2001)

The GRA+DE is a standardized, group-administered, multiple choice test of overall reading performance. Two composite scores are derived from four subtest scores to provide estimates of students' vocabulary development and comprehension skills. A measure of word recognition is combined with a measure of word meaning to derive a

vocabulary composite score, and a measure of sentence comprehension is combined with a measure of passage comprehension to derive a *comprehension composi*te score. Internal consistency for Levels 2 and 3 of the GRA+DE ranges from .96-.98. Alternate-form reliability ranges from .90-.94, and test-retest reliability ranges from .89-.93. Concurrent correlation is .87 with the California Achievement Test and .90 with the Gates-MacGinitie Reading Test (Gates, Riverside Publishing, 2000) for the Level 2 test and .86 with the Gates for the Level 3 test. Predictive correlation from fall to spring of second grade is .76 with the reading subtest of the Terra Nova.

Procedures

Procedures are described for the training of data collectors, for each of three stages of data collection, and for data preparation.

Training

Data collectors were primarily graduate students in the school psychology or special education doctoral programs at the University of Oregon. Each data collector completed a two-hour training session on the administration and scoring of LSF, NWF, PDE, SWE and WIF. Prior to the training session, data collectors read through a packet of assessments to familiarize themselves with the student and administrator materials. The trainer, an advanced doctoral student in the school psychology program, went through the administration of each measure with the data collectors and provided a model of appropriate administration and opportunities to practice. After each practice session, the data collectors were tested on their scoring accuracy. The trainer completed

the measure as a typical student and produced a scripted set of errors. The data collectors timed and scored the trainer's performance on the measure. The data collectors calculated the number of correct words (or sounds), and then the trainer reviewed each error and the correct score for the measure. Discrepancies in scoring were discussed and scoring rules were reviewed.

To determine the level of agreement between the data collectors and the trainer, the data collectors' scores for each measure were compared to the trainer's predetermined score (based on the scripted set of errors). The percent of agreement between scores for each data collector with the trainer across each measure was averaged to provide an overall estimate of agreement. The average percent agreement for 14 data collectors across the measures was 99% (range from 98% to 100%). Level of agreement for the WRC score on NWF was not evaluated because data collectors were not required to calculate this score. During the training session, data collectors were taught to indicate if a student recoded a word on NWF by drawing a continuous line under the word, but they were not required to tally the number of words recoded during training or actual administration.

Data Collection

Figure 4 presents a timeline of the data collection stages and the measures that were administered at each stage. The first stage of data collection occurred in May of 2007. The second stage of data collection included four measurement occasions (fall, winter, and spring of 2007-2008 and fall of 2008). The third stage of data collection occurred in late fall of 2008.

Timeline for Data Collection

	Stage I	Stage II				Stage III
	May 2007	Oct 2007	Jan 2008	May 2008	Oct 2008	Nov. 2008

Kindergarten Cohort	Kindergarten	First Grade			Second Grade	
	IVs: LSF, NWF, PDE DVs: SWE, WIF	—	DV: DORF	DV: DORF	DV: DORF	DV: GRA+DE

First Grade Cohort	First Grade	Second Grade			Third Grade	
	IVs: LSF, NWF, PDE DVs: SWE, WIF, DORF	DV: DORF	DV: DORF	DV: DORF	DV: DORF	DV: GRA+DE

Figure 4. Timeline for data collection in stages I, II, and III. IV = Independent Variable (Predictor Measure). DV = Dependent Variable (Criterion Measure).

Stage I of Data Collection

A team of 5 to 10 data collectors went to each of the four schools during the first week of May, 2007. Data collection took place primarily in the hallways or in empty classrooms. Each student who participated in the study worked one-on-one for 8 to 20 minutes with a trained data collector. Each student completed five or six brief measures

of early reading. The order in which the measures were presented was randomized to control for potential confounding effects such as carry-over effects that might occur when a student switches from reading real words to reading nonsense words.

Stage II of Data Collection

After each school completed the DIBELS benchmark assessments in May of 2007 and entered data into the DIBELS Data System, the school district's DIBELS Data System Administrator obtained students' DORF scores for the spring benchmark and sent the data to the DIBELS Team. The same data collection procedure was conducted after each benchmark period during the 2007-2008 school year and into the fall of 2008.

Stage III of Data Collection

In the fall of 2008, the GRA+DE was given to second and third grade students in the four schools that participated in stage I and stage II of data collection. Teachers in these grades were invited to participate in the study, and letters of consent were mailed home to the parents of students in each class. Research assistants were recruited to assist with administering the GRA+DE at two of the four schools. The other two schools elected to have teachers and literacy coaches administer the tests to the students. Test administrators received training commensurate with recommendations provided in the administration manual for the test. All students in each class were invited to complete the test, but only the scores of students who also completed stage I of data collection were retained for this study. Students typically completed the test in two 45-minute sessions.

Data Preparation

Interrater reliability. Reliability checks occurred for 3 percent (11 of 321) of the administrations during stage I of data collection. A doctoral student with advanced training in the scoring and administration of the measures sat in close proximity to the data collector and shadow-scored as the data collector completed each assessment with the student. Pearson product moment correlations ranged from .95 to 1.0 across all measures. Percent agreement for the total score on each measure ranged from 82% to 97%.

Accuracy of scoring and data entry. After completion of stage I of data collection, one of five doctoral students in the school psychology program rescored each protocol. Scorers also reviewed the data collectors' markings on the NWF protocols to determine the WRC score. When uncertainties in scoring arose, the scorers worked in pairs, or collectively, to determine the correct score. Each scorer entered data into Excel for a set of protocols. The author checked the accuracy of data entry for each protocol. After completion of stage III of data collection, answers for the GRA+DE were hand-entered into the GRA+DE scoring software (Pearson, 2007). Item-entry was verified for 20% of the protocols (n = 34). Three of the 34 protocols that were verified had a single item-entry error that resulted in a raw score change of one point. Final scores were exported from the scoring software and matched to existing data in the Excel data sheet using the DIBELS Data System identification numbers.

Data retrieval. After all data were collected and entered onto the Excel data sheet, cases with missing DORF scores were identified by their DIBELS identification

numbers. These identification numbers were sent to the district's data system administrator to determine if missing DORF scores could be obtained from the larger district data base. For the kindergarten cohort, all three DORF scores were obtained for 86 of the 109 original students in the study. For the first grade cohort, all five DORF scores were obtained for 150 of the 212 original students in the study. After missing DORF scores were retrieved and entered into the Excel data sheet, the DIBELS identification numbers were replaced with randomly generated identification numbers, and the district's DIBELS Data System administrator was given a copy of the matching sheet.

Missing data and sample sizes. Given that data collection occurred across three stages during a 20-month period, complete sets of data were collected for only a portion of the students in the original sample (58 of 109 students in the kindergarten cohort and 121 of 212 students in the first grade cohort). To address missing data concerns in the analyses, three samples were identified that correspond to each stage of data collection. See Table 5 for sample sizes.

Derived scores. Three scores were derived from the raw data for use in the analyses. Given that WIF and SWE were highly correlated (.91 for the kindergarten sample and .94 for the first grade sample), these scores were averaged to derive a *word reading fluency composite*. Next, *raw score change* was calculated for DORF at two points in time to give a broad index of fluency development. For the kindergarten cohort, the middle of first grade DORF score was subtracted from the end of first grade DORF

score. For the first grade cohort, the end of first grade DORF score was subtracted from the end of second grade DORF score.

Table 5

Sample Sizes at Each Stage of Data Collection

Subsample	Kindergarten Cohort	First Grade Cohort
Stage I[a]	106	206
Stage II[b]	83	146
Stage III[c]	62	122
Complete Data Across All Stages	58	121

[a]Students with complete sets of scores for all predictor measures and concurrent criterion measures. [b]Students with complete sets of scores for predictor measures, concurrent criterion measures, and each measure of Oral Reading Fluency [c]Students with complete sets of scores for predictor measures, concurrent criterion measures, and the Group Reading Assessment and Diagnostic Evaluation

Analysis

Four research questions were investigated for this study. The first three questions address the prediction of scores on outcome measures at specific points in time. The fourth research question addresses the prediction of growth on repeated measures of oral reading fluency over time.

Research Question One

How much variance in word reading fluency, oral reading fluency, and reading comprehension is explained by each measure of the alphabetic principle?

A linear regression model was used to examine the relationship between each predictor and each criterion, and an F test statistic was used to determine if the relationship was significant at $p < .05$.

Research Question Two

Does including a measure of letter-sounds in nonsense words (i.e. CLS) add significantly to the variance explained in word reading fluency, oral reading fluency, and reading comprehension beyond a measure of letter-sounds in isolation (i.e. LSF)?

A multiple linear regression model was used to examine the contributions of LSF and CLS as combined predictors of each of the criterion. The R^2 value, F statistic, and mean-square residual were reported for LSF and CLS as sole predictors and for LSF and CLS as combined predictors for each criterion.

Research Question Three

Does including a measure of nonsense words recoded (i.e. WRC, PDE) add significantly to the variance explained in word reading fluency, oral reading fluency, and reading comprehension beyond a measure of letter-sounds (i.e. LSF, CLS)?

A multiple linear regression model was used to examine the contributions of four different sets of measures as combined predictors of each criterion: (a) CLS and WRC, (b) CLS and PDE, (c) LSF and WRC, and (d) LSF and PDE. The R^2 value, F statistic, and mean-square residual were reported for each set of predictors.

Research Question Four

When predicting change/growth on repeated measures of oral reading fluency over time:

(a) How much variance is explained by each measure of the alphabetic principle?

(b) Does combining a measure of letter-sounds in nonsense words with a measure of letter-sounds in isolation add significantly to the variance explained?

(c) Does combining a measure of nonsense words recoded with a measure of letter-sounds add significantly to the variance explained?

The fourth research question was addressed in two different ways for each cohort. First, regression models for research questions one through three were re-run with raw score change on DORF as the dependent variable. The R^2 value, F statistic, and mean-square residual were reported for each predictor in isolation and for each set of predictors. Second, DORF scores were transformed into Lexiles, and hierarchical linear modeling (HLM; Bryk & Raudenbush, 1992) was used to evaluate the predictor-criterion relationships between measures of the alphabetic principle and slope on Lexile measures of DORF over time.

CHAPTER IV

RESULTS

This chapter describes the results from this study and is organized into three sections: *descriptive statistics*, *prediction of outcomes*, and *prediction of growth.* First, sample sizes, means, standard deviations, and ranges of performance are reported, and the distribution of scores for each measure is reviewed. In the next section, Pearson product moment correlations and scatterplot matrices are reviewed to explore the strength and nature of predictor-criterion relationships before addressing research questions one through three. The final research question is addressed in the last section, *prediction of growth.*

Descriptive Statistics

Predictor Measures

Table 6 presents the average performance for the kindergarten and first grade cohorts on measures of the alphabetic principle, administered in stage I of data collection (i.e. spring of kindergarten/first grade). Frequency histograms for the predictor variables are presented in the Appendix. Scores for Letter Sound Fluency (LSF) and Correct Letter Sounds (CLS) appear normally distributed at both grade levels, with positive skew noted for CLS. All scores on LSF were within 2.5 standard deviations of the mean while CLS had outliers in excess of 3 standard deviations beyond the mean. Floor effects are apparent for Words Recoded Completely and Correctly (WRC) and Phonemic Decoding

Efficiency (PDE) in spring of kindergarten. Fifty percent (n=53) of students scored 0 on WRC and 40% (n=44) scored 0 on PDE. Scores for PDE were more normally distributed in the spring of first grade while floor effects remained for WRC.

Table 6

Average Performance on Measures of the Alphabetic Principle

	Kindergarten Cohort (n = 106)			First Grade Cohort (n = 206)		
Measure	*M*	*SD*	*Range*	*M*	*SD*	*Range*
LSF	26.15	13.53	0-59	44.65	16.00	0-84
CLS	27.27	19.84	0-111	62.41	32.59	0-188
WRC	3.56	5.67	0-28	13.36	12.94	0-61
PDE	4.21	5.06	0-23	15.37	9.67	0-53

Two scatterplots are presented in Figure 5 to further examine WRC's floor effect at the end of first grade. The first scatterplot presents a comparison of the WRC scores with their corresponding CLS scores on the NWF measure, and the second scatterplot presents a comparison of the WRC and PDE scores. As can been seen in the first scatterplot, there were a number of students who scored 0 on WRC but produced numerous letter-sounds on CLS. Examination of the second scatterplot reveals that many students who scored 0 on WRC did *not* score 0 on PDE which also measures nonsense word recoding. Of 206 students who completed both the NWF and PDE measures in the spring of first grade, 24 students scored 0 on WRC while only 3 students scored 0 on PDE, and 2 students scored 0 on both measures. Scores on PDE for the 24 students who

