

INVESTIGATING THE RELIABILITY AND VALIDITY OF THE CONSORTIUM
ON READING EXCELLENCE (CORE) PHONICS SURVEY

by

Lorilynn Brandt

A dissertation submitted in partial fulfillment
of the requirements for the degree

of

DOCTOR OF PHILOSOPHY

in

Education

Approved:

Parker Fawson, Ed.D.
Major Professor

Sylvia Read, Ph.D.
Committee Member

D. Ray Reutzel, Ph.D.
Committee Member

Cindy Jones, Ph.D.
Committee Member

Jamison Fargo, Ph.D.
Committee Member

Byron R. Burnham, Ed.D.
Dean of Graduate Studies

UTAH STATE UNIVERSITY
Logan, UT

2009

Copyright © Lorilynn Brandt 2009

All Rights Reserved

ABSTRACT

Establishing the Reliability and Validity of the Consortium on Reading
Excellence (CORE) Phonics Survey

by

Lorilynn Brandt, Doctor of Philosophy

Utah State University, 2009

Major Professor: Parker Fawson, Ed.D.
Department: Teacher Education and Leadership

Phonics was identified as one of the critical components in reading development by the National Reading Panel. Over time, research has repeatedly identified phonics as important to early reading development. Given the compelling evidence supporting the teaching of phonics in early reading, it is critical to make sure that instructional decisions in phonics are based upon valid and reliable assessment data. This study examined the psychometric properties of the Consortium on Reading Excellence (CORE) Phonic Survey and was designed to establish instrument validity and reliability. Analyses indicated moderate to very strong validity and reliability coefficients. Additionally, a *D* study using generalizability analyses data identified the optimal assessment administration protocol for the CORE Phonics Survey to minimize the error variance and maximize the reliability under absolute and relative decision-making conditions.

(101 pages)

ACKNOWLEDGMENTS

Many people have played a part in helping me complete this doctoral degree. I have realized that the saying “it takes a village” is definitely applicable here.

I would like to give special thanks to my chair, Dr. Parker Fawson, for his extended help during these few years in offering knowledge, advice, resources, and encouragement. It has been a pleasure working with him. In addition, Dr. Ray Reutzler kept this study on track. He helped me push through some roadblocks with the statistical analysis and spent much time helping me understand the skill of research writing. His willingness to share expertise and knowledge has been invaluable and I will always appreciate having had this opportunity to be mentored by him. Thanks to Dr. Jamison Fargo who went out of his way to help interpret the statistical findings in his last hours here at Utah State University and for sticking with this committee for the duration. Thanks also to Dr. Cindy Jones who came on board later in the process; she has shared insights of her own recent journey through this that has helped immensely. Dr. Sylvia Read has taught me in many ways through her instruction, feedback, and her own writing. Finally, to John Smith who started out as a member of this committee before moving to Texas. Although not officially on the committee now, he has continued to support me and offer encouragement from afar. He has been a great mentor to me throughout the years.

I would also like to thank the many professors I have had throughout my doctoral coursework. I have had a rich learning experience. In addition, as part of that, the fellow students I have rubbed shoulders with have made the journey more enjoyable.

Warm thanks to my friends and colleagues in the Department of Teacher Education and Leadership for the privilege it has been teaching and associating with you for the last several years (1999-2009). My time as an instructor at USU has truly been a rewarding experience and I am blessed to have had this opportunity.

Finally, special heart-felt thanks to all the family and friends who have supported me throughout this process. Especially my dad, who having done this himself, was able to encourage me to the end. All my love to my children, Parker and Allison, who woke up each morning asking if I had finished my paper. It is for them that I even started this process. Their continuous patience through the years has been a source of motivation and inspiration. I love you, dearly.

Lorilynn Brandt

CONTENTS

	Page
ABSTRACT.....	iii
ACKNOWLEDGMENTS	iv
LIST OF TABLES.....	vii
LIST OF FIGURES	viii
CHAPTER	
I. INTRODUCTION.....	1
Phonics Assessment is Important	3
Validity and Reliability of Assessments	5
Problem Statement	6
Research Questions	8
II. REVIEW OF LITERATURE	11
Results of the NRP Report on Decoding.....	11
Four Purposes of Reading Assessment	16
Psychometric Properties of Reading Assessment: Validity and Reliability.....	20
The CORE Phonics Survey	24
Summary	27
III. METHODOLOGY	28
Research Questions	28
Instrument	29
Design.....	31
Procedures	36
Data Analysis	42
Summary	44
IV. RESULTS.....	45
Results	45
Summary	70

	Page
V. DISCUSSION	72
Discussion of Validity Analysis	72
Discussion of Reliability Analysis	76
Implications for Instruction	80
Limitations.....	81
Recommendations for Future Research	82
REFERENCES	84
CURRICULUM VITAE.....	88

LIST OF TABLES

Table	Page
1. Criterion Validity: Comparison of Utah State Core Curriculum and CORE Phonic Survey	50
2. Criterion Validity: Comparison of Core Reader with CORE Phonics Survey	52
3. Descriptive Analysis of Confirmatory Factor Analysis	54
4. Correlations Matrix for Confirmatory Factor Analysis	54
5. Goodness-of-Fit Tests	56
6. Standardized Model Results	58
7. <i>R</i> -Square	59
8. Cronbach's Alpha Analysis for CORE Phonics Survey	61
9. Estimated Variance Components and Standard Errors for Part 1 (Sections A-D).....	62
10. Estimated Variance Components and Standard Errors for Part 2 (Sections E-L).....	62
11. <i>D</i> Study Phi Coefficients	67
12. Standard Error of Measurement	70

LIST OF FIGURES

Figure	Page
1. Confirmatory factor analysis construct validity.....	53
2. G-stat for alphabet skills and letter sounds: Part 1	68
3. G-Stat for reading and decoding skills: Part 2	68

CHAPTER I

INTRODUCTION

Early reading success or failure sets the stage for future academic success or failure. Failure to acquire early literacy skills is a potential indicator of future reading difficulties in school (Jenkins & O'Connor, 2002). However, the ability to read is not just a necessary task for school. Success in reading is also important for lifelong achievement; how well a child learns to read may determine future opportunities, including career possibilities and the ability to accomplish the basic activities of daily life (McCardle & Chhabra, 2004). The National Institute of Child Health and Human Development (NICHD) convinced lawmakers that “the failure to learn to read reflects an educational and public health problem because this lack of success affects emotional health and overall well-being” (p. 4). Thus, currently in the United States, there is an increased focus on making sure all students are proficient readers and have the necessary skills to be successful.

One very important literacy skill that students should know to be effective readers is phonics. Phonics has been identified by many as one of the crucial early literacy skills that make a significant difference in reading success (Cunningham & Cunningham, 1992; National Reading Panel [NRP], 2000). Phonics is the relationship between letters (graphemes) and their corresponding sounds (phonemes) (Adams, 1990; Ehri, 1998). A proficient reader is able to use this knowledge of letter/sound relationships to decode unknown words in text (National Research Council [NRC], 1999; Rasinski & Padak, 1996).

Throughout the history of reading instruction, phonics has been a topic of great discussion (Smith, 2002). Differing views have emerged in the last few decades among researchers regarding its importance in literacy instruction (Chall, 1967; Erhi, 1998; Flesch, 1955; Goodman, 1980; Smith, 1979). Some have claimed that phonics instruction is not an important element of learning to read while others claim that it is a very valuable skill for early readers. Research findings, however, have consistently supported the effectiveness of phonics instruction for early readers. One well-known review of research concerning reading and phonics is Chall's *Learning to Read: The Great Debate*. Her conclusion from this comprehensive review was that instruction in phonics led to better achievement in reading. This conclusion has been supported in many subsequent research studies and reports (e.g., Adams, 1990; Anderson, Hiebert, Scott, & Wilkinson, 1985; Balmuth, 1982; Dykstra, 1968; Foreman & Moats, 2004).

In the report *Becoming Nation of Readers*, it stated that “on the average, children who are taught phonics get off to a better start in learning to read, than those who are not taught phonics (Anderson et al., 1985, p. 37). Additionally, phonics knowledge is the single best predictor of reading comprehension (Stanovich, 1990; Vellutino & Scanlon, 1991), which is the ultimate goal of reading.

Perhaps the most influential document supporting the critical importance of phonics instruction is that of the NRP Report (2000). In 1998, the U.S. Congress commissioned a panel of experts to review the current literature on reading and determine the most effective teaching methods and approaches to see whether specific instructional practices were linked to reading success. To meet this challenge, the panel adopted the

meta-analytic technique of comparing effect sizes from all studies since 1970 that used an experimental or quasi-experimental design with a control group or a multiple-baseline method and met specific rigorous research criteria.

Of those research studies reviewed, 1,373 studies were directed to phonics, 38 of those met the research criteria established by the NRP, and 66 comparisons were made. Phonics was shown to be one of the critical components of reading instruction for both early readers and older readers and those students who received decoding instruction showed positive benefits in reading performance. Even the critics of this report showed that phonics instruction outperformed treatment conditions in which a more typical or moderate level of phonics instruction was provided” (Camilli, Vargas, & Yurecko, 2003, p. 34). The findings of this report are critical in establishing the importance of phonics instruction in reading education.

Given that phonics knowledge is shown to be so important in beginning reading acquisition, it is crucial that phonics concepts be taught in the classroom as part of an effective reading instruction program. In order for teachers to effectively include phonics in their instruction, they need to know which concepts students know and which they do not. Phonics assessments provide teachers with important information which can help them screen and diagnose students’ phonics instructional needs as well as progress monitor the effectiveness of a variety of phonics interventions.

Phonics Assessment is Important

Assessments have a significant role in helping teachers determine the needs of

students to inform instructional decision making and intervention selection. They provide documentation of students' performance and progress, so that instruction can be quickly changed or modified before the student falls too far behind their reading goals.