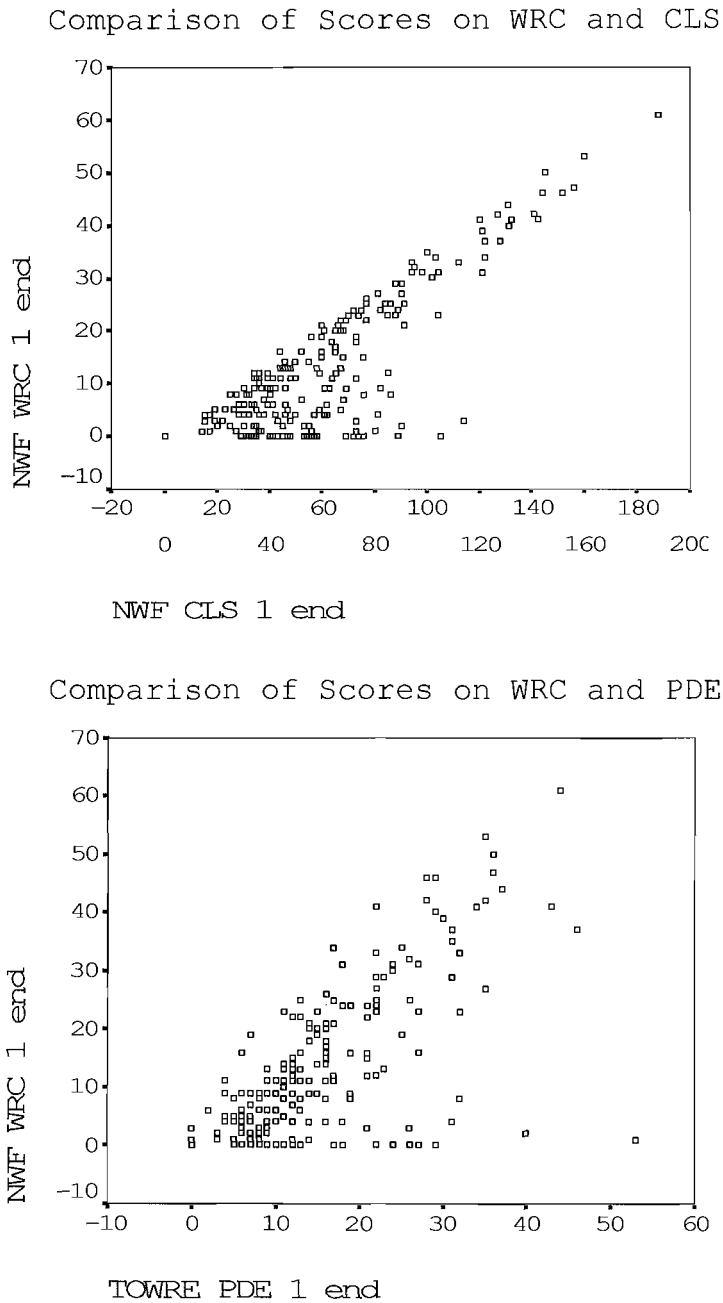


Figure 5. Scatterplots comparing scores within and across measures for the first grade cohort.

scored 0 on WRC ranged from 0 to 29 with a mean of 12.58 and standard deviation of 7.89. Investigation of outliers revealed that one student scored 73 on CLS, 53 on PDE, and 1 on WRC. Another student scored 90 on CLS, 40 on PDE, and 2 on WRC. These results indicate that WRC may not accurately reflect students' abilities to recode nonsense words in this sample since many students who did not recode nonsense words on WRC were able to recode similar nonsense words on PDE.

Criterion Measures

Table 7 presents the average performance for the kindergarten and first grade cohorts on measures of word reading fluency and oral reading fluency administered in stages I and II of data collection. Table 8 presents the average performances on the reading comprehension measure administered during stage III of data collection.

Frequency histograms for the criterion measures revealed floor effects for the word reading fluency composite in spring of kindergarten. DIBELS oral reading fluency (DORF) scores appeared positively skewed in winter and spring of first grade and fall of second grade. In winter of second grade, DORF scores appear normally distributed. Vocabulary Composite scores on the Group Reading Assessment and Diagnostic Evaluation (GRA+DE) were impacted by a ceiling effect in second grade but are more normally distributed in third grade. The Comprehension Composite scores appeared normally distributed across both second and third grades.

Table 7

Average Performance on Measures of Word Reading Fluency and Oral Reading Fluency

	Kindergarten Cohort			First Grade Cohort		
Measure	*M*	*SD*	*Range*	*M*	*SD*	*Range*
SWE	9.08	11.42	0-57	31.74	16.33	0-78
WIF	6.65	12.81	0-73	37.45	27.73	0-117
Word Reading Fluency Composite[a]	7.87	11.84	0-65	34.60	21.74	0-96.5
ORF Winter 1st Grade	25.66	32.27	0-144	—	—	—
ORF Spring 1st Grade	43.22	35.35	0-143	43.18	34.99	0-215
ORF Fall 2nd Grade	38.89	32.47	0-147	38.12	27.92	0-165
ORF Winter 2nd Grade	—	—	—	71.47	36.00	1-181
ORF Spring 2nd Grade	—	—	—	85.88	35.92	2-209
ORF Fall 3rd Grade	—	—	—	69.84	30.78	0-163
ORF Raw Score Change[b]	17.55	13.16	-12-57	43.14	18.73	1-90

Note. Refer to Table 5 for sample sizes at stages I and II of data collection. SWE = Sight Word Efficiency (subtest of the Test of Word Reading Efficiency.) WIF = Word Identification Fluency.

[a]Word Reading Fluency Composite = average of WIF and SWE. [b]ORF Raw Score Change, kindergarten = (ORF Spring of 1st –ORF Winter of 1st); ORF Raw Score Change, first grade = (ORF Spring of 2nd –ORF Spring of 1st).

Table 8

Average Performance on the Group Reading Assessment and Diagnostic Evaluation (GRA+DE)

	Second Grade (n = 62)			Third Grade (n = 122)		
Score	*M*	*SD*	*Range*	*M*	*SD*	*Range*
Vocabulary Composite RS	44.52	11.25	15-55	46.21	9.19	20-60
(SS)	(97.06)	(18.01)	(57-126)	(97.71)	(13.23)	(66-137)
Word Reading RS	21.95	6.08	4-28	27.06	3.25	13-30
Word Meaning RS	22.56	5.69	6-27	19.16	6.61	4-30
Comprehension Composite RS	22.95	10.97	3-44	28.61	9.51	8-44
(SS)	(92.83)	(15.05)	(55-125)	(96.17)	(12.96)	(67-124)
Sentence Comp. RS	10.48	5.28	1-19	14.60	3.89	3-19
Passage Comp. RS	12.47	6.25	2-27	14.02	6.54	2-26
Total Test RS	67.47	20.80	19-99	74.83	17.97	30-102
(SS)	(93.82)	(16.65)	(55-131)	(96.98)	(13.33)	(66-127)

Note. Indented scores indicate subtest scores that are summed to derive the composite raw score. Standard scores are in parentheses. RS = Raw Score. SS = Standard Score.

Prediction of Outcomes

The strength and nature of relationships among measures was investigated by calculating Pearson product moment correlations and by visually inspecting scatterplots. Associations among all the predictor and criterion measures were significant at $p < .05$.

Bivariate Relationships

Correlations among measures of the alphabetic principle and word reading fluency are presented in Table 9. For the kindergarten cohort, these correlations ranged from .50 to .82 with weaker correlations noted for LSF (.50 to .67) than for the other measures (.71 to .85). For the first grade cohort, correlations ranged from .29 to .85 with weaker correlations noted again for LSF (.29 to .48) than for the other measures (.60 to .85).

Table 10 presents correlations of alphabetic principle and word reading fluency measures with selected measurement occasions for DORF. Correlations of predictor measures with DORF ranged from .29 to .81. Weak to moderate correlations between LSF and DORF (.29-.39) and between WRC and DORF (.53) were found for the first grade cohort. The word reading fluency composite, a concurrent criterion measure, correlated more strongly with DORF (.79-.91) when compared to the correlations between the predictor measures and DORF. Correlations were also stronger with the first DORF measurement occasion than with the second measurement occasion, except in the case of LSF.

Correlations among measures of the alphabetic principle, word reading fluency, and selected measures of oral reading fluency with scores on the GRA+DE are

Table 9

Correlations Among Concurrent Measures of Alphabetic Principle and Word Reading Fluency for Kindergarten and First Grade

Measure	LSF	CLS	WRC	PDE	Word Reading Composite
		Kindergarten Cohort (n = 106)			
LSF	—				
CLS	.48	—			
WRC	.29	.82	—		
PDE	.38	.81	.67	—	
Word Reading Composite	.40	.83	.60	.85	—
		First Grade Cohort (n = 206)			
LSF	—				
CLS	.67	—			
WRC	.50	.73	—		
PDE	.59	.71	.82	—	
Word Reading Composite	.54	.85	.80	.78	—

Note. All correlations are significant at $p < .001$.

presented in Table 11. For the kindergarten cohort, correlations for WRC with the GRA+DE ranged from .36-.42 while correlations for the other predictors ranged from .50-.61. For the first grade cohort, correlations for LSF and WRC ranged from .25-.33, while correlations for PDE and CLS ranged from .50-.57. Correlations between scores on DORF and GRA+DE (.59-.80) were stronger than correlations between the predictors and GRA+DE (.25-.61).

Table 10

Correlations of Alphabetic Principle and Word Reading Fluency Measures with Oral Reading Fluency

	Kindergarten Cohort (n = 83)		First Grade Cohort (n = 146)	
Measure	ORF Winter First Grade	ORF Spring First Grade	ORF Spring First Grade[a]	ORF Spring Second Grade
LSF	.62**	.64**	.29**	.39**
CLS	.79**	.72**	.78**	.75**
WRC	.77**	.67**	.53**	.53**
PDE	.80**	.74**	.81**	.75**
Word Reading Composite	.89**	.79**	.91**	.85**

[a]Concurrent criterion measure.

*$p < .05$, **$p < .001$.

Scatterplot matrices were examined to determine the nature of bivariate relationships between predictor and criterion measures. Relationships appear more linear

between the first grade predictors and the criterion measures than between the kindergarten predictors and the criterion measures. In general, the trend of data points in each scatterplot seemed to be approximated well by a straight line.

Table 11

Correlations of Alphabetic Principle, Word Reading Fluency, and Oral Reading Fluency Measures with the GRA+DE

	Kindergarten Cohort in Fall of Second Grade (n = 62)		First Grade Cohort in Fall of Third Grade (n = 122)	
Measure	Vocab.	Comp.	Vocab.	Comp.
LSF	.61	.53	.32	.29
CLS	.61	.61	.55	.53
WRC	.42	.36	.33	.30
PDE	.54	.52	.57	.50
Word Reading Fluency Composite	.48	.54	.68	.70
ORF[e] Winter 1st Grade	.59	.73	—	—
ORF Spring 1st Grade	.70	.80	.65	.70
ORF Spring 2nd Grade	—	—	.77	.75

All correlations are significant, $p < .01$.

Research questions one, two, and three address the prediction of outcomes for measures of word reading fluency, oral reading fluency, and reading comprehension using linear regression models. Model assumptions were reviewed prior to addressing the research questions. Examination of histograms revealed that the residuals for each

regression model were approximately normally distributed around their mean value zero. Scatterplots of standardized residuals against standardized predicted values did not reveal any major violations to the assumption of homoscedasticity. In addition to regression diagnostics, a global test of independence was conducted for each dependent variable to determine the total variance explained by the combination of all four predictors. In Table 12, the first row under each criterion measure presents the total variance explained when all predictors are entered simultaneously into the model.

Research Question One

How much variance in word reading fluency, oral reading fluency, and reading comprehension is explained by each measure of the alphabetic principle?

A linear regression model was used to examine the relationship between each predictor and each criterion. Table 12 presents the variance explained by each predictor on each dependent variable.

Summary of results for research question one. Each predictor explained significant portions of variance in each criterion when entered independently into the regression model ($p < .05$). LSF and WRC explained more variance as independent predictors in the spring of kindergarten than in the spring of first grade, while PDE and CLS performed similarly across both cohorts. For the kindergarten cohort, LSF was a weaker predictor of the word reading fluency composite than CLS, WRC, and PDE, explaining only 29% of the variance compared to the variance explained by the other predictors, 72%, 64% and 61% respectively. WRC in the spring of kindergarten stood out as a weaker predictor of the GRA+DE explaining13% to 17% of the variance compared

Table 12

Variance in Each Criterion Explained by All and Each Measure of the Alphabetic Principle

Model	Cohort	R^2	*F* Statistic	*df*
	Word Reading Fluency (n = 106 for K, 206 for 1st)			
LSF, CLS, WRC, PDE	K	.81	106.18**	101
	1	.80	203.05**	201
LSF	K	.29	42.66**	104
	1	.16	38.71**	204
CLS	K	.72	266.51**	104
	1	.69	455.61**	204
WRC	K	.64	185.46**	104
	1	.37	117.34**	204
PDE	K	.61	163.36**	104
	1	.72	511.03**	204
	ORF Spring of 1st Grade (n = 83 for K, 206 for 1st)			
LSF, CLS, WRC, PDE	K	.62	31.82**	78
	1	.77	168.13**	201
LSF	K	.41	56.55**	81
	1	.09	18.91**	204
CLS	K	.52	88.25**	81
	1	.56	261.15**	204
WRC	K	.44	64.76**	81
	1	.29	82.81**	204
PDE	K	.55	97.20**	81
	1	.73	536.87**	204
	ORF Spring of 2nd Grade (n = 146 for 1st)			
LSF, CLS, WRC, PDE	1	.64	61.98**	141
LSF	1	.15	25.93**	144
CLS	1	.57	187.46**	144

Table 12 (continued)

Model	Cohort	R^2	*F* Statistic	*df*
	ORF Spring of 2nd Grade (n = 146 for 1st)			
WRC	1	.29	57.30**	144
PDE	1	.56	183.44**	144
	GRA+DE Vocabulary (n = 62 for K, 122 for 1st)			
LSF, CLS, WRC, PDE	K	.45	11.83**	57
	1	.38	17.94**	117
LSF	K	.38	36.36**	60
	1	.10	13.29**	120
CLS	K	.37	34.84**	60
	1	.31	53.13**	120
WRC	K	.17	12.46*	60
	1	.11	14.97**	120
PDE	K	.29	24.22**	60
	1	.32	56.28**	120
	GRA+DE Comprehension (n = 62 for K, 122 for 1st)			
LSF, CLS, WRC, PDE	K	.42	10.19**	57
	1	.33	14.49**	117
LSF	K	.28	23.71**	60
	1	.09	11.28*	120
CLS	K	.37	35.17**	60
	1	.28	45.89**	120
WRC	K	.13	8.67*	60
	1	.09	11.92*	120
PDE	K	.27	22.27**	60
	1	.25	39.85**	120

*$p < .05$, **$p < .001$.

to the other predictors which explained 25% to 38% of the variance. In first grade CLS and PDE explained the most variance in each criterion measure often explaining twice the variance that was explained by WRC or LSF when each was entered as an independent predictor.