A valid and reliable phonics assessment can give teachers valuable information. First, phonics assessment helps to pinpoint specific areas of student need so instruction and practice can be appropriately focused. Second, it helps provide screening and diagnostic information throughout the year. Third, assessment provides evidence of the effectiveness of instructional interventions (Consortium on Reading Excellence, 2004).

Knowing that assessment is an important tool for monitoring student progress, the congressional law of *No Child Left Behind 2001* included the mandate that all students in third, fifth, and seventh grade in public schools take annual statewide standardized tests. These tests are required as a means of showing evidence that students adequately progressed in reading during the school year. The ultimate goal is to identify the reading needs of students so that interventions can be made to improve student outcomes. Thus, schools and teachers are required to show documentation of student learning through these tests. The results of these high stakes assessments are used as a measure of the annual yearly progress (AYP) of the students and school and often determine the degree of federal funding they receive. States can also opt to use criterion referenced tests (CRTs) for demonstrating their AYP. Thus if CRTs are used, it is important that any phonics assessments given by teachers throughout the year provide information about what students know about the concepts to be tested on the CRT. Consequently, educators have increased interest in accessing and using valid and reliable phonics assessments

which can help identify students' needs throughout the year giving evidence that students are adequately learning and are prepared for the phonics knowledge on these and other end-of-year tests.

Significant strides have been made to more effectively assess young children's early literacy skills (Good & Kaminski 2003; Wagner, Torgesen, & Rashotte, 1999). Since the NRP report (2000) was published, educators have begun to align teaching and assessment with its findings, which revealed that effective reading instruction should include concepts and strategies that help students to develop phonological awareness skills, alphabetic understanding, reading fluency, vocabulary and comprehension. However, unless the tests that are selected for use are appropriately assessing these skills, teachers cannot be sure that students are indeed learning and progressing. Appropriate phonic assessments, would be defined as those that are easy to use, valid, and reliable.

Validity and Reliability of Assessments

Validity refers to whether a test truly assesses what it claims to (Gall, Gall, & Borg, 2007) and that the construct being assessed is appropriate, accurately represented, and meaningful (Rathvon, 2004). To be considered a valid test, both experts and empirical evidence must support the construct. Reliability refers to how consistently the test measures the construct and is repeatable. That means that the test results remain the same regardless of the rater, occasion or test format. If a test is not valid or reliable, the results cannot be considered representative of a student's knowledge nor can it be relied upon for making accurate decisions for monitoring students' progress. Therefore,

schools and teachers need to know the validity and reliability of tests when selecting them for use.

Many teachers, however, may not be qualified or knowledgeable enough on this topic to investigate validity and reliability of assessments due to lack of training in this area and therefore may be using tests because they are popular or easily available. Unfortunately, this will not ultimately help students meet their literacy goals as these results may be giving wrong or inaccurate information. For example, if a phonics test is given with little validity evidence, the teacher may falsely assume that students understands phonics skills, when in reality the test is not addressing all the concepts that are part of the phonics construct. On the other hand, if a test is not reliable, the scores may vary each time it is given making it hard to identify a student's true understanding of phonics or to track progress in phonics knowledge. This is highly problematic in an era when accountability for student achievement is increasing and teachers are required to show evidence of students' progress. Therefore, it would be helpful to have valid and reliable assessments easily accessible to teachers that have already demonstrated adequate psychometric properties.

Problem Statement

Because the NRP identified phonics instruction as a critical or essential component of effective reading instruction, the need for valid and reliable assessments of phonics has resulted. One such phonics assessment that has recently been developed is the *Consortium on Reading Excellence (CORE) Phonics Survey*. It is included in a

compendium of reading assessments in the book entitled, *CORE Assessing Reading: Multiple Measures*. It is an informal test that examines various decoding concepts and skills routinely used in beginning reading (Bailey, 1967; Clymer, 1996; CORE, 2008). The CORE Phonics Survey is quick, easy to administer, and affordable. The cost is approximately \$40.00 and can be copied for use within a school. It takes approximately ten minutes to administer. It would be considered a user-friendly test.

Because the CORE Phonics Survey is very useable and assesses skills identified in the NRP report (2000) as important, it has gained much popularity with educators. It is currently being used quite extensively as part of the *Utah Reading First* instructional reform program. All of the *Reading First* schools in Utah currently use it to some degree as part of their assessment plan (interview with Rebecca Donaldson, November 2008). Several more Utah schools, not part of the *Reading First* program, also use this assessment (Cache County School District, November 2008). Moreover, it was found to be a popular assessment tool used by schools in other states across the nation. In a general internet search of Google and Yahoo, over 200,000 items surfaced that specifically reference this assessment, many of which were schools or district web pages that promoted the use of this assessment as part of their educational plan or *Reading First* proposals. Some of these states were Idaho, Utah, Colorado, Washington, California, Montana, and Hawaii. Other hits on the webpage were from colleges or universities outlining it as part of their teacher education courses.

The popularity of the CORE Phonics Survey indicates its pervasive use as a phonic assessment tool in schools and universities. Given this pervasiveness, one

wonders whether this test has sound psychometric properties and whether the scores obtained from it can be regarded as valid or reliable. The early reading progress of literally thousands of students, in Utah alone, is being evaluated with this assessment and decisions are being made about students' phonics knowledge based on the results of this test.

In an effort to pursue further information about the CORE Phonics Survey, a call was made to the company, *Consortium on Reading Instruction*, and publisher, *Area Press*, on November 15, 2008. Both confirmed that they did not have any data on the validity and reliability of the CORE Phonic Survey. Next, a library search was done to see if any empirical research had been published in educational journals. Nothing resulted from any of these searches indicating that the validity and reliability of this assessment has been previously investigated. Therefore, the purpose of this study is to address this problem by thoroughly investigating the validity and reliability of the CORE Phonics Survey. Such a study will make an important contribution to teachers and to educational research in general because without this information, teachers using the CORE Phonics Survey cannot be certain that the scores derived there from can guide decisions about students' phonics knowledge or their decoding instructional needs.

Research Questions

The research questions for this study will be driven by two different theoretical test theories. The first is *Classical Test Theory* (CTT) which assumes that every score on a test is composed of two components, the true score (the score that would be obtained if

there were no errors) and the measurement error (anything that prohibits the true score from showing). Although classical testing acknowledges error variance, it does not try to define or reduce it. Therefore it is not possible to have absolute confidence in what the true score is. Classical test theory is the testing approach that most studies of reliability and validity depend on (Reynolds, Livingston, & Willson, 2009).

Modern Test Theory (MTT) tries to determine how that error variance can be reduced rather than just acknowledging it. Advantages of modern test theory are that (a) it allows researchers to estimate reliability of each measure rather than assuming all are equally reliable, (b) it yields various measure of goodness of fit for the overall model, (c) it compares alternative explanatory models systematically to test hypothesis about which factors influence observed correlations in the matrix and how these interrelate, and (d) it provides a way of partitioning the variance of the measure into separate trait, method and error components (Grimm & Yarnold, 2000). Both of these testing theories will be addressed as part of this study. Each section will be outlined as to whether validity or reliability issues are being address and whether the selected test is one of Classical Testing Theory or of Modern Testing Theory.

Validity Research Questions

Classical Test Theory

1. What is the evidence for consensus or content validity of the CORE Phonics Survey as measured by convergence or agreement among expert reviewers?
2. What is the evidence of criterion validity for the CORE Phonics Survey as

measured by the percent of overlap between items on it and the phonics items in the Utah State Core Curriculum and the Scott Foresman basal reading series?

Modern Test Theory

1. What is the evidence for construct validity on the CORE Phonics Survey using a confirmatory factor analysis to validate a hypothesized two-factor model?

Reliability Research Questions

Classical Test Theory

1. What is the evidence of interrater reliability on the CORE phonics Survey as measured by a Pearson's r correlation coefficient?

2. What is the evidence of test-retest reliability on the CORE Phonics Survey as measured by a Pearson's r correlation coefficient?

3. What is the evidence of internal consistency reliability on each subtest of the CORE Phonics Survey as measured by a Cronbach's alpha coefficient?

Modern Test Theory

1. To what degree do the raters and occasions contribute to variance among scores on the CORE Phonics Survey as measured by a G study?

2. What is optimal number of occasions and raters when administering the CORE Phonics Survey to minimize error variance and optimize the reliability of the resulting rating as measured by a D study?

CHAPTER II

REVIEW OF LITERATURE

Chall's (1983) six-stage reading development model helps us understand the decoding development of early readers. This theoretical framework is the underlying premise that children need to learn phonics to make progress in reading. Therefore, it was important that supporting evidence be identified to reiterate that phonics instruction helps children to do so. This will be the first purpose of the review of literature. Then, since phonics instruction can only be as good as the assessment instruments used to inform that instruction, the second purpose of the review of literature was to define and discuss (a) the purposes of reading assessment generally, and (b) the necessary psychometric properties of valid and reliable reading assessment instruments. Finally, an investigation was done to verify whether any existing psychometric evidence exists to support the widely used phonics assessment, the CORE Phonics Survey.

Results of NRP Report on Decoding

The NRP (2000) synthesized the findings of existing studies on the effects of phonics instruction on young learners' reading achievement. Specifically, the NRP conducted a literature search of experimental studies that compared the effectiveness of systematic phonics instruction to that of unsystematic phonics instruction. Systematic phonics instruction refers to instruction that has a sequential progress and a clearly identified set of skills, concepts, or strategies to be taught. For studies to be included in the NRP meta-analysis, each had to meet rigorous criteria. Thirty-eight studies that met

these criteria and were analyzed. Effect sizes were calculated to quantify the size of the effect of the treatment and determine and decide if that effect size was statistically greater than zero at $p < .05$. An effect size is “the degree to which the phenomenon is present in the population or the degree to which the null hypothesis is false” (Cohen, 1988, pp. 9-10). Estimates of effect size provide essential information because they provide information about the relative magnitude of outcomes. The scale of significance for the effect size is defined as follows: .00 to .19 is described as trivial effect; .20 to .49, small; .50 to .79, moderate; .80 or higher, large (Cohen).