Research Question Two

Does including a measure of letter-sounds in nonsense words (i.e. CLS) add significantly to the variance explained in word reading fluency, oral reading fluency, and reading comprehension beyond a measure of letter-sounds in isolation (i.e. LSF)?

A multiple regression model was used to examine the contributions of LSF and CLS as combined predictors of each of the criterion. Results from this analysis are presented in Tables 13-15. Each table addresses a single or set of criterion measures.

CLS explained significant additional unique variance in each criterion measure beyond the variance explained by LSF ($p < .05$). The additional unique variance explained by CLS ranged from 6% on the GRA+DE Vocabulary Composite (kindergarten) to 53% on the word reading fluency composite (first grade). When CLS was entered into the model first, LSF explained additional unique variance in two of the nine prediction models (DORF in the spring of first grade and the GRA+DE Vocabulary Composite in the fall of second grade (kindergarten cohort only)). The additional unique variance explained by LSF in these models ranged from 4% to 7%.

Summary of results for research question two. These results indicate that including a measure of letter sounds in nonsense words (CLS) significantly improved upon the variance explained in reading outcomes beyond a measure of letter sounds in

isolation (LSF). Both predictor measures were necessary for the prediction of two of nine criterion measures while a single measure of letter sounds in nonsense words was sufficient for the prediction of the seven remaining criterion measures.

Table 13

Model Summary for Predicting Word Reading Fluency from Measures of Letter-Sounds in Isolation and Letter-Sounds in Nonsense Words

Model	Cohort	R^2	R^2 Change	F Change	*df*	*MS* Residual
1. LSF	K	.29		42.66**	104	100.40
	1	.16		38.71**	204	399.26
2. LSF, CLS	K	.72	.43	158.64**	103	39.91
	1	.69	.53	348.75**	203	147.62
1. CLS	K	.72		266.51**	104	39.74
	1	.69		455.61**	204	146.91
2. CLS, LSF	K	.72	.00	0.56	103	39.91
	1	.69	.00	0.02	203	147.62

Note. $n = 106$ for kindergarten cohort, 206 for first grade cohort.

*$p < .05$, **$p < .001$.

Table 14

Model Summary for Predicting Oral Reading Fluency from Measures of Letter-Sounds in Isolation and Letter-Sounds in Nonsense Words

Model	Cohort	R^2	R^2 Change	*F* Change	*df*	*MS* Residual
	ORF Spring of First Grade (n = 83 for K, 206 for 1st)					
1. LSF	K	.41		56.55**	81	744.91
	1	.09		18.91**	204	1126.05
2. LSF, CLS	K	.56	.15	27.20**	80	562.86
	1	.57	.48	225.85**	203	535.66
1. CLS	K	.52		88.25**	81	605.41
	1	.56		261.15**	204	539.63
2. CLS, LSF	K	.56	.04	7.12*	80	562.86
	1	.57	.01	2.51	203	535.66
	ORF Spring of Second Grade (n = 146 for 1st)					
1. LSF	1	.15		25.93**	144	1100.82
2. LSF, CLS	1	.57	.42	137.23**	143	565.68
1. CLS	1	.57		187.46**	144	564.34
2. CLS, LSF	1	.57	.00	0.66	143	565.68

*$p < .05$, **$p < .001$.

Table 15

Model Summary for Predicting Measures of Vocabulary and Comprehension from Measures of Letter-Sounds in Isolation and Letter-Sounds in Nonsense Words

Model	Cohort	R^2	R^2 Change	F Change	df	MS Residual
		GRA+DE Vocabulary (n = 62 for K, 122 for 1st)				
1. LSF	K	.38		36.36**	60	80.09
	1	.10		13.29**	120	76.61
2. LSF, CLS	K	.44	.06	6.70*	59	73.14
	1	.31	.21	36.05**	119	59.29
1. CLS	K	.37		34.84**	60	81.37
	1	.31		53.13**	120	58.97
2. CLS, LSF	K	.44	.07	7.75*	59	73.14
	1	.31	.00	0.37	119	59.29
		GRA+DE Comprehension (n = 62 for K, 122 for 1st)				
1. LSF	K	.28		23.71**	60	87.70
	1	.09		11.28*	120	83.36
2. LSF, CLS	K	.39	.11	10.77*	59	75.41
	1	.28	.19	31.65**	119	66.40
1. CLS	K	.37		35.17**	60	77.14
	1	.28		45.89**	120	65.97
2. CLS, LSF	K	.39	.02	2.37	59	75.41
	1	.28	.00	0.21	119	66.40

*$p < .05$, **$p < .001$.

Research Question Three

Does including a measure of nonsense words recoded (i.e. WRC or PDE) add significantly to the variance explained in word reading fluency, oral reading fluency, and reading comprehension beyond a measure of letter-sounds (i.e. LSF or NWF-CLS)?

A multiple regression model was used to examine four combinations of predictors: (a) LSF and WRC, (b) LSF and PDE, (c) CLS and WRC, and (d) CLS and PDE. Results are presented in Tables 16-18. Each table addresses a single or set of criterion measures.

Both measures of nonsense word recoding (WRC and PDE) explained significant additional variance beyond the variance explained by measures of letter sounds (LSF or CLS) for criterion measures of word reading fluency and oral reading fluency for both cohorts. One or both measures of nonsense word recoding explained significant additional variance beyond the variance explained by LSF or CLS in criterion measures of vocabulary and comprehension for both cohorts.

Additional regression models were run to determine if a measure of letter-sounds was still needed to maximally predict reading outcomes when a measure of nonsense word recoding was already entered into the model. For criterion measures of word reading fluency and oral reading fluency, both types of predictor measures (letter-sounds and nonsense words recoded) contributed unique and significant variance to the model with one exception: LSF did not make a unique and significant contribution to the prediction of word reading fluency in the presence of PDE when both predictor measures were administered in the spring of kindergarten ($F(2, 103) = 2.35$, $p = .13$). For criterion

Table 16

Model Summary for Predicting Word Reading Fluency from Measures of Letter-Sounds and Nonsense Words

Model	Cohort	R^2	R^2 Change	*F* Change	*df*	*MS* Residual
1. LSF	K	.29		42.66**	104	100.40
	1	.16		38.71**	204	399.26
2. LSF, WRC	K	.67	.38	115.38**	103	47.81
	1	.42	.26	91.64**	203	276.44
3. LSF, PDE	K	.62	.33	89.04**	103	54.37
	1	.72	.56	410.44**	203	132.78
1. CLS	K	.72		266.51**	104	39.74
	1	.69		455.61**	204	146.91
2. CLS, WRC	K	.79	.07	33.79**	103	30.22
	1	.71	.02	13.86**	203	138.20
3. CLS, PDE	K	.78	.07	30.76**	103	30.90
	1	.78	.09	81.26**	203	105.43

Note. $n = 106$ for kindergarten cohort, 206 for first grade cohort.

*$p < .05$, **$p < .001$.

Table 17

Model Summary for Predicting ORF from Measures of Letter-Sounds and Nonsense Words

Model	Cohort	R^2	R^2 Change	*F* Change	*df*	*MS* Residual
		ORF in Spring of 1st Grade (n = 83 for K, 206 for 1st)				
1. LSF	K	.41		56.55**	81	744.91
	1	.09		18.91**	204	1126.05
2. LSF, WRC	K	.55	.14	25.67**	80	570.99
	1	.31	.22	65.96**	203	854.09
3. LSF, PDE	K	.59	.18	35.98**	80	520.24
	1	.73	.64	474.64**	203	338.99
1. CLS	K	.52		88.25**	81	605.41
	1	.56		261.15**	204	539.63
2. CLS, WRC	K	.55	.03	4.52*	80	580.21
	1	.58	.02	9.48*	203	518.09
3. CLS, PDE	K	.60	.08	16.00**	80	510.81
	1	.74	.18	134.64**	203	326.05
		ORF in Spring of 2nd Grade (n = 146 for 1st)				
1. LSF	1	.15		25.93**	144	1100.82
2. LSF, WRC	1	.36	.20	45.40**	143	841.38
3. LSF, PDE	1	.57	.42	140.81**	143	558.54
1. CLS	1	.57		187.46**	144	564.34
2. CLS, WRC	1	.58	.02	5.34*	143	547.82
3. CLS, PDE	1	.62	.05	20.10**	143	498.26

*$p < .05$, **$p < .001$.

Table 18

Model Summary for Predicting Measures of Vocabulary and Comprehension from Measures of Letter-Sounds and Nonsense Words

Model	Cohort	R^2	R^2 Change	*F* Change	*df*	*MS* Residual
		GRA+DE Vocabulary (n = 62 for K, 122 for 1st)				
1. LSF	K	.38		36.36**	60	80.09
	1	.10		13.29**	120	76.61
2. LSF, WRC	K	.41	.03	3.28	59	77.15
	1	.17	.07	9.56*	119	71.51
3. LSF, PDE	K	.43	.05	5.30*	59	74.73
	1	.33	.23	39.88**	119	57.86
1. CLS	K	.37		34.84**	60	81.37
	1	.31		53.13**	120	58.97
2. CLS, WRC	K	.37	.00	.47	59	82.10
	1	.34	.03	5.42*	119	56.88
3. CLS, PDE	K	.39	.03	2.39	59	79.53
	1	.33	.04	6.90*	119	56.21
		GRA+DE Comprehension (n = 62 for K, 122 for 1st)				
1. LSF	K	.28		23.71**	60	87.70
	1	.09		11.28*	120	83.36
2. LSF, WRC	K	.31	.02	1.98	59	86.29
	1	.14	.05	7.39*	119	79.14
3. LSF, PDE	K	.35	.07	6.36*	59	80.51
	1	.26	.17	27.35**	119	68.35

Table 18 (continued)

Model	Cohort	R^2	R^2 Change	F Change	df	MS Residual
	GRA+DE Comprehension (n = 62 for K, 122 for 1st)					
1. CLS	K	.37		35.17**	60	77.14
	1	.28		45.89**	120	65.97
2. CLS, WRC	K	.37	.00	0.00	59	78.44
	1	.31	.04	6.30*	119	63.18
3. CLS, PDE	K	.39	.02	1.70	59	76.25
	1	.29	.02	2.44	119	65.18

*$p < .05$, **$p < .001$.

measures of reading comprehension, results were mixed. Of 16 models that were run, 7 models indicated that only a measure of letter-sounds was necessary for the prediction of vocabulary and comprehension performance; 7 models indicated that both types of predictor measures were necessary, and 2 models indicated that only a measure of nonsense word recoding (PDE) was necessary for the prediction of vocabulary and comprehension.

Summary of results for research question three. These results indicate that including a measure of nonsense words recoding (i.e. WRC or PDE) adds significantly to the variance explained in word reading fluency, oral reading fluency, and reading comprehension beyond a measure of letter-sounds (i.e. LSF or CLS). The results also indicate that measures of letter-sounds explain unique and significant portions of variance beyond the variance explained by measures of nonsense words recoded when predicting word reading fluency and oral reading fluency. When predicting performance on criterion

measures of reading comprehension administered 20 months after the predictor measures, one type of measure may or may not explain additional variance in the presence of the other type of measure. This combination of results suggests that both types of predictor measures should be used to maximally predict performance on measures of word reading fluency and oral reading fluency, but both types of measures may not be necessary for the prediction of performance on measures of reading comprehension that are temporally and developmentally distal from the predictor measures.

Prediction of Growth

The fourth research question addresses the prediction of growth on measures of oral reading fluency over time. Prior to answering the research question, growth trajectories were inspected to evaluate the linearity of DORF growth over time.

Modeling Growth

Scores from the fall of each year were hypothesized to be lower than scores from the spring of the previous year due to the influence of summer vacation and the change in passage difficulty. Additionally, previous research indicates that DORF growth within second grade may reflect a positive curvilinear growth pattern from fall to winter (Baker et al., 2008). Figure 6 presents boxplots of DORF scores at each measurement occasion. There were three measurement occasions for the kindergarten cohort and five measurement occasions for the first grade cohort.