Performance on six phonics-based outcomes was considered: decoding regularly spelled real words, decoding pseudo words, reading real words that included irregular spellings, comprehending text, and reading connected text orally. Effect sizes in most of these measures were positive and significantly greater than zero, indicating that in most studies the group receiving systematic phonics instruction evidenced higher reading achievement than the control group who did not. The effect sizes were, however, significantly higher for studies with kindergarteners and first graders than with those of second through sixth graders. This finding suggests two things. First is that phonics is time sensitive information and needs to be learned early. Second, this finding suggests that phonics instruction is a better prevention from reading difficulties than it is as a cure once reading difficulties have resulted. The categories that had the strongest effect size for both early readers and later readers were decoding regular words and decoding pseudo words.

Effect sizes were also calculated for various related subsets of the studies

reviewed to break down how phonics instruction affected students' performance in various situations and across distinct characteristics. The first subset addressed was the time of the testing, either at the end of the program or the end of the year. Both effects showed to be statistically greater than zero and moderate in size, regardless of whether effects were measured at the end of the program (.41) or the end of the first year (.44). This indicates that systematic phonics instruction helps children learn to read more effectively than those who do not receive it and that the timing of the testing does not make much difference.

Phonics instruction also improved reading ability in both early readers and older readers. Effect sizes were statistically greater than zero for both, but were larger for studies with kindergarteners and first graders (0.55) than for studies with second through sixth grades (0.27). This indicates that although phonics instruction does have a positive effect on both ages, its strongest impact is in the early stages of reading acquisition.

Another subset analyzed the effect that phonics instruction had on students with differing reading abilities. Effects were statistically significant for all groups with the exception of second through sixth low-achieving students (0.15). At-risk and typically achieving readers in kindergarten and first grade both showed moderate to large differences when receiving phonic instruction. At-risk first graders were most affected by phonic instruction, with a strong effect size (0.74). Second through sixth grade low achieving student showed the smallest benefits (0.15). Effect sizes were small to moderate for the second- through sixth-grade students who are typically achieving readers (0.27) and students with reading disabilities (0.32). Thus, phonic instruction

improves reading ability more than no phonics instruction not only for beginning readers but also among typically progressing readers above first grade and older readers with reading disabilities. However, phonics did not enhance reading ability as much among low-achieving readers in Grades 2 through 6.

Studies reporting the socioeconomic status (SES) of participants were also examined. Effects were strong for children of low SES (0.66) and middle SES (0.44). This indicates that phonics instruction helps children in learning to read regardless of SES level.

Effect size results were similar when considering the sizes of the group receiving the instruction, whether it is individuals (0.57), small groups (0.34), or classrooms (0.39). This means that classroom instruction may be just as effective as tutoring without the increased expense and difficulty of one-on-one teaching.

Effects were also examined for three types of systematic phonics programs. One category was synthetic phonics, which involves teaching students to sound out letters and blend sounds into words. This effect size was strong at (0.45). Another category was to analyze and blend larger units of words such as onset, rimes, or spelling patterns. This effect size was moderate (0.34). Finally, a miscellaneous category included traditional spelling or basal programs or instruction on word analysis, which had a moderate effect size of (0.27). Effect sizes for all three categories were statistically greater than zero and would thus indicate that all of the types of systematic phonics programs were more effective than nonsystematic or not phonic program at all. As long as programs are systematic, it does not seem to matter which approach is used.

The type of instruction given to the control group in each study also varied from study to study. Effect sizes for each type of group were calculated. Control groups were categorized as basal groups and had an effect size of 0.46, regular curriculum was 0.41, whole language was 0.31, whole word was 0.51, or miscellaneous 0.46. The effect sizes for all of these was a moderate to strong positive indicating that phonics-instructed groups performed better than the other types of groups. Effect sizes were also statistically greater for groups receiving systematic phonics.

Finally, studies also differed in their design, specifically their method of assigning students to experimental groups. Effect sizes were calculated to investigate how the design impacted the outcomes. Some studies randomly assigned students to treatment and control groups while others used preexisting groups. Additionally, some studies used large sample sizes whereas others worked with fewer students. Effect sizes for the more rigorous designs using larger groups and random assignment, were as large as or larger (0.45) than the effect sizes of the less rigorously assigned groups (0.43). This is not much of a difference and would indicate that the positive effects of phonics instruction were not due to comparison with weaker designs.

In summary, findings of the NRP's (2000) meta-analysis support the conclusion that systematic phonics instruction helps all children to learn to read more quickly, easily, and with greater success than nonsystematic or no phonics instruction. The impact was significantly greater in early grades (K-1) when phonics was the method used to start students out, than in later grades (2-6) after they had made some progress in reading with other methods. The instructional approach or specific program used to teach phonics in

early grades made little difference. Synthetic phonics showed the strongest effect sizes but even these were not significantly different from the other five phonics approaches. As long as students received systematic phonics instruction, it did not make a significant difference which approach was used (McCardle & Chhabra, 2004).

Since this meta-analysis provides strong evidence that phonics instruction is an important part of early reading, then the assessment instruments used to determine the phonics instruction students' need becomes an essential part of providing effective, systematic phonics instruction. If a phonics assessment instrument is not giving the teacher accurate information on what the student needs, then phonics instruction will still not result in student progress. Effective phonics assessments need to be selected according to their (a) intended purpose and (b) the validity and reliability of the scores obtained. Both of these elements are necessary for a phonics assessment to appropriately guide the teacher to high quality phonics instruction that meets students' learning needs. When each of these is considered, the assessment can be an effective tool for planning future decoding instruction, which has previously been shown to be important for early readers. Therefore, the general purposes and psychometric properties of reading assessments are discussed below.

Four Purposes of Reading Assessment

The purpose of reading assessment is to identify the strengths and weaknesses in students' reading abilities throughout the learning process. The results from these assessments provide teachers with necessary information for effective instruction. The

National Reading First Assessment Committee (Kame'enui, 2000) concluded that a comprehensive school-wide early literacy assessment system should include assessments to accomplish four purposes: screening, progress monitoring, diagnosis, and measuring student outcomes. Because different reading assessment instruments serve different purposes, assessments should be thoughtfully selected to insure an appropriate match.

Screening Assessments

Screening assessments help identify children who are at risk for experiencing reading difficulties. They are usually done at the beginning of the school year so that students needing extra support can be promptly identified. Those who score below benchmark (appropriate grade level performance) are given additional instructional support to help get them back on track. The information obtained from screenings guides future decisions about instruction and needed interventions.

Progress Monitoring

A second purpose for early literacy assessment is *progress monitoring*. These assessments are given repeatedly throughout the year to provide a quick indication of a student's progress, checking for evidence of continual, adequate growth. If students are not sufficiently progressing, teachers can quickly adjust instruction as needed. All students' progress should be monitored regularly; however, struggling students should be monitored more frequently. If a student's results indicate a problem, teachers can administer a more comprehensive, or diagnostic assessments, to help pinpoint the exact

area of need.

Diagnosis

A third purpose for early literacy assessment is diagnosis. Diagnostic assessments provide more detailed, comprehensive information about students' skills and instructional needs. These are often administered when results from screening and progress monitoring indicate a problem. Unlike screening and progress monitoring, diagnostic assessments measure a variety of literacy component skills or abilities, which give teachers additional information needed to adjust or intensify instruction. A diagnostic assessment on phonics, for example, may include beginning with the most basic phonetic concepts (cvc words for example) and gradually progressing on to more difficult phonics concepts, such as multisyllable words. Diagnostic assessments give teachers an idea what students know about each area of phonics and where to begin or continue instruction. As teachers analyze the results of these assessments, they are able to pinpoint the exact area of needed focus. The data from these assessments helps teachers to develop tightly focused individual instruction.

Outcome Measurement

A final purpose of literacy assessment is to measure effectiveness of the instruction program on clearly identified student outcomes. Outcome measurements, such as end-of-year state core testing, provide teachers, parents, and administrators evidence of the students' overall performance for the year. These results should not be

surprising after the screening, progress monitoring and diagnostic testing done throughout the year. These previous types of testing should be predictive of outcome measures. The information gained from these assessments is often used to make policy decisions about instructional programs, funding, etc. Therefore, it is important for results to be both valid and reliable. Thus, outcome measurements are often standardized tests that have undergone evaluations of their reliability and validity. These four different types of assessment, screening, progress monitoring, diagnosis, and outcome measurement can facilitate better instruction and student learning. As teachers select the type of assessment that best fit the intended purpose, they will more easily and quickly get the desired information, which will guide instruction. Then, to verify that instruction is working, assessment must be used again to monitor student progress toward meeting literacy goals. Data gathered from assessments should be used to guide future instruction for students who struggle and to scaffold lessons to help resolve their problems early and efficiently (McCardle & Chhabra, 2004).

In order to effectively accomplish this assessment-based instruction, it is important not only to use the appropriate type of assessment but also to use valid and reliable assessments. Even if an assessment is appropriate for the purpose of the assessment, it may not be giving teachers accurate information about what the students knows and thus have limited use for planning future instruction. Teachers can only feel confident in their instructional decisions if the reading assessments being used are psychometrically sound. Because there are many reading assessments available for use, educators should strive to select only those that are valid and reliable.

Psychometric Properties of Reading Assessment:

Validity and Reliability

The second important factor to consider in selecting reading assessments is the evidence of validity and reliability of the obtained scores. Selection of valid and reliable reading assessments is essential for providing accurate and impactful early reading instruction. If a test is not valid and reliable, we cannot be sure that it is pointing us instructionally in the right direction. Validity refers to whether a test truly assesses what it is suppose to (Gall et al., 2007) and that the construct, or unobservable quality, being considered is accurately represented, appropriate, and meaningful (Rathvon, 2004). To be valid, both experts and empirical evidence must support the construct being measured. Reliability refers to how consistently the test measures the construct and is repeatable. There are several types of validity and reliability evidence that can be reported. Often validity and reliability will be established in several ways to strengthen the confidence that can be put in that test. Below, different types of validity and reliability evidences are discussed which, if done, would strengthen an assessment tool.