Visual inspection of boxplots and comparisons of mean performance at each measurement occasion (see Table 7) revealed a modest dip in mean performance from

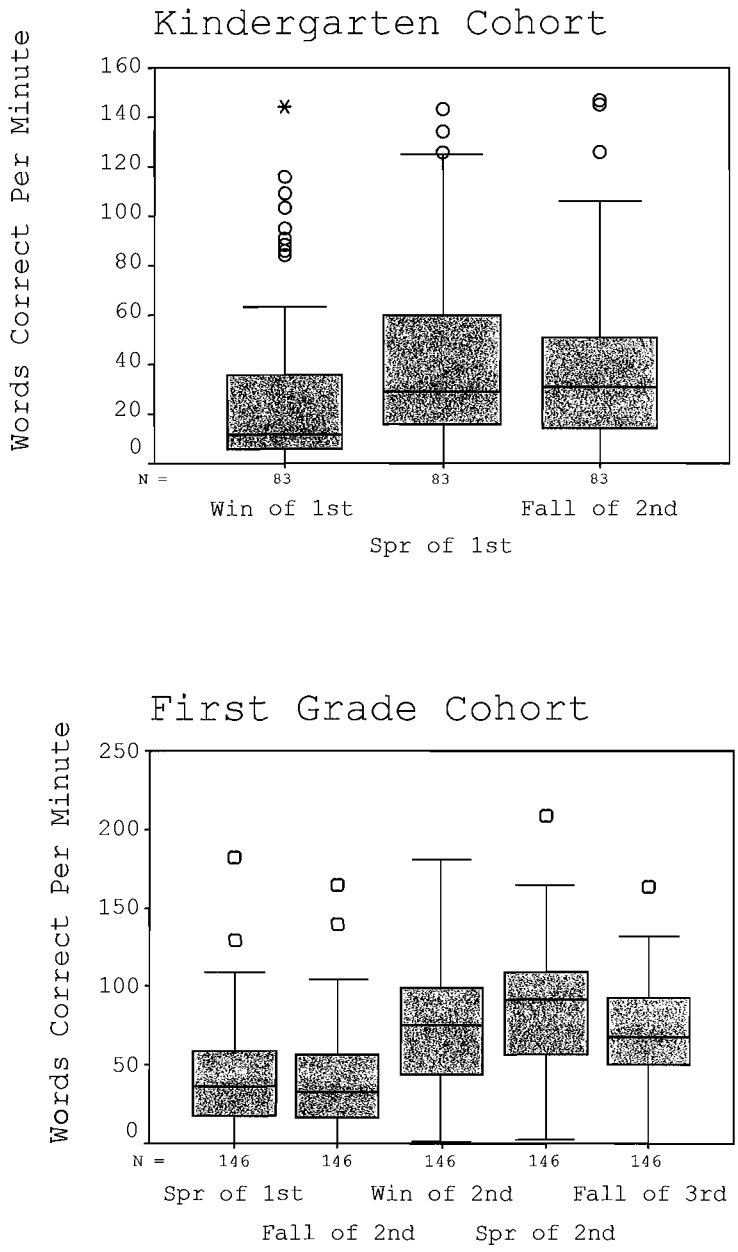


Figure 6. Boxplots of DORF growth.

spring of first grade to fall of second grade for both cohorts and a more noticeable dip in scores from spring of second grade to fall of third grade for the first grade cohort. The first grade cohort's scores over time also indicate a positive curvilinear growth pattern in second grade as observed by Baker et al. (2008).

Given the nonlinear growth in DORF observed for both cohorts, two methods for modeling growth were applied and results were compared. First, raw score change on DORF was used as a broad measure of change in oral reading fluency performance over time. Raw score change for the kindergarten cohort equaled the difference in scores from the middle of first grade to the end of first grade. Raw score change for the first grade cohort equaled the difference in scores from the end of first grade to the end of second grade. Descriptive statistics for raw score change are presented in Table 7.

As a second approach to modeling growth, DORF scores were replaced with corresponding Lexile measures (MetaMetrics, Inc., 2008). Lexile measures are numeric representations of individuals' reading abilities. MetaMetrics, Inc. recently conducted a study linking Lexile measures to specific scores on DORF in grades 1-3. The transformation of DORF scores into Lexile measures accounted for changes in passage difficulty on the DORF measures across grade levels and consequently improved the linearity of scores (see Figure 7). A regression line was fit to the Lexile measures across measurement occasions. The slope of the regression line served as a second outcome variable for predicting growth on Lexile measures of oral reading fluency (LORF) from initial performance on measures of the alphabetic principle.

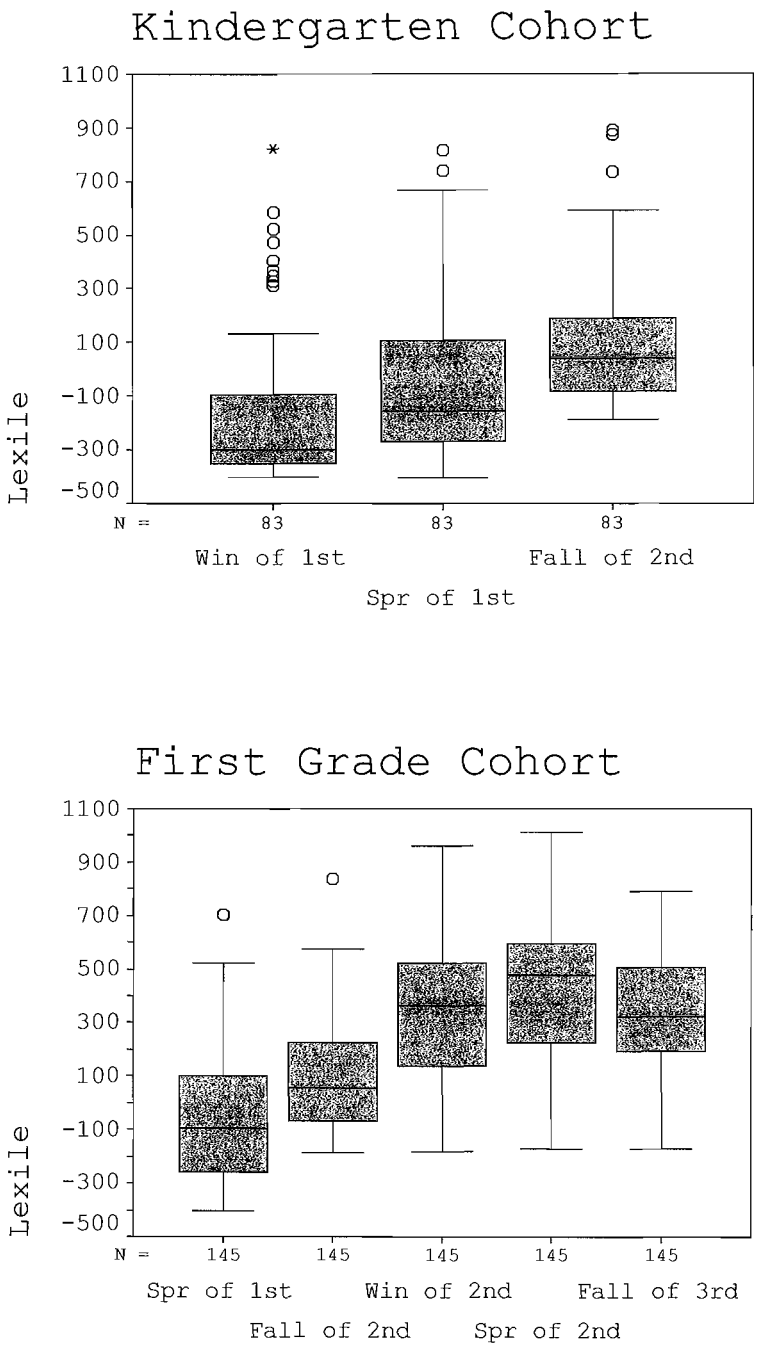


Figure 7. Boxplots of LORF growth.

Research Question Four

When predicting change/growth on repeated measures of oral reading fluency over time:

(a) How much variance is explained by each measure of the alphabetic principle?

(b) Does combining a measure of letter-sounds in nonsense words with a measure of letter-sounds in isolation add significantly to the variance explained?

(c) Does combining a measure of nonsense words recoded with a measure of letter-sounds add significantly to the variance explained?

This research question was addressed using two different outcome variables (raw score change and slope).

Predicting Raw Score Change

Strength of predictive relations for each measure of the alphabetic principle with raw score change. Raw score change on DORF from the middle of first grade to the end of first grade was unrelated to initial performance on measures of the alphabetic principle in the spring of kindergarten (p >.30 for CLS, WRC, and PDE; p = .07 for LSF). Raw score change from the end of first grade to the end of second grade was significantly related to initial performance on LSF and CLS in the spring of first grade (.28 correlation with LSF, .17 correlation with CLS, p <.05) but not significantly related to WRC (p = .054) or PDE (p = .18). Linear regression was used to examine the power of LSF and CLS as individual predictors of raw score change in the spring of first grade. LSF explained 8% of the variance in raw score change ($F(1,144) = 11.96, p < .001$), and CLS explained 3% of the variance in raw score change ($F(1,144) = 4.13, p < .05$).

Strength of predictive relations of letter sounds in isolation and letter sounds in nonsense words with raw score change. A multiple regression model was used to examine the contributions of LSF and CLS as combined predictors of raw score change. Although neither measure was a significant individual predictor at the end of kindergarten, when combined, LSF and CLS accounted for 8% of the variance in raw score change from the middle to end of first grade ($F(2,80) = 3.44, p < .05$). Results from the analysis of LSF and CLS at the end of first grade are presented in Table 19. Adding a measure of letter-sounds in nonsense words (CLS) to a measure of letter sounds in isolation (LSF) did not improve the prediction of raw score change for the first grade cohort. As an independent predictor, CLS explained nearly 3% of the variance in raw score change, but this predictive relation was not maintained when LSF was entered into the model. These results indicate that (a) at the end of kindergarten, both types of measures are needed to predict raw score change from the middle to the end of first grade, (b) at the end of first grade, only a measure of letter-sounds in isolation is needed to predict raw score change from the end of first to the end of second, and (c) the percent of variance explained in raw score change was minimal for both cohorts (8%).

Strength of predictive relations of letter sounds in nonsense words and nonsense words recoded with raw score change. A multiple regression model was used to examine four combinations of predictors: (a) LSF and WRC, (b) LSF and PDE, (c) CLS and WRC, and (d) CLS and PDE. For the kindergarten cohort, one combination of predictors explained significant variance in raw score change. When entered simultaneously into the prediction model, LSF and WRC explained 11% of the variance in raw score change for

the kindergarten cohort ($F(2,80) = 4.74, p < .05$). For the first grade cohort, none of the four combinations of predictors significantly improved upon the variance explained by LSF and CLS as independent predictors of raw score change.

Table 19

Model Summary for Predicting Raw Score Change from Measures of Letter-Sounds in Isolation and Letter-Sounds in Nonsense Words in First Grade

Model	R^2	R^2 Change	F Change	df	MS Residual
1. LSF	.077		11.96**	144	326.13
2. LSF, CLS	.079	.002	0.282	143	327.76
1. CLS	.028		4.13*	144	343.37
2. CLS, LSF	.079	.051	7.86*	143	327.76

Note. $n = 146$

*$p < .05$, **$p < .001$.

Summary of results for predicting raw score change. For the kindergarten cohort, two combinations of predictors, LSF-CLS and LSF-WRC, made significant contributions to the variance explained in the model, but no single measure held significant predictive relations. For the first grade cohort, LSF and CLS were significant individual predictors, but the predictive strength was not increased by adding any combination of predictors to the model. Overall, the predictors explained a minimal proportion of the variance in raw score change across both cohorts, ranging from 3% to 11%.

Predicting Slope

Table 20 presents descriptive statistics for LORF scores at each measurement occasion. Visual inspection of the boxplots for LORF scores (see Figure 7) indicated improved linearity of growth across each measurement occasion except for the fall of third grade. Even after accounting for a change in passage difficulty, mean performance in the fall of third grade was below mean performance in the spring of second grade. Although a quadratic model would represent this dip in performance, such a model would not adequately describe the first four data points which appear linear in nature. Therefore, the fifth measurement occasion was dropped from analyses and a linear model was applied to the first four measurement occasions.

Table 20

Descriptive Statistics for Lexile Measures of Oral Reading Fluency (LORF)

	Kindergarten Cohort (n = 83)			First Grade Cohort (n = 145)		
Measurement Occasion	*M*	*SD*	*Range*	*M*	*SD*	*Range*
Winter of 1st	-184.30	274.87	-403-824	—	—	—
Spring of 1st	-34.73	301.11	-403-815	-47.04	241.80	-403-704
Fall of 2nd	98.51	237.67	-186-890	92.84	204.44	-186-1022
Winter of 2nd	—	—	—	337.01	263.55	-179-1139
Spring of 2nd	—	—	—	436.30	252.81	-171-1014
Fall of 3rd	—	—	—	338.05	225.27	-173-1020

Note. One student's scores were removed from the data set because ORF scores were too high to correspond to Lexiles.

HLM (Bryk & Raudenbush, 1992) was used to evaluate the predictor-criterion relationships between measures of the alphabetic principle and growth on LORF across measurement occasions. The student version of HLM6 (Raudenbush, Bryk, Cheong & Congdon, 2004) was used to analyze measurement occasions and student level variables. A Level 1 unconditional model was applied to the measurement occasions nested within each student, and a Level 2 conditional model was applied to the student-level variables (i.e. initial performance on measures of the alphabetic principle). Equations for each level are

Level 1 model:

$$\text{LORF}_{ti} = \Pi_{0i} + \Pi_{1i}(\text{MONTHS}_{ti}) + e_{ti}$$

Level 2 model:

$$\Pi_{0i} = \beta_{00} + \beta_{01}(\text{PREDICTOR}_i) + r_{0i}$$

$$\Pi_{1i} = \beta_{10} + \beta_{11}(\text{PREDICTOR}_i) + r_{1i}$$

where LORF is the outcome variable, MONTHS represents the measurement occasions for each student, and PREDICTOR represents initial performance on a measure of the alphabetic principle. These models allowed for the simultaneous analysis of the influence of time and scores on measures of the alphabetic principle on LORF growth.

Model assumptions and estimation method. Model assumptions were tested by inspecting scatterplots of the distribution of LORF scores across time and conducting a test of homogeneity of variance. Scatterplots of the distribution of LORF scores over four measurement occasions indicated that variance increased over time. The test of homogeneity of variance was significant (kindergarten cohort: χ^2 (1,82) = 121.67, p

=.003 (first grade cohort: χ^2 (1,144) = 253.77, p <.001)) and confirmed that variance in LORF scores is not homogeneous over time. Violation of this assumption may limit confidence in the parameter estimates generated by the HLM model. Full Maximum Likelihood estimation was employed for all analyses. The intercept was defined as the first LORF measurement occasion for each cohort. For the kindergarten cohort, the intercept is the LORF score at the middle of first grade (i.e. 8 months after the predictor measures were administered). For the first grade cohort, the intercept is the LORF score at the end of first grade (i.e. at the same point in time as the predictor measures).