Types of Validity Evidence

Content Validity

Content validity demonstrates that the construct, or unobservable quality, is valid by showing the degree of agreement among specialists on the appropriateness of the items on the instrument. These specialists consider whether the items cover the breath of the content area and adequately represent a sample of the content being assessed (Gay,

1985). They also evaluate whether the test items and format are appropriate for those using the instrument. For example, a test that is intended to measure the quality of decoding instruction in first grade should not only cover material in the first grade core but should also be presented in an age appropriate manner for first graders. Another example would be that a national reading test might not be a valid measure of local reading instruction, although it might be a valid measure of national reading standards.

Criterion-Related Validity

Criterion-related validity is evidenced by comparing the instrument with some future or current criteria, thus the name criterion related. Validity based on future criteria is called predictive validity; validity based on current criteria is called concurrent validity. Questions to be answered when establishing this type of validity are “Does the measure relate to other measures or manifestations of the same construct?” or “Does the test predict an individual’s future performance in specific abilities?”

Construct Validity

Construct validity is the most important kind of validity. If a measure has construct validity it measures what it purports to measure. Establishing construct validity is a long and complex process involving defining the construct (unobservable quality) of interest and then identifying observable items that adequately measure and define that quality. A statistical analysis is then done to test and modify the assessment to show whether an agreement exists between a theoretical concept (construct) and the test. Therefore, a researcher might spend a great deal of time attempting to “define” the

construct in order to reach an acceptable level of construct validity.

Construct validity can be broken down into two sub-categories: Convergent validity and discriminate validity. Convergent validity is the similarity or agreement among ratings or information, gathered independently of one another. Discriminate validity is the lack of a relationship among measures that theoretically should not be related.

To understand whether a piece of research has construct validity, three steps should be followed. First, the theoretical relationships must be specified. Second, the empirical relationships between the measures of the concepts must be examined. Third, the empirical evidence must be interpreted in terms of how it clarifies the construct validity of the particular measure being tested (Carmines & Zeller, 1979, p. 23).

Types of Reliability Testing

Reliability refers to how consistently the test measures the construct and is repeatable. If five different examiners give an assessment, the results of all the raters would be similar on a test that was highly reliable. All measurement procedures have the potential for error, so the aim is to minimize that error. An observed test score is made up of the true score plus measurement error. Knowing the reliability of a test helps to distinguish how much of a test result is due to error in measuring and how much is due to true differences in performance or knowledge of the student. The goal of estimating reliability, or consistency, is to determine how much of the variability in test scores is due to measurement error and how much is due to variability in true scores. Once sources of

error are identified, researchers can try and eliminate that error as much as possible.

There are several standard techniques that researchers use to ensure reliability and identify and eliminate error.

Test-Retest Reliability

The test-retest method of estimating a test's reliability involves administering the test to the same group of people at least twice with a short span of time between testing. The first set of scores is then correlated with the second set of scores. If the correlations between the scores are close to 0 this would indicate low reliability while correlations closer to 1 indicate a high reliability.

Interrater Reliability

This method of testing reliability is done by comparing the scores given by different raters for the same task. If the scores given by each rater are similar, this indicates a more reliable test. Teachers can be more confident in their instructional planning if test scores were relatively the same regardless of the rater.

Internal Consistency Reliability

Internal consistency refers to whether the questions in the test consistently measure the same content. It is commonly measured using the Cronbach's alpha, which is a statistical coefficient based on the correlations between the items on the test. The closer the alpha level is to 1, the greater the reliability of the test. Generally, an alpha of .80 is considered a moderate benchmark for reliability. An alpha of .90 indicates a high reliability coefficient and .70 is a low level of reliability.

A test does not necessarily have to be low or high in both validity and reliability at the same time. A test may be low in reliability and high in validity or vice versa. For example, a phonics test that accurately measures all the phonics skills it was meant to measure would be valid, but not reliable if each rater gets a different score. Conversely, a test that shows consistent results among raters would be reliable, but if the questions do not adequately reflect the construct, the test is not valid. Teachers need to be aware of the strengths and limitations of assessments in order to correctly interpret the scores, and make sound instructional decisions. Therefore, test validity and reliability should be considered when determining how much confidence to put in test scores.

The CORE Phonics Survey

There are many reading and phonics assessments available to teachers. Some of these have undergone rigorous evaluation to establish their purpose and the validity and reliability of the scores obtained; some have not. Since quality phonics instruction is informed by assessments that have purposes and psychometric evidence, it is important to select and use only assessments that do. One available phonics assessment is the CORE Phonics Survey. As established in Chapter I, this assessment is a popular decoding assessment; therefore, it is especially important to find out if it has clearly established purposes or psychometric properties. This is the third purpose of this review of literature.

To investigate the purposes of this assessment, the *Assessing Reading: Multiple Measures* (Honing, Diamond, & Nathan, 2008) was reviewed. The purposes of the CORE Phonics Survey are outlined in the introduction preceding this test (p. 41). It

states that this test can be used for all four purposes of assessment, screening, outcome measurement, diagnosis, and progress monitoring. As a screening measure or an outcome measure, it can provide data about growth and mastery at the end of an instructional period. As a diagnostic tool, it can indicate whether or not a student needs instruction in selected phonics concepts, or if further assessment is needed. It may also be used to track progress from earlier skills to grade level mastery. It is stated that the CORE Phonics Survey is not meant to replace all other phonics assessments but is to be used to augment other tests (p. 41).

Since the purposes of this assessment are already defined, the remaining focus of this literature review is to inquire as to whether there are any existing psychometric properties for this assessment tool.

To begin this search, a phone interview was conducted on December 15, 2008, with Nancy Martin, test consultant for *CORE*, concerning any technical information available on the CORE Phonics Survey. She stated that this test was designed to follow the guidelines of current research yet be quick and easy to use in classrooms. However, at this point in time this assessment had no established diagnostics. Another phone interview was made that same day to the test publisher, *Arena Press*. The personnel there had no information concerning psychometric properties of this assessment either. This interviewing process with *CORE* and *Arena Press* was conducted twice. Once on December 15, 2008 at the onset of this research study and again on August 12, 2009 as it was coming to an end. This was to verify whether any additional information had come to light over those 10 months. It had not.

Since the authors and publishers were not aware of information concerning validity and reliability of the CORE Phonics Survey, a library search was conducted to find out if educational researchers had investigated the psychometric properties of this assessment and had published such information.

A comprehensive library search was also done concerning the CORE Phonics Survey which included looking in the following databases: *Academic Premier, CQ Researcher, Digital Dissertations, Educational Full text, ERIC, Professional Development Collection, Psychology and Behavioral Sciences Collection, PsychINFO, and Web of Science*. Nothing resulted from any of these searches. There was no evidence in any of these searches that any studies or information dealing with the psychometric properties of the CORE Phonics Survey existed. At this point in time it does not have psychometric evidence.

Validity measures would strengthen the CORE Phonic Survey. Experts on phonics could give opinion as to whether this survey includes all the concepts of the construct of phonics. Criterion-related validity could be established by comparing the concepts of the CORE Phonics Survey with those of already existing phonics criteria such as the Utah State Language Arts Core Curriculum or a scope and sequence in a national basal reader. Finally, the assumed two-factor construct of the CORE Phonics Survey could be investigated through statistical analysis. Any or all of these validity-testing procedures would add strength to this assessment tool.

Evidence of reliability could also be strengthened. The CORE Phonics Survey would be strengthened by determining the degree of variation between scores on different

testing sessions, between different raters, or between test items. Generalizability theory testing would also give insight to which of these factors added significant error variance.

Clearly this phonics assessment tool was lacking the scientific research base needed to support its established wide use. Therefore the purpose of this study was to establish the validity and reliability of the CORE Phonics Survey.

Summary

This review of literature accomplished three purposes. First, it established that phonics instruction is important for young readers as evidenced in the findings of the National Reading Panel 2000 meta-analysis on decoding. This document gave overwhelming evidence of the benefits of phonics instruction in early reading acquisition. However, phonics instruction, regardless of how beneficial, can only be as effective as the assessments that drive it. Therefore, assessments must be carefully selected for their (a) intended purpose and for their (b) psychometric properties. These qualities were both discussed. Finally, an investigation was conducted to investigate any existing literature on the validity and reliability of the CORE Phonics Survey. This investigation showed that no empirical evidence exists at this point in time.

CHAPTER III

METHODOLOGY

There is no empirical evidence for the valid or reliable use of scores obtained from administering the CORE Phonics Survey, which has become a widely used assessment. The purposes of this chapter are to describe the process by which scores obtained from the CORE Phonics Survey will be evaluated for validity and reliability.

Research Questions

Validity Research Questions

Classical Test Theory

1. What is the evidence of content validity of the CORE Phonics Survey as measured by convergence or agreement among expert reviewers?
2. What is the evidence of criterion validity for the CORE Phonics Survey as measured by the percent of overlap between items on it and the phonics items in the Utah State Core Curriculum?

Modern Test Theory

1. What is the evidence for construct validity on the CORE Phonics Survey using a confirmatory factor analysis to validate a hypothesized two-factor construct?

Reliability Research Questions

Classical Test Theory

1. What is the evidence of interrater reliability on the CORE phonics Survey as measured by a Pearson's r correlation coefficient?
2. What is the evidence of test-retest reliability on the CORE Phonics Survey as measured by a Pearson's r correlation coefficient?
3. What is the evidence of internal consistency reliability on each subtest and the total CORE Phonics Survey as measured by a Cronbach's alpha coefficient?

Modern Test Theory

1. To what degree do the raters and occasions contribute to variance among scores on the CORE Phonics Survey as measured by a G study?
2. What is optimal number of occasions and raters when administering the CORE Phonics Survey to minimize error variance and optimize the reliability of the resulting rating as measured by a D study?

Instrument

The CORE Phonics Survey (Honing et al., 2008) was a nonstandardized series of subtests addressing phonics related skills for early readers. This assessment is one in a compilation of reading assessments for early readers found in the *Handbook of Assessing Reading: Multiple Measure for Kindergarten through Eighth Grade* (Honing et al.). The survey assesses phonics skills that have a high rate of application in beginning reading.

There are 12 subtests, which are grouped into two major categories or factors. The first factor is alphabet skills (subtests A-D) and the second factor is entitled Reading and Decoding Skills (subtests E-L). On the subtests E-L, pseudo words are included in the list of words so that children must rely on their knowledge of letters and sounds, rather than on memory. One third of the words in each of these sections are pseudo words.

Alphabet skills (factor 1): This category includes four subtests that include (a) naming the uppercase letters, (b) naming the lowercase letters, (c) saying the consonant sounds, and (d) saying the long vowel sounds and short vowel sounds.

Reading and decoding skills (factor 2): This category includes eight subtests. In each category, students are to read both real words and pseudo words. The categories include (a) short vowels in CVC words; (b) consonant blends with short vowels; (c) short vowel, digraphs, and –tch trigraph; (d) r-controlled vowels; (e) long vowels spellings; (f) variant vowels; (g) low frequency vowel and consonant spellings; and (h) multisyllabic words.

The *CORE Phonics Survey* is an inexpensive test. The cost of a test manual is about \$40 and the pages are reproducible for use throughout the school. It is also quick to administer and score. It typically takes about 10 minutes to administer it to one student. This assessment can be used for screening, progress monitoring and diagnostic purposes. The results from the data are helpful in planning instruction and instructional groupings in the primary grades.

Design

Determining the Validity of the CORE Phonics Survey

Classical Test Theory

Content validity. Content validity is determined by expert judgment. There is no formula for computing it and there is no standard way of expressing it quantitatively. Experts in the area of phonics were asked to make a judgment concerning how well the items on the test represent the construct of phonics. This judgment was based on whether all phonics concepts are represented and whether the format is appropriate for beginning readers. Selected evaluators assumed the role of “expert” and evaluated the CORE Phonics Survey for content validity. The process by which content validity was evaluated in the present study will be explained shortly.

Criterion validity. Criterion validity is the degree to which the scores on a test are related to the scores on another test or to some other valid criterion available at the same time. The design of this validity testing was to determine the percent of overlap between the CORE Phonics Survey, the Utah State Language Arts Core, and *Scott Foresman* (2008) basal reader scope and sequence on decoding instruction. The process by which criterion validity was evaluated in the present study will be explained shortly.

Modern Test Theory

A confirmatory factor analysis (CFA) is a statistical method used to define unobserved variables, called latent variables, which can account for the covariance among items of observed variables. It is a special form of factor analysis that is

hypothesis driven. A CFA was done to confirm or disconfirm the hypothesized two-factor construct of phonics used by the CORE Phonics Survey, which assumes that phonics knowledge has two main factors: (a) alphabetic knowledge and (b) reading and decoding skills. Since these are unobservable variables, they need to be defined by observable tasks. These observable tasks are the test items that are listed under each heading on the test. The first factor, *alphabet skills and letter sounds*, is defined by the ability to identify letter names, consonant sounds, long vowel sounds and short vowel sounds. The second factor, *reading and decoding skills*, is defined by the ability to decode CVC words, consonant blends with short vowels, short vowels, diagraphs and -tch trigraph, r-controlled vowel, long vowel spellings, variant vowels, low frequency vowel and consonant spellings, and multisyllabic words. A CFA shed light as to how and if the decoding skills for each category correlate together and cluster around these two general factors, thus indicating that those skills/items are indeed pointing to the same construct.

A CFA was analyzed using the items from 500 student tests to show whether they clustered onto their respective factors with a high enough correlation to confirm that the hypothesized two-factor structure of the CORE Phonics Survey is correct.

As part of the analysis, goodness-of-fit tests were run to see how well the hypothesized model matched the observed data. Results showing a failure to reject the null hypothesis would be an indication of a good model fit.

Determining the Reliability of the CORE Phonics Survey

To investigate or determine the reliability of the CORE Phonics Survey, Classical Testing Theory (CTT) and Modern Test Theory (MTT) were used.

Classical Test Theory

Test-retest reliability. Test-retest is part of CTT in which researchers establish the degree to which test scores are consistent over time. It indicates the variation in test scores obtained from the same individuals that occurs from one testing session to another as a result of errors of measurement (Gay, 1985). In other words it shows evidence that the score a person obtains on a test at some moment in time is the same score, or close to the same score, that the person would get if the test were administered some other time. For this study, a Pearson's r was calculated to determine the degree of correlation and variance in CORE Phonics Survey scores that were given on two different occasions to the same group of students. Two classes of students (44 students total) were tested twice, two weeks apart to identify possible sources of variance in scores due to the testing occasion.

Interrater reliability. CTT was used to establish the correlation between scores on the CORE Phonics Survey given by different raters. Interrater reliability refers to the reliability of two (or more) independent scorers (Gay, 1985). A Pearson's r was calculated to determine the degree of correlation and variance in scores due to the difference in the raters. Twenty-five students (of the 44 students tested for test-retest reliability) were also scored by two raters during the two testing periods to identify

possible variance in scores due to the rater.

Internal consistency reliability. Internal consistency is a measure of item inter-correlation. This correlation is represented in a coefficient of reliability called the Cronbach's alpha. Five hundred tests were analyzed to determine the item inter-correlation for each subheading (subtests A-L) on the CORE Phonic Survey. A correlation of at least (.70) indicates there is evidence that the items are measuring the same underlying construct. Each part of the CORE Phonics Survey (A-L) was analyzed separately as its own testlet.

Modern Test Theory

Generalizability testing. Generalizability studies are conducted in two stages. The first stage is called a *G* study; the second is called a *D* study. The *G* study is a statistical test that not only establishes the general degree of correlation found in classical testing, but also aims to identify the sources of variance in the scores. *G* theory assumes that each student's observed score is comprised of a universe score (the student's average score over all items of measurement), with one or more sources of error. Therefore, the power of *G* theory is that it allows one to evaluate the extent to which generalizations might be made from the student's observed score to the universe of observations that are confined to the factors measured in the *G* study (Rathvon, 2004). Because the possible sources of variance are defined from the beginning, it is possible to determine which source(s) of variance could be reduced by changing aspects of the process or instrument. Reduced or low variance makes a more reliable assessment. A smaller error variance provides for a higher power of hypothesis test and narrower confidence interval.

The second stage of generalizability study is the *D* study. This analysis uses the information from the *G* study to predict the effect of decreasing or increasing the number of levels of each facet on reliability. It can also show the effects of using an alternative design.

Through *G* theory, traditional distinctions of reliability and validity are overcome. A universe (true score), its facets (sources of possible variance), and conditions for admissible observations are defined through careful explication of the construct, which is the traditional domain of validity theory (Grimm & Yarnold, 2000). The use of the terms “dependability” and “generalizability” instead of “reliability” are used to show a degree of unifying principles of reliability and validity.

For this study, a *G* test was done to determine the sources of variance, called facets, which influence the generalizability of the test. A fully crossed, two-factor design was used: 25 students x 2 occasions x 2 raters. This means that 25 students were tested on two occasions by two raters. The students were the objects of measurement and the occasions and raters represented possible sources of error variance in the scores. These factors were selected because the researcher felt that they could be ones that contribute to the difficulty of establishing instrument reliability (Rathvon, 2004). Then a *D* Study was done to explore the degree of reliability when the number of occasions and raters were changed.

Procedures

Determining the Validity of the CORE Phonics Survey

Classical Test Theory

Data collection. Three experts in the field of reading were contacted to give an evaluation of the appropriateness and completeness of the items of the CORE Phonics Survey. These were John A. Smith, Lloyd Eldredge, and Barbara Fox. All of these experts have published national textbooks on the topic of phonics. Two of these experts responded to the request. The responses of both were collected and published in this report to provide evidence of the content validity of this assessment.

The Utah State Language Arts Core Curriculum and the Scott Foresman (2008) phonics scope and sequence were collected to compare the concepts included in each and compare the overlap.

Content validity testing. The reviews from two decoding experts were obtained to verify that the items in the CORE phonic Survey tested important, generally-accepted phonics content. Both experts have published national textbooks on the topic of phonics.

Criterion validity testing. The CORE Phonics Survey was compared to the Utah State Language Arts Core and Scott Foresman (2008) basal reading series. All the concepts from CORE Phonics Survey were listed beside corresponding concepts of the Utah State Language Arts Core, then again beside the phonics concepts in the scope and sequence of the Scott Foresman basal reading series. A percentage was calculated to determine the degree of overlap between each respective identifier of the construct of

phonics and the CORE Phonics Survey.

Modern Test Theory

Data collection. Tests were collected from Schools A, School B, and School C which had been completed at the beginning of the school year 2008. These tests had been administered by school reading specialists and other trained school personnel. Tests gathered from School D were administered to all first and second graders in January 2009 by 18 different raters. Fifteen of the raters were students from a local university who volunteered to help. They were trained on how to administer the CORE Phonics Survey. The training consisted of 1 hour of instruction that included: (a) instruction on test procedures and appropriate dialogue, (b) review on the word pronunciation from the lists, (c) practice giving the test in partners, and (d) viewing a video clip of the test being administered to a young student. While watching the video, each section of the test was paused to allow for discussion and questions. These university student volunteers, along with the reading specialist and two paraprofessionals of School D, tested four first grade classrooms, and four second grade classrooms of that school. There were 351 tests collected.