Unconditional linear growth model. Table 21 presents the baseline statistics for modeling growth on LORF. For the kindergarten cohort, the average LORF score at the middle of first grade was -176, and the average rate of growth from one measurement occasion to the next was 31Lexiles. For the first grade cohort, the average LORF score at the end of first grade was -65, and the average rate of growth from one measurement occasion to the next was 43 Lexiles.

In the second section of Table 21, the chi-square statistic and its corresponding p value for each parameter provide an indication of whether students vary significantly in their initial LORF score and in their rate of growth (i.e. slope). Results suggest that significant between-student variance exists for initial status on LORF for both cohorts and for the rate of growth for the first grade cohort (p < .05) while between-student variance in the rate of growth for the kindergarten cohort was not significant (p = .08).

The third section of Table 21 provides an indication of the reliability of the parameter estimates. The reliability for the intercept is strong (.95 (.88)) indicating that

Table 21

Unconditional Model of Growth on Lexile Measures of Oral Reading Fluency (LORF)

Fixed Effect	Cohort	*Coefficient*	*SE*	*t Ratio*	*p Value*
Mean initial status, β_{00}	K	-176	31.42	-5.59	<0.001
	1	-65	18.66	-3.48	0.001
Mean growth rate, β_{10}	K	31	1.40	22.25	<.001
	1	43	0.99	43.24	<.001

Random Effect		*Variance Component*	*df*	χ^2	*p Value*
Initial status, r_{0i}	K	77562.07	82	1561.93	<.001
	1	44385.17	144	1195.11	<.001
Growth rate, r_{1i}	K	28.87	82	100.80	.078
	1	35.97	144	194.57	.003
Level-1 error, e_{ti}	K	5474.78			
	1	8075.22			

Reliability of OLS Regression Coefficient Estimate

Initial status, Π_{0i}	K	0.95
	1	0.88
Growth rate, Π_{1i}	K	0.18
	1	0.26

Note. Initial status on LORF for the kindergarten cohort is middle of first grade. Initial status for the first grade cohort is end of first grade. Growth rate is the expected increase in LORF from one benchmark period to the next (approximately 4-5 months).

n = 83 for kindergarten cohort, 145 for first grade cohort.

exploration of individual differences in initial LORF performance is warranted. However, the same conclusion cannot be made for the exploration of individual differences in slope. The reliability estimates for slope (.18 (.26)) indicate that variance in the growth parameters may be attributable to model error rather that actual individual differences in rates of growth over time.

Given the low degree of confidence that individual differences in slope are meaningful and warrant further investigation, additional analyses comparing the predictive strength of different measures of the alphabetic principle are not reported. However, exploratory analyses were conducted to gain additional insight into the modeling of individual differences in growth on LORF.

Exploratory analyses for the modeling of growth. Each measure of the alphabetic principle was entered at level 2 as an individual predictor of growth. For the kindergarten cohort, all four predictors had significant influence on the slope at $p < .05$. For the first grade cohort, only LSF had significant influence on the slope ($p = .007$), and its influence was minimal. Overall, PDE at the end of kindergarten had the greatest degree of influence on the slope. PDE in the spring of kindergarten and spring of first grade will be used as an example for modeling individual differences. Table 22 presents the results of a conditional model where PDE is entered as a single, level-2 predictor of LORF growth.

Results in Table 22 indicate that when initial performance on PDE at the end of kindergarten is 0, initial performance on LORF at the middle of first grade is predicted to be -354 Lexiles. A score of 1 on PDE at the end of kindergarten indicates that LORF performance at the middle of first grade would be -311 Lexiles (-354 + 43). The rate of

Table 22

Linear Model of Growth on LORF Predicted by PDE

Fixed Effect	Cohort	*Coefficient*	*SE*	*t Ratio*	*df*	*p Value*
Model for initial status, Π_{0i}						
BASE, β_{00}	K	-353.94	24.17	-14.65	81	<.001
	1	-369.21	23.85	-15.48	143	<.001
PDE, β_{10}	K	42.68	3.56	11.98	81	<.001
	1	20.12	1.37	14.70	143	<.001
Model for growth rate, Π_{1i}						
BASE, β_{10}	K	36.11	1.60	22.63	81	<.001
	1	42.14	1.99	21.19	143	<.001
PDE, β_{11}	K	-1.17	0.24	-4.96	81	<.001
	1	0.04	0.11	0.31	143	.759

Note. Initial status on LORF is middle of first grade (end of first grade). Initial status on PDE is end of kindergarten (end of first grade). Growth rate is the expected increase in LORF from one benchmark period to the next (approximately 4-5 months). BASE, β_{00} = intercept when PDE = 0. BASE, β_{10} = slope when PDE = 0.

n = 83 for kindergarten cohort, 145 for first grade cohort.

growth on LORF across measurement occasions equals 36 Lexiles when PDE equals 0 and 35 Lexiles when PDE equals 1 (i.e. 36 - 1). The *t* Ratio and corresponding *p* value for these parameter estimates suggest that they are significantly different from 0. Results for the first grade cohort indicate that performance on PDE in spring of first grade significantly influences level of performance on LORF at the same point in time but not rate of growth on LORF from end of first to end of second ($p = .76$). Given these results, a conclusion *could* be drawn that better performance on PDE at the end of kindergarten leads to less growth on LORF over time, but better performance on PDE at the end of first grade has no effect on growth. To further examine PDE's influence on growth in kindergarten, individual growth trajectories were plotted based on the results from Table 22.

Figure 8 presents a graph of growth trajectories for students at varying levels of initial status on PDE at the end of kindergarten. Mean performance on PDE for this sample was 4.21 with a standard deviation of 5.06. Growth trajectories for students whose scores were at the mean, one standard deviation above the mean, and at the floor for PDE are plotted on the graph. All three growth trajectories appear to have nearly parallel slopes. The difference in rate of growth for students who scored a 0 on PDE at the end of kindergarten compared to students who scored a 9 on PDE (one standard deviation above the mean) is 11 Lexiles per measurement occasion. Remembering that approximately four to five months pass between measurement occasions and that 11 Lexiles equates to approximately 1.5 words read correctly, it appears as though this difference in growth rates is not educationally meaningful.

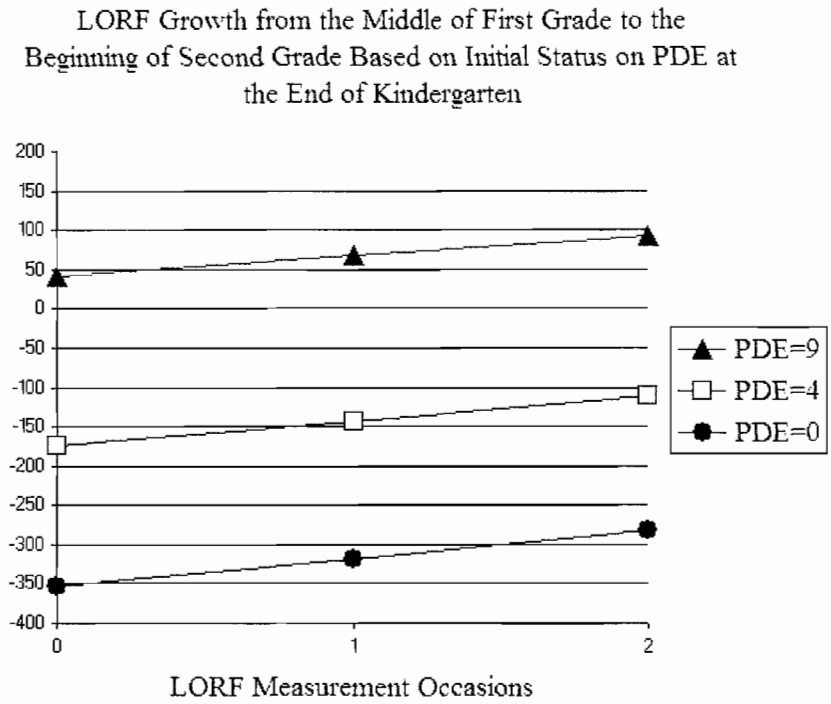


Figure 8. Growth trajectories based on initial performance on PDE. Measurement occasion 0 = middle of first grade, 1 = end of first grade, 2 = beginning of second grade.

Summary of results for predicting slope. HLM was utilized to evaluate the predictor-criterion relationships between measures of the alphabetic principle and slope on LORF across measurement occasions. Results from the unconditional model indicate significant between-student variance for the rate of growth for the first grade cohort (p <.05) but not for the rate of growth for the kindergarten cohort (p = 0.08). Reliability estimates for slope (.18 (.26)) indicate that variance in these growth parameters may not be attributable to actual individual differences in rates of growth over time. Although exploratory analyses indicated that measures of the alphabetic principle explained significant portions of the individual differences in growth, the reliability estimates for

the slopes and visual inspection of growth trajectories call into question the meaningfulness of these results.

Summary of Results

Research questions one, two, and three addressed the prediction of outcomes for measures of word reading fluency, oral reading fluency, and reading comprehension using sequential regression. Each measure of the alphabetic principle explained significant portions of variance in each criterion when entered independently into the regression model ($p < .05$) (Research Question 1).

Adding a measure of letter-sounds in nonsense words (CLS) to a measure of letter-sounds in isolation (LSF) significantly improved upon the variance explained in reading outcomes. Both types of alphabetic principle measures were necessary for the prediction of two of nine criterion measures while a single measure of letter-sounds in nonsense words (CLS) was sufficient for the prediction of the seven remaining criterion measures (Research Question 2).

Including a measure of nonsense words recoded (i.e. WRC or PDE) with a measure of letter-sounds (i.e. LSF or CLS) added significantly to the variance explained in reading outcomes (Research Question 3). When predicting word reading fluency and oral reading fluency, both types of alphabetic principle measures had unique and significant predictive relations. When predicting performance on measures of vocabulary and comprehension administered 20 months later, results were mixed with some results

indicating that each measure may not have unique and significant predictive relations in the presence of the other measure.

The fourth research question addressed the prediction of growth on measures of oral reading fluency over time. Raw score change and slope are different conceptualizations of individual growth on oral reading fluency. Both conceptualizations of growth were used as predictive criterion in this study to investigate the methodology of modeling growth. When using raw score change as a predictive criterion, some predictor measures had no significant associations with the criterion, and those measures that did have significant associations explained minimal variance (3-11%) as individual or combined predictors of raw score change. When investigating slope as a predictive criterion, results indicated that individual differences in slope were minimal and not very reliable making the prediction of these differences difficult.

CHAPTER V

DISCUSSION

The primary purpose of this study was to examine the power of different measures of the alphabetic principle for predicting later reading development. Results indicate that scores from each measure held significant power for determining later reading development, and combining scores from certain measures maximized predictive relations. This study also investigated the methodology of modeling growth on oral reading fluency to gather evidence for the use of growth as a meaningful criterion measure of reading development. Results indicate that raw score change on DIBELS Oral Reading Fluency (DORF) within a grade level was largely unrelated to scores from measures of the alphabetic principle. Transforming DORF scores into Lexile measures improved the linearity of the data within and across grade levels, but individual differences in slope were not reliably detected in this sample making this criterion difficult to predict. This chapter will expand on these results by interpreting each finding as a different line of evidence contributing to evaluative judgments of the inferences and uses of scores derived from measures of the alphabetic principle. This chapter is organized into four sections: *Interpretation of Findings*, *Limitations, Directions for Future Research,* and *Conclusions.*

Interpretation of Findings

As discussed in Chapter I, the science and art of test validation requires the collection of numerous lines of theoretical and empirical evidence to support the inferences and uses of test scores (AERA, APA, & NCME, 1999). The measures selected for investigation in this study were presumed to be indicative of the alphabetic principle. The utility of measuring the alphabetic principle stems from research indicating that development of the alphabetic principle is essential for becoming a proficient reader (Stanovich, 1986; Torgesen, 2002). By measuring this construct during the early stages of reading development, educators are able to identify students in need of additional support and consequently provide these students with systematic and explicit instruction in the alphabetic principle to prevent later reading difficulties. Concurrent and predictive related evidence from previous studies, summarized in Chapter II, provided initial support for the inference that scores from measures investigated in this study are indicative of the alphabetic principle and are useful for predicting later reading development. Results from this study provide additional lines of theoretical and empirical evidence to examine the inferences and uses of these assessments in greater detail.

The first section will discuss construct-related evidence linking measures from this study to Ehri's (1999) theory of sight word development. The second section will discuss criterion-related evidence supporting the inference that measuring multiple facets of the alphabetic principle improves power for predicting later reading development. Evidence pertaining to the relevance and utility of the assessment tools will also be discussed in this section. The third section will discuss construct-related evidence

addressing the methodology of measuring growth on oral reading fluency. As discussed in Chapter I, the predictive validity of alphabetic principle measures is interpretable only if the criterion measures are considered meaningful indicators of later reading development, and additional research is needed to investigate DORF growth as a meaningful criterion measure.

Mapping Measures onto Theory

Linking measures of the production of letter-sounds and the recoding of nonsense words to a theoretical rationale for the development of the alphabetic principle provides evidence supporting the inference that these assessment tools are indicative of the construct they are presumed to measure and allows further investigation into the different facets of the construct. Convergent and divergent evidence linking measures from this study to Ehri's (1999) theory of sight word development are discussed.

Convergent evidence. Results from this study provide evidence supporting a link between measures of the alphabetic principle and phases of Ehri's (1999) theory of sight word development. Specifically, a measure of letter sounds in isolation (i.e. Letter Sound Fluency (LSF)) appeared to quantify a behavior that develops before behaviors quantified by the other assessment tools suggesting that this measure might be more indicative of an earlier phase of Ehri's (1999) theory of development.