The students who completed the tests to be used for this study came from four different elementary schools in the Rocky Mountain Region. Two of the schools were in metropolitan, low SES areas. School A was a K-6 school and had a population of 730 students with 57% Caucasian, 35% Hispanic, 5% Pacific Islanders, 1% Black and 2% other ethnic races. The SES status of the school was low, with 73% of students received free or reduced lunch, qualifying as a Title I school. School B was a K-6 school and had

a population of 700 students with 78% Caucasian, 18% Hispanic, 2% Pacific Islander, and 2% other racial backgrounds. Forty-four percent of the students in School B received free/reduced lunch. The tests from these students came from grades 2-5 and were previously administered in the fall of 2008 as part of their school-wide reading program. Permission was obtained to copy these tests for this study without student identification.

The participants from the third and fourth schools were from elementary schools in a mid-size city (less than 250,000), middle-class neighborhoods. School C was a K-5 school and had a population of 63% Caucasian, 31% Hispanic, 4% Asian, 1% Black, and 1% Native American. The school population was below average SES with approximately 60% of the students in this school receiving free or reduced lunch therefore qualifying as a Title I school. The student tests from this school were also completed at the beginning of the 2008-2009 school year. School D was a K-2 school of about 600 students. The school population was 49% Caucasian, 39% Hispanic, 6% Pacific Islander, 3% Black, and 3% Native American. The SES in this school was average, with approximately 30% of students qualifying for free/reduced lunches, and it did not qualify as a Title I school. This school had 351 first and second graders who participated in this study by taking the CORE Phonics Survey. Student identity was not made known to the researcher and all identifying information was deleted from the test score sheets.

Construct validity testing. A CFA was done to confirm or disconfirm the hypothesized two-factor construct of phonics used in the CORE Phonics Survey. Data from 500 tests were used to run a G test, a type of analysis of variance (ANOVA).

Determining the Reliability for the CORE Phonics Survey

Classical Test Theory

Data collection. Forty-four students (two classrooms) from School D were administered the CORE Phonics Survey. Students were tested during computer time and independent work time. Two weeks later these same students were given the test again at approximately the same time of the day and under the same circumstances. Test scores were analyzed to determine the variance between scores given on different occasions.

Twenty-five students (of the 44 students tested above from School D) were given the CORE Phonics Survey with two raters scoring it. Test data were analyzed to determine variance between scores given by different raters.

For the internal reliability testing, 500 tests total were collected. One hundred forty-nine tests were collected from School A, School B, and School C which had been administered at the beginning of the 2008-2009 school year. Tests were completed by students ranging from second to fifth grades. Student names were eliminated before tests were collected for this analysis. Raters were trained within their own respective schools and consisted of about one half hour instruction.

Three hundred fifty-one tests were administered and gathered from School D. Raters were university student volunteers which had previously had about 1 hour of training. Training consisted of explanation, hands-on practice, and video clips accompanied by discussion.

Test-retest reliability testing. Forty-four students were tested on two occasions, 2 weeks apart. A Pearson's r was calculated to see if the scores would remain relatively

constant over the different testing occasions.

Interrater reliability testing. Twenty-five students were scored by two raters as they took the CORE Phonics Survey. A correlation coefficient (Cronbach's alpha) was calculated to see if the scores remained the same across both raters. If scores were highly consistent, this indicated a high degree of reliability.

Internal consistency reliability testing. Each section of the CORE Phonics Survey (parts A-L) was statistically analyzed to determine the degree of correlation between all the items in each separate part of the test. The correlation coefficient used is called a Cronbach's alpha. Cronbach's alpha is a function of the number of test items and the average inter-correlation among them. Thus, if the number of items is high and/or inter-item correlations are high, the alpha increases. A high Cronbach's alpha indicates that the items are measuring the same underlying construct, indicating high reliability. If the datum is multidimensional, or made of several unrelated parts, the Cronbach's alpha will generally be low for the items, indicating low reliability.

Modern Test Theory

Data collection. Tests were collected from Schools A, School B, and School C which had been completed at the beginning of the school year 2008. These tests had been administered by school reading specialists and other trained school personnel. Tests gathered from School D were administered to all first and second graders in January 2009 by 18 different raters. Fifteen of the raters were students from a local university who volunteered to help. They were trained on how to administer the CORE Phonics Survey. The training consisted of 1 hour of instruction that included: (a) instruction on test

procedures and appropriate dialogue, (b) review on the word pronunciation from the lists, (c) practice giving the test in partners, and (d) viewing a video clip of the test being administered to a young student. While watching the video, each section of the test was paused to allow for discussion and questions. These university student volunteers, along with the reading specialist and two paraprofessionals of School D, tested four first grade classrooms, and four second grade classrooms of that school. There were 351 tested collected.

The students who completed the tests to be used for this study came from four different elementary schools in the Rocky Mountain Region. Two of the schools were in metropolitan, low SES areas. School A was a K-6 school and had a population of 730 students with 57% Caucasian, 35% Hispanic, 5% Pacific Islanders, 1% Black and 2% other ethnic races. The SES status of the school was low, with 73% of students received free or reduced lunch, qualifying as a Title I school. School B was a K-6 school and had a population of 700 students with 78% Caucasian, 18% Hispanic, 2% Pacific Islander, and 2% other racial backgrounds. Forty-four percent of the students in School B received free/reduced lunch. The tests from these students came from grades 2-5 and were previously administered in the fall of 2008 as part of their school-wide reading program. Permission was obtained to copy these tests for this study without student identification.

The participants from the third and fourth schools were from elementary schools in a mid-size city (less than 250,000), middle-class neighborhoods. School C was a K-5 school and had a population of 63% Caucasian, 31% Hispanic, 4% Asian, 1% Black, and 1% Native American. The school population was below average SES with approximately

60% of the students in this school receiving free or reduced lunch therefore qualifying as a Title I school. The student tests from this school were also completed at the beginning of the 2008-2009 school year. School D was a K-2 school of about 600 students. The school population was 49% Caucasian, 39% Hispanic, 6% Pacific Islander, 3% Black, and 3% Native American. The SES in this school was average, with approximately 30% of students qualifying for free/reduced lunches, and it did not qualify as a Title I school. This school had 351 first and second graders who participated in this study by taking the CORE Phonics Survey. Student identity was not made known to the researcher and all identifying information was deleted from the test score sheets.

Generalizability testing. A two-facet fully crossed design was used in a generalizability study 25 students x 2 occasions x 2 raters. This means that 25 students were tested on two occasions with two raters. The students were the object of measurement. A *G* test was performed using GENOVA and determined the sources and amount of error variance that was contributed by the various facets: students, raters, occasions, students by occasions, students by raters, and occasions by raters, and students by occasions by raters.

Data Analysis

Determining the Validity of the CORE Phonics Survey

Classical Test Theory

To analyze the data for content validity, the responses of two experts on the topic of phonics were compared and reported. Furthermore to analyze the data for criterion

validity, a percentage was figured showing the degree of overlap between the items of the CORE Phonics Survey and the Utah State Language Arts Core; then again between the CORE Phonic Survey with the Scott Foresman Basal Reader.

Modern Test Theory

To investigate construct validity, the CFA was run using the statistical program *M-Plus*. After the analysis was complete a chi-square test-of-fitness was also performed to determine if the hypothesized two-factor model of the CORE Phonics Survey held up to the statistical data. Acceptable criteria for the factor structure model was $p \geq .05$.

Determining the Reliability of the CORE Phonics Survey

Classical Test Theory

To investigate test-retest and interrater reliability, a Pearson's r was run. To investigate internal consistency reliability, a Cronbach's alpha was used.

Modern Test Theory

For the generalizability study, a three-way (student by rater by occasion) random effects ANOVA was used to compute estimates of the seven variance components. The components include student, rater, occasion, student by occasion, student by rater, rater by occasion, and residual interactions. The computer program of SPSS version 17 was used to run a GENOVA analysis to perform the G study and a D study.

Summary

The purpose of this chapter was to describe the procedures for determining the degree of validity and reliability of this highly used reading assessment. Several different types of validity and reliability testing were used to thoroughly accomplish this. The types of validity that were addressed were content, criterion, and construct. The types of reliability that were addressed in the study were test-retest, interrater, internal consistency, and generalizability.

CHAPTER IV

RESULTS

The purpose of this study was to establish the validity and reliability of the CORE Phonics Survey. Although this is a widely used assessment tool, the review of literature indicates that there have been no psychometric studies done on this to date. There were several types of validity and reliability testing done to determine the soundness of this test. This chapter contains a report of the results and findings relative to the eight questions stated in Chapter I.

To explore these questions, several different types of statistical analyses were conducted. For classical test theory Pearson's r and Cronbach's alpha were done. In addition to these analyses, a response from a questionnaire from an expert in reading and decoding was included and a comparison between the CORE Phonics Survey, the Utah State Core Curriculum, and the Scott Foresman basal reading series. For modern test theory, generalizability tests (G and D studies) and confirmatory factor analysis (CFA) was done. Below is shown the outcomes of each of these results. Each will be explained in the order of the outlined research questions.

Findings

Determining the Validity of the CORE Phonics Survey

Classical Test Theory

Question 1. What is the evidence for consensus or content validity of the CORE

Phonics Survey as measured by convergence or agreement among expert reviewers?

John A. Smith is an expert in the field of reading instruction and responded to the request to evaluate the CORE Phonics Survey. He is currently a professor and department head at University of Texas Arlington. He has written a textbook entitled *Early Literacy Instruction; A comprehensive Framework for Teaching Reading and Writing, K-3*, which includes extensive instruction on teaching phonics skills to young readers. In his response concerning the accuracy and completeness of the CORE Phonics Survey, John A. Smith gave the following critique of this test.

Overall, I like this test very much. I think it does a fine job of assessing student knowledge of the common and useful phonics spelling patterns.