In describing the transition from the pre-alphabetic to the partial alphabetic phase of development, Ehri (2005) stated:

> Children progress to the partial alphabetic phase when they learn the names or sounds of alphabet letters and use these to remember how to read words. However, they form connections between only some of the letters

> and sounds in words, often only the first and final sounds, which are easier to detect. (p. 173)

Ehri (2005) draws attention to the fact that detecting individual letters within words is a skill that develops over time beginning with the identification of the first and final sounds in words. Most students learn letter-names and letter-sounds in isolation before they begin to recognize these same letter-names and letter-sounds in the context of words (Texas Education Agency, 2002). Given this understanding of early reading development, producing letter-sounds in isolation is presumed to be an easier skill that develops before the skill of producing letter-sounds in nonsense words. Predictive relations and frequency distributions for performance on LSF and the Correct Letter Sounds (CLS) scores on Nonsense Word Fluency (NWF) in this study support this hypothesis.

Predictive relations can be influenced by both *temporal* distance and *developmental* distance. As can been seen in Table 10, correlation coefficients for most measures of the alphabetic principle decrease as the temporal distance of the criterion measures increases (e.g. correlation coefficients for PDE are .81 in spring of first grade and .75 in spring of second grade). Evidence supporting the influence of *developmental* distance is found in Tables 9 and 10. For the first grade cohort, both the word reading fluency composite and DORF were administered at the same time as the predictor measures (eliminating temporal distance as a factor), yet correlations are slightly stronger with the word reading composite than with DORF for each of the predictors. This pattern makes sense when considering that the skills measured by the predictors are more

developmentally proximal to the skill of reading lists of words than to the skill of reading connected text.

This line of reasoning can be applied to examining more subtle developmental differences among the skills measured by LSF and CLS. Both measures were administered at the same point in time, yet LSF had lower predictive relations than CLS for eight of nine criterion measures (three of four in kindergarten and all five in first grade) and comparable predictive relations for the ninth criterion (GRA+DE Vocabulary Composite for kindergarten cohort). LSF also had lower predictive relations than WRC and PDE for seven of nine criterion measures, but to ease interpretation, LSF will only be compared to CLS.

Re-examination of the frequency distributions for LSF and CLS helps to clarify why these differences may have occurred (see Appendix). The distribution for CLS had a positive skew with outliers in excess of three standard deviations beyond the mean and was similar to the distributions of scores for most of the criterion measures. Given the restricted range of scores on LSF, this measure may not have distinguished above-average performance from exceptional performance on the criterion measures to the same degree as CLS did. To translate in terms of student performance, results suggest that students with advanced skills in the spring of kindergarten or first grade may have reached a ceiling on the LSF measure producing the maximum number of letter-sounds feasible in one-minute. However, these same students did not reach the same performance ceiling when producing letter-sounds in nonsense words, presumably because these students could elect to blend multiple sounds together or recode whole

words which are more efficient approaches to the task. Had LSF been given earlier in kindergarten when the skill of producing letter-sounds is presumed to be less developed, students may have been less likely to reach a ceiling. Speece and Ritchy (2005) noted the same ceiling effect for LSF in their sample of first graders and suggested that "once students begin reading words with some competency, word-level skills (both accuracy and fluency) become the best predictors of fluency" (p. 396).

Divergent evidence. Interestingly, evidence from this study indicates a pattern of differences in predictive relations for WRC and PDE that does not align with the theoretical rationale discussed in Chapter I. WRC's predictive relations were hypothesized to be similar to the estimates for PDE since both measures involve the recoding of nonsense words. However, this study found that WRC had lower predictive relations than PDE for three of the four criterion measures for the kindergarten cohort and all five criterion measures for the first grade cohort.

Surmising from these findings that skills required for performing on PDE must develop before skills required for performing on WRC does not make sense when considering the actual behaviors being measured. The first 14 nonsense words for PDE are comparable to all of the nonsense words presented on the NWF measure (i.e. two and three-letter, decodable words) so performance across these two measures should be quite similar at least when WRC performance is compared to performance on the first 14 nonsense words on PDE. Frequency distributions clarify why the theoretical rationale for the alignment of predictive relations for WRC and PDE was not supported (see Appendix).

In the spring of kindergarten both WRC and PDE were greatly impacted by floor effects making neither measure sensitive for detecting individual differences in later reading development for students whose scores were in the lower half of the distribution. In the spring of first grade, the floor effect for PDE went away, while WRC's floor effect remained. Investigation of descriptive statistics and scatterplots revealed that many students who scored 0 on WRC in the spring of first grade were able to recode at least some of the basic decodable nonsense words on PDE. Outlier scores indicate that a few students were very efficient and skilled at recoding both basic and advanced nonsense words (based on their performance on PDE), but elected to produce letter-sounds rapidly instead of recoding words on the NWF measure.

Research is currently being conducted to explore why the unexpected low performance on WRC is occurring. One hypothesis is that the directions for the task may prompt students to elect to read sound-by-sound even if they are able to recode. Dynamic Measurement Group, the developers of the DIBELS, is looking at the effects of revised test-directions on WRC scores to investigate this hypothesis. Another related hypothesis is that the approach to the NWF task varies across different samples, possibly due to the way the task is understood and presented by school personnel. While Fuchs et al. (2004) reported that lower performing students in their sample increased their efficiency with using a sound-by-sound strategy for completing NWF without progressing to blending or recoding sounds, another study found that students with high scores on CLS tended to recode the nonsense words (Harn et al., 2008). Again, revised directions and possibly additional training for test users could improve the consistency of results across samples.

Using Single or Combined Measures of the Alphabetic Principle

Criterion-related evidence from this study indicates that two behaviors associated with development of the alphabetic principle (i.e. production of letter-sounds from nonsense words and the recoding of nonsense words) should be measured in order to maximally predict later reading development for students at the end of kindergarten and at the end of first grade. This finding is based on two lines of evidence, each of which will be discussed in the following paragraphs followed by a discussion of the relevance and utility of the assessment tools that measure these behaviors.

Evidence from this study indicates that measuring letter-sounds in the context of nonsense words appears to have more power for predicting reading outcomes than measuring letter sounds in isolation for students at the end of kindergarten and at the end of first grade. The power of CLS as a single predictor exceeded the power of LSF as a single predictor for eight of the nine criterion measures. Frequency histograms indicate that the CLS measure had greater breadth for detecting individual differences in student performance than LSF which was impacted by ceiling effects for both cohorts. When both measures were entered into sequential regression models, CLS overshadowed LSF in the prediction of two of the four criterion measures for the kindergarten cohort and all five of the criterion for the first grade cohort. When LSF did explain unique and significant variance in the presence of CLS, the variance explained in the criterion measures was minimal (4% for DORF in spring of first grade and 7% for the GRA+DE Vocabulary Composite).

Although neither WRC nor PDE has much utility in the spring of kindergarten for differentiating between students who may have some risk for later reading difficulties and students who may have moderate to severe risk for later reading difficulties, they do appear useful for differentiating between students with average and advanced skills in the spring of kindergarten. In spring of first grade, performance on PDE was more normally distributed making it more appropriate for differentiating between levels of risk-status, while performance on WRC remained positively skewed. PDE made unique and significant contributions to the prediction of every criterion measure beyond the contribution of CLS, while WRC had added value in predicting some criterion measures but not others. Predictive relations from this study indicate that PDE may be the preferred measure of nonsense word recoding to use in spring of kindergarten and spring of first grade, and adding scores from PDE to scores from CLS maximizes the prediction of later reading development.

Although predictive relations point toward the use of CLS and PDE as combined predictors in the spring of kindergarten and first grade, other factors related to validity should also be evaluated when considering appropriate inferences and uses of test scores (Good & Jefferson,1998; Messick,1995). First, alternative hypotheses for these findings should be investigated (Kane, 2001). One alternative hypothesis is that, had more students who could recode nonsense words elected to do so on the NWF measure, neither PDE nor WRC may have explained additional, unique variance beyond CLS because the CLS scores would have increased in accordance with the WRC scores (e.g. recoding a three-letter nonsense word yields 3 points for CLS and 1 point for WRC). Recoding is

likely to be the most efficient approach to the task that yields the greatest number of letter-sounds produced in one-minute. Additional research is needed to determine if there is added value in measuring the skill of recoding (either on WRC or PDE) for a sample of students who are encouraged to recode on NWF or if the CLS score sufficiently encompasses this skill (i.e. higher scores on CLS resulted from efficient recoding). Harn et al. (2008) provided initial evidence that there is added value in directly quantifying the recoding of words in addition to indirectly accounting for this skill with the CLS score, but further investigation is needed.

The validity argument for selecting one assessment tool over another has focused primarily on criterion-related validity. Good and Jefferson (1998) wrote "Criterion-related validity is desirable *but not sufficient* for determining the validity of any measure. The *relevance* and *utility* of any measure also must be considered when establishing the validity of that measure" [emphasis in original] (p. 67-68). Relevance refers to the degree to which information derived from the assessment "*directly* addresses or answers the questions posed in the assessment process" and utility refers to the "*benefits* of assessment relative to its *costs"* [emphasis in original] (p. 68). This study has framed measures of the alphabetic principle as relevant for predicting later reading development and potentially useful for remediating alphabetic principle deficits by calling attention to students most in need of additional instruction. Closer examination of issues pertaining to relevance and utility sheds light on differences between the PDE and NWF measures.

While both PDE and NWF appear relevant for identifying students in need of additional instructional support, only NWF is relevant for setting goals for these students.

Unlike PDE, which compares scores to a set of national norms, NWF has empirically based benchmarks that identify the score a student should meet or exceed by certain grade levels and times of year in order to maximize the student's likelihood for continued reading success (Good et al., 2002). Knowing exactly where a student should be by when appears more useful for ensuring that the assessment serves as a catalyst for remediation than simply knowing how the student currently compares to a national sample. Although the administration directions for NWF are still being examined and the WRC scoring option is still being researched, preliminary evidence indicates that students who are able to produce 50 correctly letter-sounds in one minute by the middle of 1st grade but are unable to recode at least some of those sounds into 15 complete and correct nonsense words may be in need of additional instruction in blending (Dynamic Measurement Group, no date). The general guideline of 15 WRC by the middle of first grade combined with the established benchmark of 50 CLS serves as a goal for instruction and as a reference point for evaluating the effectiveness of instruction. Given that some students who *can* recode may elect *not* to recode on the NWF measure, teachers may need to compare performance on NWF to performance on other assessments to confirm if students need additional instruction in blending. NWF is also relevant for evaluating a student's response to additional instruction. NWF has more that 20 equivalent forms for use as progress monitoring assessments that can be given weekly to monitor growth, while PDE only has two equivalent forms making it inappropriate for use as a progress monitoring tool.

Each of the factors discussed in the preceding paragraph contributes to a validity argument that supports the use of both scores on the NWF measure instead of the use of CLS and PDE *if* the purposes for using the scores include developing goals for remediation and monitoring response to instruction. If the sole purpose is to identify students in need of additional instruction in the alphabetic principle, then the use of CLS with PDE would be supported given evidence that PDE added to the predictive relations of CLS to a greater degree than WRC added to CLS for this sample.

Methodology of Measuring Growth on Oral Reading Fluency

While ORF and many measures of reading comprehension have established research bases supporting their use as meaningful criterion measures of later reading development, *growth* on oral reading fluency over time is not well established as a meaningful predictor or criterion due to the challenges associated with its measurement. This study made two modest contributions to research exploring the methodology of modeling growth within and across grade levels.

Raw Score Change

First, this study revealed that raw score change for DORF may not be an appropriate criterion measure of later reading development. Results indicate that initial performance on measures of the alphabetic principle had little if any effect on students' amount of change in DORF scores from the middle to the end of first grade or from the end of first grade to the end of second grade. Raw score change was the only criterion measure in this study that did not correlate significantly with *all four* predictor measures. The two correlations that were significant were in the low range. The relevance of raw

score change as a criterion measure of later reading development is called into question by these results.

Previous research has also questioned the use of change scores as indicators of individual growth because there is no way to know the shape of this growth (Rugutt, 2001), and change scores tend to be less reliable than either of the original scores they were derived from (Lord, 1958). Evidence from this study indicates that as students get older, the variance in their performance on DORF increases. Heteroscedasticity introduced in the change score from unequal variance in performance on the two measurement occasions may have also impacted results. Another possibility is that no actual individual differences in change occurred in this sample. This hypothesis will be discussed in greater detail in the next section.

Lexile Measures of Oral Reading Fluency

Transforming DORF scores to Lexile measures improved the linearity of growth within and across grade levels, at least through spring of second grade, making linear approaches to modeling data more appropriate. Previous research suggests that DORF growth is typically nonlinear both within and across grade levels (Baker et al., 2008; Briggs, Good, & Rogers, 2007). Summer vacation and a change in the readability levels of the passages from one grade to the next are hypothesized to influence the nonlinear pattern of the data. If a polynomial equation were applied to the raw DORF score data set to accurately represent the acceleration and deceleration of DORF growth over time, each change in slope for the model would result in loss of one degree of freedom from error, consequently reducing the power of the equation for predicting outcomes. Linear models

are preferred in statistical analyses of growth because such models preserve the power of the model for predicting beyond the range of the data, and this study provides evidence that linear models may be appropriate once DORF scores are transformed to Lexile measures. However, this study also calls attention to concerns regarding the reliability of estimating individual differences in growth on LORF.

Shin, Espin, Deno, and McConnell (2004) identified three factors that are know to influence the reliability of growth parameters in HLM: "(a) number of data points, (b) heterogeneity of true growth parameters of individual students, and (c) measurement error" (p. 142). The following paragraphs will discuss the degree to which each of these factors may have influenced the reliability of the growth parameter in this study.

The reliability estimate for the growth parameter was stronger for the first grade cohort which had four data points than for the kindergarten cohort which only had three data points. It stands to reason that additional data points could have improved the reliability of the growth estimate, but its worth noting that these data points stretch across five to eight months of instruction (excluding summer vacation) during which differences in reading fluency growth should be easily detected by three or four measurement occasions.

Shin et al.'s (2004) second factor pertains to the hypothesis mentioned previously that true growth rates did not differ among students in this sample. This hypothesis is highly concerning because it suggests that students who started out with less skills than their peers continued to lag behind without showing any sign of catching up. If the purpose of the assessment is to identify students most in need of additional instruction

and this instruction actually occurs, than students with low performance on initial measures of the alphabetic principle should have growth rates that are steeper than the growth rates of students who started out with skills in the average range. When individual differences in growth are not detected because true growth rates are actually quite similar across students of varying skill levels, this could indicate that the assessment tool may not be an effective catalyst for improving outcomes for students in need of additional support.

The third factor identified by Shin et al. (2004) is measurement error. In a recent study investigating the measurement of growth on oral reading fluency across three different progress monitoring tools, Ardoin and Christ (2009) found that student performance on the DIBELS progress monitoring passages resulted in greater standard error of individual slopes and greater standard error of the estimate than was found for performance on the other progress monitoring tools. Ardoin and Christ attributed this finding to greater differences in passage difficulty among the DIBELS progress monitoring passages than among the other passage sets. Transforming DORF scores to Lexile measures in this study should have negated at least some of the effects of variation in passage difficult, but Ardoin and Christ's study does provide compelling evidence that measurement error may have influenced the reliability estimate for growth rates in this study.

Summary of Measuring Growth on Oral Reading Fluency

Findings from this study contributed to the research base on measuring *growth* in oral reading fluency by calling attention to potential limitations in (a) the use of raw score

change as a broad index of growth and (b) in the reliability estimate for growth parameters in hierarchical linear modeling of LORF progress. This study also introduced a promising new approach for accounting for changes in passage difficulty: the transformation of DORF scores to Lexile measures. With additional research on the methodology of measuring growth, this construct could serve as a meaningful criterion for evaluating measures of the alphabetic principle. Important questions that could be addressed in future research include: (a) Do students who produce letter-sounds efficiently but do not recode have slower rates of growth on oral reading fluency than students who do recode? (b) Do the lowest performing students on measures of the alphabetic principle have the fastest rates of growth on oral reading fluency (suggesting that the assessment is serving as a catalyst for effective instruction that helps students catch up to their peers)?

Summary of Findings

Each finding discussed brought to light different lines of validity evidence gathered from this study. The first section focused on construct-related evidence. This study provided initial evidence that relative differences in performance on measures of the alphabetic principle seemed to map on to different phases of development identified by Ehri (1999) and subsequently increased understanding of the alphabetic principle as a multi-faceted construct. Predictive relations for WRC diverged from the theoretical rationale, which lead to further investigation of performance on this measure in comparison to performance on a similar measure, PDE. This study found that many students who could recode (based on their PDE scores) elected not to recode on the NWF

measure calling into question WRC's relevance as a measure of recoding skill. The second section focused on criterion-related evidence and extended the discussion to include evidence pertaining to relevance and utility. Two behaviors were identified (production of letter-sounds in nonsense words and recoding of nonsense words) as important to measure in order to maximize the strength of predictive relations. Different validity arguments can be made for the selection of one assessment tool over another based on the purpose for using the tool. The third section focused on construct-related evidence for the measurement of growth on ORF. This study found that raw score change did not function as a meaningful criterion of later reading development and that transforming DORF scores to Lexile measures improved the linearity of individual growth patterns within and across grade levels; however low reliability of the growth estimate was prohibitive to the prediction of growth. These two findings are modest contributions to research investigating the methodology of measuring growth.

Limitations

Multiple limitations were present in this study. First, predictive relations for each measure of the alphabetic principle are constrained to spring of kindergarten and spring of first grade, and generalization of these estimates to other grade levels or times of year may be limited. Skills associated with the alphabetic principle develop rapidly from entry into kindergarten through second grade when students are likely to have progressed to the *consolidated* phase of development (Ehri & Snowling, 2004). Measures appropriate for use as screening tools in the spring of kindergarten may not be appropriate for use as

screening tools in second grade when most students may have already mastered the skills being assessed.

There were a number of limitations with data collection that should be noted. Although all data collectors received training on the administration of each measure, reliability checks were conducted for only 3% of the test administrations during stage I of data collection. Best practice suggests 20% to be the minimum and 33% to be preferred (Kennedy, 2005). During stages II and III, schools were responsible for administering and scoring the DORF benchmark assessments and entering these scores into the DIBELS data base. Although reliability checks may have been done by the schools, no reliability checks on the DORF assessments were conducted as part of this study.

In addition to conducting more reliability checks, a more sensitive process for analyzing reliability could have been utilized. This study used percent agreement for the total raw score as the criterion for evaluating interrater reliability for training and data collection and Pearson product moment correlation coefficients as an additional measure of interrater reliability during data collection. As Kennedy (2005) pointed out, both approaches neglect to investigate whether observers ever agreed on the occurrence of individual instances of behavior. Using percent agreement for the occurrence and nonoccurrence of errors could have provided a more sensitive index for evaluating interrater reliability.

Two limitations of this study posed concerns for the accuracy of the WRC scores. Although data collectors were trained on how to mark the NWF protocols to indicate whether or not a student recoded a word (i.e. drawing a continuous line under each sound

within the word), the data collectors did not calculate a score for words recoded at the time of data collection. These scores were derived after data collection was completed, and inferences had to be made on occasion as to the intent of ambiguous markings by the original data collector. Additionally, although the correlation of rater's scores on WRC was quite high (.99), investigation of actual agreement on the total raw score revealed that there was only 84% agreement for the total number of words recoded on the ten protocols that were shadow-scored. This suggests that additional training was needed to clarify the marking procedure on the protocols for when students recoded words.

Directions for Future Research

In addition to the recommendations for future research already discussed, additional research is needed to broaden the scope of test score validation for measures of the alphabetic principle. First, research is needed to look at the predictive relations of these measures across a wider distribution of grade levels, possibly from preschool through second grade to determine when floor and ceiling effects are likely to occur for different measures and to make further connections between the behaviors being measured and the phases of development they are presumed to represent. For example, future research could investigate the strength of predictive relations of LSF in late preschool and early kindergarten to determine if LSF could have utility for detecting which students are beginning to transition out of the *pre-alphabetic* phase into the *partial alphabetic* phase. It would also be important to investigate whether CLS has utility at

these times as well, or if floor effects on CLS might prevent it from detecting individual differences in skills for students in the lowest quartile of the distribution.

A second issue that needs to be addressed in future research is the stability of the WRC scores. No research is currently available to determine if students typically recode the same number of words across equivalent forms of the NWF measure administered within a short time-span. Moving forward with investigating other lines of validity evidence for the WRC score may not be warranted if the scores are highly unreliable from one measurement occasion to the next.

Results from this study highlighted several questions pertaining to the methodology of modeling individual differences in ORF growth that will need to be investigated in future research. First, can individual differences in LORF slopes be reliability detected in other samples? Would other measurement material result in more reliable estimates of growth? If reliable growth estimates can be obtained, do statistically significant individual differences in slopes actually translate to educationally meaningful differences in rates of growth for students? Finally, is deceleration in LORF scores for fall of third grade found in other samples? After transforming DORF scores to Lexile measures for this sample, a significant dip in performance remained in the fall of 3rd grade for the first grade cohort. This dip was not seen in the fall of 2nd grade for the kindergarten cohort. Additional research is needed to investigate the transformation of DORF scores to Lexile measures in the fall of third grade. This dip could reflect an actual regression of skills that occurred for one grade level but not the other, or it could be indicative of concerns with score transformation at this level.

Conclusions

"As a salient social value, validity assumes both a scientific and a political role that can by no means be fulfilled by a simple correlation coefficient between test scores and a purported criterion " (Messick,1995, p. 742). Although analyses in this study focused primarily on predictive relations, these estimates are recognized as just once type of evidence that contributes to an evaluative judgment about the appropriateness of inferences and uses for tests. This study has also included construct-related evidence linking measures of the alphabetic principle to a theory of word reading development, evidence pertaining to the relevance and utility of different assessments of the alphabetic principle, and evidence pertaining to the methodology of measuring growth on oral reading fluency. Although additional research is needed to formulate a strong validity argument for the use of one assessment tool over another tool for identifying students in need of additional instruction in the alphabetic principle, there are some general conclusions that can be taken from this study and from existing research to guide teachers in the assessment of skills indicative of the alphabetic principle. First and foremost, previous research provides compelling evidence that the alphabetic principle is an essential component of learning to read (NRP, 2000). By measuring the development of the alphabetic principle through nonsense word reading, teachers are able to gather important information to identify students who are at-risk for reading difficulties (Rack et al. 1992).

Findings from this study suggest that kindergarten and first grade teachers should measure their students' letter-sound knowledge in the context of nonsense words. Although initial instruction and practice may focus on pairing a single sound with a single letter, students should quickly begin to generalize this skill to the context of words because this is the context that serves as the foundation for reading. As discussed in Chapter I, teachers should use nonsense words instead of real words in assessments of the alphabetic principle to ensure that scores reflect students' decoding skills and not their skill for recalling words from memory. This is not to say that instruction should necessarily focus on nonsense words. On the contrary, decoding of sounds in words is most rewarding when students are able to make connections to the pronunciation and meaning of real words. Instruction and assessment in recoding nonsense words supplements instruction and assessment in authentic text by ensuring that students have the necessary decoding skills to make connections for the pronunciation and meaning of more challenging words that are not yet a part of students' sight vocabularies.

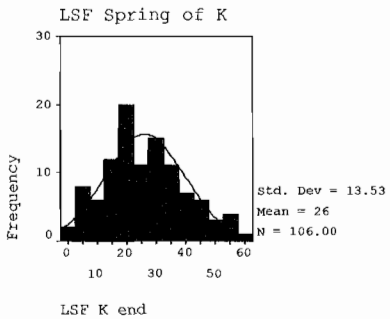
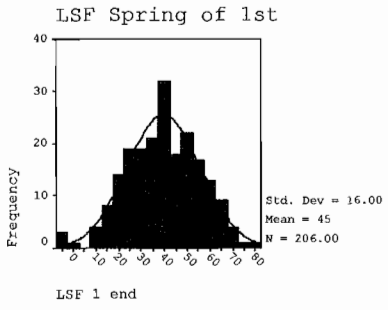
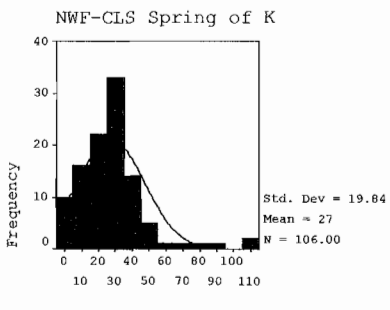
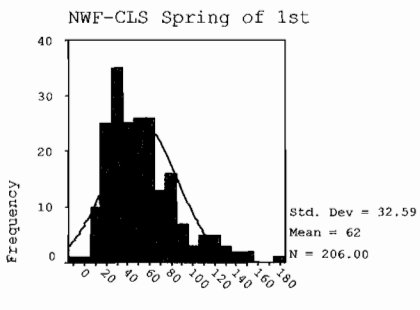
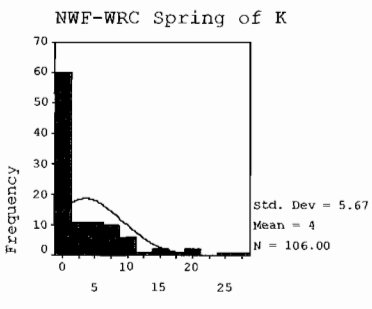
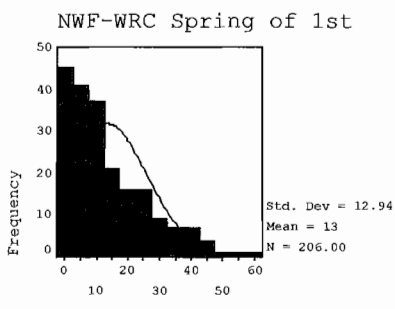
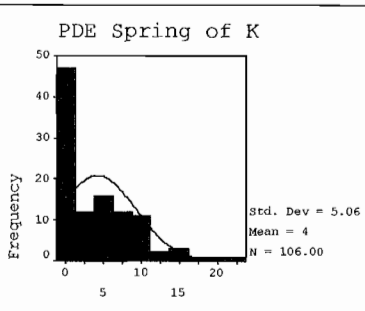
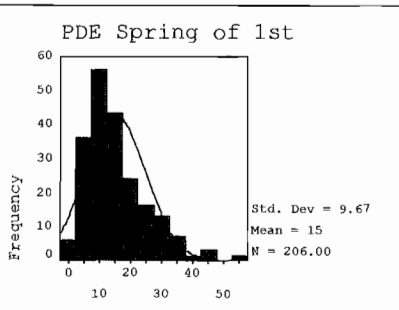
Findings from this study also provide initial evidence for recommending that teachers monitor students' progression of specific skills pertaining to the development of the alphabetic principle from identifying isolated sounds in words, to partially blending sounds within words, and finally to completely and correctly recoding whole words. Students who lag behind their peers in this progression of skills may benefit from targeted instruction to move them toward more efficient approaches to decoding.