They [authors of the CORE Phonics Survey] didn't test for common chunks (-ing, -ight, -all). They may consider that part of sight vocabulary, rather than phonics.

They [authors of the CORE Phonics Survey] didn't test for students' knowledge of contractions, but they probably consider that to be structural analysis, not phonics. I agree.

John A. Smith
04/09

J. Lloyd Eldredge is the second expert in the field of phonics instruction that participated in this content validity investigation. He was a professor in the David O. McKay School of Education at Brigham Young University where he taught both graduate and undergraduate literacy courses. He has written several books on the topic of phonic, including: *Teaching Decoding: Why and How (2005)*, *Phonics for Teachers: Self Instruction Methods Activities (2003)*, *Teaching Decoding in Holistic Classrooms (1995)*, and *Decoding Strategies (1993)*. In his response concerning the accuracy and completeness of the CORE Phonics Survey, J. Lloyd Eldredge gave the following

critique of this test.

Phonics is the association of graphemes and phonemes. Implicit phonics programs teach students how to correctly associate the letters of written words with the sounds those letters represent. Explicit phonics programs teach students this association knowledge, but also teach them to isolate the phonemes and blend them so they can identify unfamiliar words by “sounding them out.” The CORE Phonics Survey has content validity for phonics as it is defined by both implicit and explicit phonics advocates.

Phonics can only be used with syllables or single syllable words since the syllable pattern (closed syllable, open syllable, and VCe syllable) influences the vowel sound in the word or syllable. Therefore, if students are to be able to use phonic to “sound out” words they must be able to identify syllable boundaries in multi-syllabic words. The CORE Phonics Survey has content validity for phonics advocates who recognize the limitations of phonics teaching on students’ decoding abilities when they are not taught to “chunk” multi-syllabic words into appropriate syllables.

Among the most important phonics elements to assess we find the following:

- 1) single consonants found in both the initial and final position of written words,
- 2) consonant blends found in both the initial and final position of written words,
- 3) consonant digraphs found in both the initial and final position of written words,
- 4) single vowels found in various positions of written words,
- 5) vowel digraphs found in various positions of written word,
- 6) vowel diphthongs found in various positions of written words,
- 7) murmur diphthongs (sometimes referred to as r-controlled vowels) found in various positions of written words, and
- 8) silent letters found in various letter combinations.

While the CORE Phonics Survey is not an extensive diagnostic assessment, it does represent the most important phonics elements involved in phonics.

Some frequently used graphemes (letters and letter combinations representing phonemes) represent two phonemes (those that represent more than two phonemes are generally not taught by teachers because it is unproductive). Some of the vowel graphemes representing two sounds are included in the assessment (real words: toe, leap, tie, blow, few, down, moon, cook, and sweat; pseudo words: loe, beap, voot and rew). I would suggest that provisions are made for the student to respond with both sounds or that different words are used in the assessment.

Finally, all pseudo words used in the test should, in my opinion, represent the structure of real words. A pseudo word such as “nik” does not represent the structure of real words. The CORE Phonics Survey has done a very good job with pseudo words. The only questionable pseudo words would be: loe and rew.

In sum, I would consider the CORE Phonics Survey assessment to have content validity.

J. Lloyd Eldredge
09/10/09

Question 2. What is the evidence of criterion validity for the CORE Phonics Survey as measured by the percent of overlap between items on it and the phonics items in the Utah State Language Arts Core Curriculum?

Standard 4 of the Utah State Language Arts Core Curriculum on phonics instruction and the scope and sequence of the Scott Foresman (2008) Core Reading program were examined to determine if the content and skills required by the Utah State Office of Education and those included in a national reading program matched those in the CORE Phonics Survey. Because criterion-related validity can be established by comparing an instrument or test to other current manifestations of the same construct, these were viable sources of comparison for establishing the criterion validity of the CORE Phonics Survey.

The Utah State Language Arts Core Curriculum on phonics is the required curriculum for beginning readers in all of the Utah public schools. Phonics instruction begins in kindergarten and continues through second grade. It is expected that students should have mastered these concepts by the end of second grade; after second grade the skills are taught as needed, or reviewed to maintain them, or to remediate struggling readers.

The Scott Foresman Core Reading program is one of the five top programs used nationally for reading instruction. Decoding skills are taught in kindergarten through second grade. In third grade, the skills are only taught as remediation skills.

Table 1 shows the comparison between the skills of the Utah State Core Curriculum and the CORE Phonics Survey and Table 2 shows the comparison with the Scott Foresman Core Reading program.

There is a 90% overlap in the skills between CORE Phonics Survey and the Utah State Language Arts Core Curriculum. The following is the breakdown of the items of difference.

Kindergarten: All of the skills between the two decoding criteria are consistent with one another with one exception. In the Utah State Core Curriculum (USCC), kindergarten *Objective b* requires the students to be able to match the short vowel and consonant sounds to the letters. Although the CORE Phonics Survey (CPS) has the students pronounce the short vowels and consonants, it does not require the students to hear the sound first and then match the sound to a letter.

First grade: All of the skills in these two decoding criteria are consistent with each other with only the following exceptions. The USCC first grade *Objective 1a* requires the student to write the alphabet letter for the given sound. CPS requires the student to identify the sounds of each letter only without writing it. Additionally, in *objective 1c* there are no CV words included in the Core Phonics Survey except as applied in multisyllable words. Lastly, there are no suffixes, or word families in the CORE Phonics Survey as stated in first grade USCC *objectives 1e, 1i, 1j*.

Second grade: All of the skills in the USCC and CPS match with the exception of the following skills. In USCC *objective 2b*, students are required to accurately use vowel digraphs in two syllable words. All of the stated vowels digraphs are used in one-

Table 1

Criterion Validity: Comparison of Utah State Core Curriculum and CORE Phonic Survey

Utah State Core Curriculum on decoding	CORE Phonics survey
Kindergarten skills	
a. Name all upper- and lower-case letters of the alphabet in random order	Part A-Part B
b. Match consonant and short vowel sounds to the correct letter	None
c. Blend simple CVC sounds into one-syllable words	Part E
First-grade skills	
a. Write letters to represent spoken sounds of all letters in random order	None
b. Identify and pronounce sounds for consonants, consonant blends, and consonant digraphs in words	Part C, E, F, G
c. Identify and pronounce sounds for short and long vowels, using patterns (cv, vcv, cvc, cvvc, cvcv, cvc-silent e) AND vowel digraphs (ea, ee, ie, oa, ai, oy, oo ow) in words	Part E, I, J, K
d. Identify and pronounce sounds for r-controlled vowels accurately in one-syllable words (ar, or, er)	Part H
e. Identify and blend initial letters sounds with common vowel patterns to pronounce one-syllable words	Parks E-K
f. Identify and read grade-level contractions and compound words	None
g. Identify sound patterns and apply knowledge to decode one-syllable words (blends, digraphs, vowel patterns, r-controlled vowels)	Park F-K
h. Demonstrate an understanding of representing same sounds with different patterns in one-syllable words (ee, ie, ea, e)	Part I, K
i. Use knowledge of root words and suffixes to decode words (-ful, -ly, -er)	None
j. Use letter patterns to decode words (phonograms/word families/onset and rime: -ack, -ail, -ake)	None
Second-grade skills	
a. Identify and pronounce all vowel diphthongs (oi, oy, aw, au) AND consonant digraphs (ch, sh, th, wh) accurately in words	Part J, G
b. Identify and pronounce sounds for short and long vowels, using pattern (cvc, cvvc, cvcv, cvc silent e) AND vowel digraphs (ea, ee, ie, oa, ai, ay, oo, ow) accurately in two syllable words	Part E, I, L

(table continues)

Utah State Core Curriculum on decoding	CORE Phonics survey
c. Identify and pronounce r-controlled vowel pattern in words (ar, or, er)	Part H
d. Identify and blend letter sounds to pronounce words	Part E-K
e. Identify and read grade-level contraction and compound words	None
f. Identify sound patterns and apply knowledge to decode words (blends, digraphs, r-controlled)	Parts F, H, I, J, K
g. Demonstrate understanding of representing same sound with different patterns by decoding these patterns accurately in isolate and in text (ee, ea, ei, e)	Part I, K
h. Use knowledge of root words and prefixes (re, un, mis) AND suffixes (s, es, ed, ing, est, ly)	Part L
i. Use letter and syllable patterns to pronounce multisyllabic words	Part L

syllable words throughout the CPS, but only ai, ay, oo are included in Section L of multisyllabic words. Again, contractions and compound words are not included in CPS.

Additionally, the CORE Phonics Survey includes more skills of phonics not mentioned in the Utah State Core Curriculum on Phonics. Variant vowels combinations (part J) such as ew, ow, ue, ou, oo, ew, ou, aw are also included as important skills. Low-frequency vowel and consonant spelling (part K) include the additional phonics concepts of silent letters (kn, gh, wr, gn, b, wr,), soft c/g sounds, and y as a vowel.

There is a 93% overlap in the concepts between the CORE Phonics Survey and the Scott Foresman Core Reading Program if only phonics concepts are taken into account. All of the concepts between the two decoding criteria are consistent with one another except for two. The CV vowel pattern is not represented by itself, only in multisyllable words. The concept of syllable c + le is not represented in the CORE Phonics Survey at all.

Table 2

Criterion Validity: Comparison of Core Reader with CORE Phonics Survey

Scott Foresman core reader (2007)	CORE phonics survey
Name all upper and lowercase letters	Part A, B
Letter sounds	Part C, D
Blend CVC words	Part E
Initial and final consonant blends	Part F
Digraphs- th, sh, wh, ch, tch	Part G
CVCe	Part I
Hard/soft g and j	Part K
Sounds of y	Part K
CV	none
Silent letters Ng, kn, wr	Part F, K
R-controlled vowels	Part H
Long vowel spellings: ai, ay, oa, ow, ee, ie, igh	Part I
Diphthongs: Ew, eu, ui, ow, ou, oi, oy, au, aw	Part J
Syllable c+le	none
Open syllables VCV	Part L
Short ea	Part K
Vowels oo	Part J
Silent consonants – kn, wr, gn, mb	Part K
/F/ Ph, gh	Part G
Compound words	Part L
Inflected endings e before ed, ing	Part L
Suffixes -s, -ing, -es, -ed, -er, -est, ness, less, ly	none
Contractions	none

If structural analysis is considered as part of phonics instruction there is an 87% overlap in skills. The CORE Phonics Survey does not account for suffixes and contractions.