Finally, to maximize the utility of assessments of the alphabetic principle for improving student outcomes, teachers should select tools that help them identify

ambitious goals for student learning and monitor progress toward those goals. Of central importance for any educational assessment tool is that scores from that tool serve as catalysts for improving outcomes for students.

APPENDIX

FREQUENCY HISTOGRAMS FOR PREDICTOR VARIABLES

Spring of Kindergarten	Spring of First Grade
	
	
	
	

REFERENCES

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

Adams, M. J. (2001). Alphabetic anxiety and explicit, systematic phonics instruction: A cognitive science perspective. In S. B. Neuman & D. K. Dickinson (Eds.), *Handbook of early literacy research, Volume 1* (pp. 66-80). New York, NY: The Guilford Press.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (1999). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association

American Guidance Service (2001). *Group reading assessments and diagnostic evaluation.* Circle Pines, MN: American Guidance Service.

Baker, S. K., Smolkowski, K., Katz, R., Fien, H., Seeley, J. R., Kame'Enui, E. J., et al. (2008). Reading fluency as a predictor of reading proficiency in low-performing, high-poverty schools. *School Psychology Review*, *37*(1), 18.

Briggs, R. N., Good, R. H., & Rogers, F. (2007, March). *Missing phonics? What difference does it make for reading trajectories?* Paper presented at the annual convention of the National Association of School Psychologists, New York, NY.

Bryk, A. S., & Raudenbush, S. W. (1992). *Hierarchical linear models: Applications and data analysis methods*. Thousand Oaks, CA: Sage Publications.

Chard, D. J., Stoolmiller, M., Harn, B. A., Wanzek, J., Vaughn, S., Linan-Thompson, S., et al. (2008). Predicting reading success in a multilevel schoolwide reading model: A retrospective analysis. *Journal of Learning Disabilities*, *41*(2), 174.

Children's Educational Services. (1987). *Test of reading fluency*. Minneapolis, MN: Author.

Ardoin, S. P., & Christ, T. J. (2009). Curriculum-based measurement of oral reading: Standard errors associated with progress monitoring outcomes from DIBELS, AIMSweb, and an experimental set. *School Psychology Review*, *38*(2), 266-283.

CTB/McGraw-Hill. (2002). *TerraNova, second edition: Reading subtest.* Monterey, CA: CTB/McGraw-Hill.

Deno, S. L., Mirkin, P. K., & Chiang, B. (1982). Identifying valid measures of reading. *Exceptional Children, 49*, 36-45.

Dynamic Measurement Group (n.d.). *New fields added to the revised version of DIBELS 6th edition*. Retrieved July 12, 2009, from https://dibels.uoregon.edu/measures/NewFields_nwf.pdf

Ehri, L. C. (1997). Sight word learning in normal readers and dyslexics. In B. Blachman (Ed.), *Foundations of reading acquisition and dyslexia: Implications for early intervention* (pp. 163-189). Mahwah, NJ: Lawrence Erlbaum Associates.

Ehri, L. C. (1999). Phases of development in learning to read words. In J. Oakhill & R. Beard (Eds.), *Reading development and the teaching of reading: A psychological perspective* (pp. 79–108). Malden, MA: Blackwell Publishing.

Ehri, L. C. (2005). Learning to read words: Theory, findings, and issues. *Scientific Studies of Reading, 9*(2), 167-188.

Ehri, L. C., & Roberts, T. (2006). The roots of learning to read and write: Acquisition of letters and phonemic awareness. In D. K. Dickinson & S. B. Neuman (Eds.), *Handbook of early literacy research, Volume 2* (pp. 113–131). New York, NY: The Guilford Press.

Ehri, L. C., & Snowling, M. J. (2004). Developmental variation in word recognition. In C. A. Stone, E. R. Silliman, B. J. Ehren, K. Apel (Eds.), *Handbook of language and literacy: Development and disorders* (pp. 433-458). New York, NY: The Guilford Press.

Elliott, J., 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*(1), 33-49.

Foorman, B. R., Francis, D. J., Shaywitz, S. E., & Shaywitz, B. A. (1997). The case for early reading intervention. In B. A. Blachman (Ed.), *Foundations of reading acquisition and dyslexia: Implications for early intervention* (pp. 243-264). Mahwah, NJ: Lawrence Erlbaum Associates.

Francis, L., Shaywitz, S. E., Stuebing, K. K., Shaywitz, B. A., Fletcher, J. M. (1996). Developmental lag versus deficit models of reading disability: A longitudinal, individual growth curves analysis. *Journal of Educational Psychology, 88*(1), 3-17.

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*(3), 239-256.

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*(1), 7-22.

Good, R. H., Baker, S., & Peyton, J. (2009). Making sense of nonsense word fluency: Determining adequate progress in early first grade reading. *Reading & Writing Quarterly: Overcoming Learning Difficulties*, 25(1), 33-56.

Good, R. H., & Jefferson, G. (1998). Contemporary perspectives on curriculum-based measurement validity. In M. R. Shinn (Ed), *Advanced applications of curriculum-based measurement* (pp.61-88). New York, NY: The Guilford Press.

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

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

Good, R. H., Kaminski, R. A., Shinn, M., Bratten, J., Shinn, M., Laimon, L., et al. (2004). *Technical adequacy and decision making utility of DIBELS* (Technical Report No. 7). Eugene, OR: University of Oregon.

Good, R. H., Simmons, D. C., & Kame'enui, E. J. (2001). The importance and 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*(3), 257-288.

Good, R. H., Simmons, D. C., Kame'enui, E. J., Kaminski, R. A., & Wallin, J. (2002). *Summary of decision rules for intensive, strategic, and benchmark instructional recommendations in kindergarten through third grade* (Technical Report No. 11). Eugene, OR: University of Oregon.

Harcourt Educational Measurement (2007) *AIMSweb Test of early literacy: Letter Sound Fluency*. San Antonio, TX.

Harcourt Educational Measurement (2002). *Stanford Achievement Test.* San Antonio, TX.

Harn, B. A., Stoolmiller, M., & Chard, D. J. (2008). Measuring the dimensions of alphabetic principle on the reading development of first graders: The role of automaticity and unitization. *Journal of Learning Disabilities*, *41*(2), 143.

Juel, C. (1988). Learning to read and write: A longitudinal study of 54 children from first through fourth grades. *Journal of Educational Psychology*, *80*(4), 437-447.

Kame'enui, E. J., & Simmons, D. C. (2001). Introduction to this special issue: The DNA of reading fluency. *Scientific Studies of Reading*, *5*(3), 203-210.

Kame'enui, E. J., Good III, R. G., & Harn, B. A. (2005). Beginning reading failure and the quantification of risk: Reading behavior as the supreme index. In W. L. Heward, T. E. Heron, N. A. Neef, S. M. Peterson, D. M. Sainato, G. Cartledge, I. Gardner, R., L. D. Peterson, S. B. Hersh, & J. C. Dardig (Eds.), *Focus on behavior analysis in education: Achievements, challenges, and opportunities* (pp.69-88). Upper Saddle River, NJ: Prentice Hall/Merril.

Kane, M. T. (2001). Current concerns in validity theory. *Journal of Educational Measurement*, *38*(4), 319-342.

Kennedy, C. H. (2005). *Single-case designs for educational research*. Boston, MA: Pearson Education, Inc.

Larson, S. Hammill, D., & Moates, L. C. (1999). *Test of Written Spelling, Fourth edition.* Austin, TX: ProEd.

Messick, S. (1986). The once and future issues of validity: Assessing the meaning and consequences of measurement. In H. Wainer & H. Braun (Eds.), *Test validity* (pp 33-45) Hillsdale, NJ: Lawrence Erlbaum Associates.

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist*, *50*(9), 741-749.

Metametrics, Inc. (2008). *Linking DIBELS Oral Reading Fluency with the Lexile framework for reading*. Retrieved July 12, 2009 from http://www.dibels.org/papers/DIBELSLexilesLinkingReport.pdf

National Center for Education Statistics (2007, September 25). *The Nation's report card: Reading 2007* Retrieved May 23, 2008, from http://nces.ed.gov/pubsearch/pubsinfo.asp?pubid=2007496

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.

No Child Left Behind Act of 2001. (PL 107-110). *Title I, part B, Student reading skills improvement grants, subpart 1, Reading First.*

Pearson Education, Inc. (2007). *GRADE scoring and reporting software, Version 3.4.* Minneapolis, MN: Pearson Education, Inc.

Rack, J. P., Snowling, M. J., & Olson, R. K. (1992). The nonword reading deficit in developmental dyslexia: A review. *Reading Research Quarterly*, *27*(1), 28-53.

Raudenbush, S.W., Bryk, A.S., Cheong, Y.F., & Congdon, R.T. (2004). *HLM 6: Student version.* Retrieved April 10th, 2007 from http://www.ssicentral.com/hlm/student.html.

Riedel, B. W. (2007). The relation between DIBELS, reading comprehension, and vocabulary in urban first-grade students. *Reading Research Quarterly*, *42*(4), 546-567.

Ritchey, K. D., & Speece, D. L. (2006). From letter names to word reading: The nascent role of sublexical fluency. *Contemporary Educational Psychology*, *31*(3), 301-327.

Riverside Publishing (2000). *Gates-MacGinitie Reading Test, 4th Edition.* Chicago, IL: Riverside Publishing

Rugutt, J. K. (2001, April). A study of individual patterns of longitudinal academic change: Exploring the structural equation modeling (SEM) and hierarchical linear modeling (HLM). Paper presented at the Annual Meeting of the American Educational Research Association, Seattle, WA.

Salvia, J. & Yssledyke, J. (2003). *Assessment in special and inclusion education* (9th Edition). New York: Houghton Mifflin.

Share, D. L., & Stanovich, K. E. (1995). Cognitive processes in early reading development: Accommodating individual differences into a model of acquisition. *Issues in Education*, *1*(1), 1–57.

Shin, J., Espin, C. A., Deno, S. L., & McConnell, S. (2004). Use of hierarchical linear modeling and curriculum-based measurement for assessing academic growth and instructional factors for students with learning difficulties. *Asia Pacific Education Review*, *5*(2), 136-148.

Speece, D. L., & Case, L. P. (2001). Classification in context: An alternative approach to identifying early reading disability. *Journal of Educational Psychology*, *93*(4), 735-49.

Speece, D. L., Mills, C., Ritchey, K. D., & Hillman, E. (2003). Initial evidence that letter fluency tasks are valid indicators of early reading skill. *Journal of Special Education*, *36*(4), 223-234.

Speece, D. L., & Ritchey, K. D. (2005). A longitudinal study of the development of oral reading fluency in young children at risk for reading failure. *Journal of Learning Disabilities*, *38*(5), 387-399.

Simmons, D. C. & Kame'Enui, E. J. (1999). *Curriculum maps: Mapping instruction to achieve instructional priorities in beginning reading: Kindergarten – grade 3.* Eugene, OR: University of Oregon, College of Education, Institute for the Development of Education Achievement.

Stage, S. A., Sheppard, J., Davidson, M. M., & Browning, M. M. (2001). Prediction of first-graders' growth in oral reading fluency using kindergarten letter fluency. *Journal of School Psychology*, *39*(3), 225-237.

Stanovich, K. E. (1986). Matthew effects in reading: Some consequences of individual differences in the acquisition of literacy. *Reading Research Quarterly*, *21*(4), 360-407.

Texas Education Agency (2002). *The alphabetic principle*. Retrieved July 12, 2009 from http://www.readingrockets.org/article/3408

Torgesen, J. K. (2000). Individual differences in response to early interventions in reading: The lingering problem of treatment resisters. *Learning Disabilities Research and Practice*, *15*(1), 55-64.

Torgesen, J. K. (2002). The prevention of reading difficulties. *Journal of School Psychology*, *40*(1), 7-26.

Torgesen, J. K., & Burgess, S. R. (1998). Consistency of reading-related phonological processes throughout early childhood: Evidence from longitudinal-correlational and instructional studies. In J. L. Metsala & Ehri, L. C. (Eds.) *Word recognition in beginning literacy* (pp. 161-188). Mahwah, NJ: Lawrence Erlbaum Associates.

Torgesen, J. K., Rashotte, C. A., & Alexander, A. W. (2001). Principles of fluency instruction in reading: Relationships with established empirical outcomes. In M. Wolf (Ed.), *Dyslexia, fluency, and the brain* (pp. 333–355). Timonium, MD: York Press.

Torgesen, J. K., Wagner, R. K., & Rashotte, C. A. (1999). *TOWRE, Test of Word Reading Efficiency: Examiner's manual.* Austin, TX: PRO-ED.

Troia, G. A. (2004). Phonological processing and its influence on literacy learning. In C. A. Stone, E. R. Silliman, B. J. Ehren, K. Apel (Eds.), *Handbook of language and literacy: Development and disorders* (pp. 271–301). New York, NY: The Guilford Press.

Wagner, R. K., & Torgesen, J. K. (1987). The nature of phonological processing and its causal role in the acquisition of reading skills. *Psychological Bulletin, 101*(2), 192-212.

Wagner, R. K., Torgesen, J. K, Rashotte, C. A., Hecht, S. A., Barker, T. A., Burgess, S. R., et al. (1997). Changing relations between phonological processing abilities and word-level reading as children develop from beginning to skilled readers: A 5-year longitudinal study. *Developmental Psychology, 33*(2), p. 468-479.

Woodcock, R. W. (1987). *Woodcock Reading Mastery Test* (Rev. ed.). Circle Pines, MN: American Guidance Service.

Woodcock, R. W., Johnson, M. B. (1989). *Woodcock-Johnson Psych-Educational Battery-Revised* (WJ-R). Chicago, IL: Riverside Publishing.

Woodcock, R.W., McGrew, K.S., & Mather, N. (2001). *Woodcock Johnson III Tests of Achievement: Examiner's manual.* Chicago, IL: Riverside Publishing.