Modern Testing Theory

Question 1. What is the evidence for construct validity on the CORE Phonics Survey using a confirmatory factor analysis to validate a hypothesized two-factor construct?

A confirmatory factor analysis (CFA) is a statistical procedure that confirms or disconfirms the hypothesized structure of an assessment. The CORE Phonics Survey has defined phonics as two latent variables or factors: (a) alphabet skills and letter sounds and (b) reading and decoding skills. Factor 1, alphabetic skills and letter sounds, includes the observable traits of (a) letter names uppercase, (b) letter names lowercase (c) consonant sounds and (d) vowel sounds. Factor 2, called reading and decoding skills, includes the observable traits of (e) CVC words (f) blends (g) diagraphs, tri- (h) r-controlled vowels, (i) long spellings, (j) variant vowels, (k) low frequency spellings, and (l) multisyllable words. A CFA analysis showed whether this construct does indeed fit this model. Figure 1 shows how each trait clustered together confirming a two-construct structure of phonics. Table 3 shows the output of descriptive analysis of the CFA; Table 4 shows degree of correlation between different sections of the CORE Phonics Survey.

.67

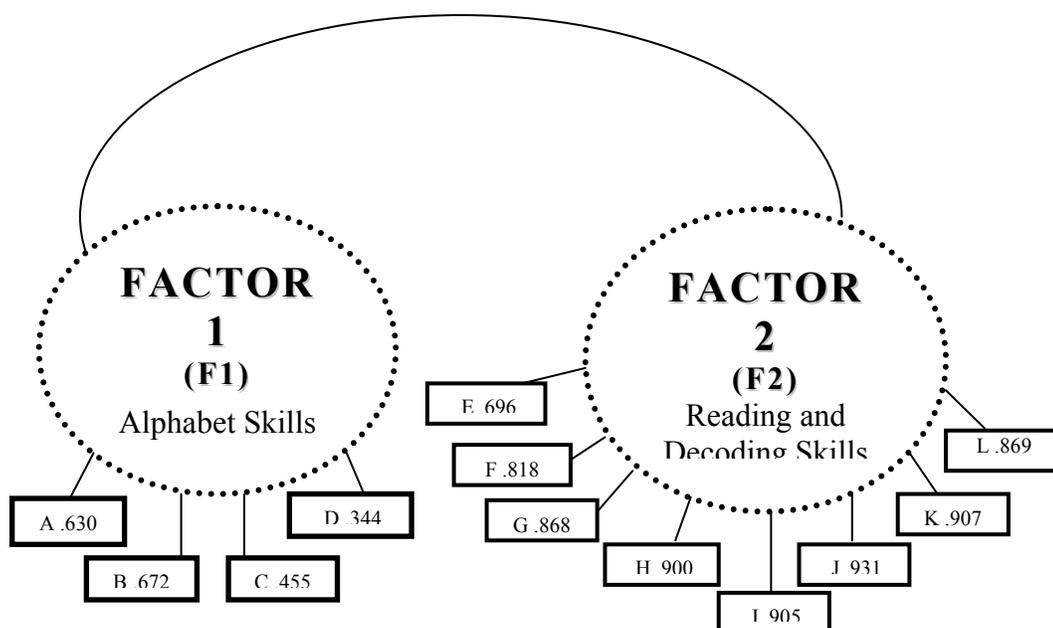


Figure 1. Confirmatory factor analysis construct validity.

Table 3

Descriptive Analysis of Confirmatory Factor Analysis

Survey section	Min	1 st quartile	Median	Mean	3 rd quartile	Max
Part A	21.00	26.00	26.00	25.77	26.00	26.00
Part B	17.00	25.00	26.00	25.43	26.00	26.00
Part C	0.00	21.00	22.00	21.95	23.00	23.00
Part D	1.00	9.00	10.00	9.28	10.00	10.00
Part E	1.00	13.75	15.00	13.79	15.00	15.00
Part F	0.00	11.00	13.00	12.05	14.00	15.00
Part G	0.00	8.00	13.00	11.98	15.00	15.00
Part H	0.00	8.00	13.00	10.91	15.00	15.00
Part I	0.00	9.00	13.00	11.26	15.00	15.00
Part J	0.00	7.00	12.00	10.23	14.00	15.00
Park K	0.00	2.00	9.00	8.01	13.00	15.00
Part L	0.00	2.00	14.00	11.61	20.00	24.00

Table 4

Correlations Matrix for Confirmatory Factor Analysis

Variable	A	B	C	D					
Observed variable F1									
A. Letters-upper	1.00	--	--	--					
B. Letters-lower	.48	1.00	--	--					
C. Consonants	.21	.31	1.00	--					
D. Vowels	.17	.18	.26	1.00					
Observed variable F2	E	F	G	H	I	J	K	L	
E. CVE	1.00	--	--	--	--	--	--	--	
F. Blends	.74	1.00	--	--	--	--	--	--	
G. Digraphs, tri-	.70	.79	1.00	--	--	--	--	--	
H. r-controlled	.62	.74	.80	1.00	--	--	--	--	
I. Long vowel	.63	.73	.79	.84	1.00	--	--	--	
J. Variant vowels	.62	.73	.79	.83	.87	1.00	--	--	
K. Low frequency	.56	.70	.75	.81	.80	.86	1.00	--	
L. Multisyllable	.55	.69	.72	.76	.75	.82	.89	1.00	

As a preliminary step to conducting the CFA, a set of preliminary exploratory factor analyses (EFA) were conducted to determine whether the data would be better fit using either a one-factor or a three-factor model. The EFA did not converge for either model, however (results not reported). Only when specifying a two-factor model, did the model estimation converge to a solution. The results of the EFA served to further support the result of the CFA analysis.

Tests of Model Fit

There are several fit indices that provide information as to whether a model is a good fit and matches the observed data. Bentler (1990) and Thompson (1998) noted the problem with interpreting just one fit statistic and advise researchers to consult multiple fit statistics in order to consider different aspects of fit. This study consulted the chi-square statistic, the Bentler Comparative Fit Index (CFI), Tucker-Lewis Index (TLI), the root mean square residual (RMSEA), and the standardized root mean square residual (SRMR).

Chi-squared model (χ^2) is a classic goodness-of-fit measure to determine model fit. The null hypothesis is that the implied or predicted covariance matrix is equivalent to the observed sample covariance matrix. A large chi-square and rejection of the null means the model estimates do not sufficiently produce sample covariance; the model does not fit the data very well. By contrast, a small chi-square and failure to reject the null hypothesis is a sign of good model fit (Hu & Bentler, 1999). The chi-squared statistic (χ^2) is equal to 324.025 with 53 degrees of freedom and a p value of (0.00). This χ^2 is significant so the specified CFA model is not supported by the data and is not a

good fit. However, the χ^2 test is widely recognized to be problematic (Jöreskog, 1969). One reason for this is that it is sensitive to sample size, and it becomes more and more difficult to retain the null hypothesis as the number of cases increases. Therefore, other commonly reported tests of fit were also performed.

Other tests of fit were calculated (see Table 5). The *root mean square error of approximation* (RMSEA) likewise was consulted as a determinate of model fit. The criterion for a good model fit to the data for RMSEA are values less than .08. The RMSEA was calculated as .097, a bit higher than ideally acceptable.

The *standardized root mean square residual* (SRMR) is the standardized difference between the observed covariance and the predicted covariance. A value of zero indicates a perfect fit. This measure tends to be smaller as sample size increases and as the number of parameters in the model increases. A value less than .05 is considered a

Table 5

Goodness-of-Fit Tests

Test	Value
Chi-square test of model fit	
Value	324.025
Degrees of freedom	53
<i>p</i> value	0.0000
CFI/TLI	
CFI	.910
TLI	.887
RMSEA	
Estimate	.097
90% C.I.	.091 - .112
Probability RMSEA \leq .05	.000
SRMR	.042

good fit. The SRMR in this model was .042 indicating a good fit.

The confirmatory fit index (CFI) outcome indicates a good fit to the data. As the CFI approaches 1.0 the better the fit of the model to the data. The criterion for a good model fit to the data for CFT are values that exceed .90 (Stevens, 1996). The CFI in this model was .91, which falls into the acceptable range of model fit. Finally, the TLI was .887, also a bit lower than the lower range of acceptable model fit (TLI > .95).

Although not all of the tests of goodness of fit hit the exact cut-off values, Hu and Bentler (1999) provided a rule of thumb for deciding which statistics to report and choosing cut-off values for declaring significance. When RMSEA values are close to .08 or below and CFI and TLI are close to .90 or greater, the model may have a reasonably good fit. Therefore, although goodness-of-fit was not established with the χ^2 , the RMSEA, CFI and TLI together support an adequate goodness-of-fit for the predicted model and the data results.

Standardized Model Results

Standardized factor loadings are presented in Table 6. Variant vowels (part J) have the highest standardized factor loading of .931 and thus appears to be the most reliable indicator of reading and decoding skills (F2). By contrast, vowel sounds has the lowest standardized factor loading, .344. This suggests that it is not as strongly reliable an indicator of alphabet skills and letter sounds (F1). All of the factor loadings of F1 and F2 are significant ($p < 0.00$). The correlation of factor 1 with factor 2 is .68, a moderate correlation (Cohen, 1988).

Table 6

Standardized Model Results

Variable	Estimate	S.E.	Est./S.E.	Two-tailed <i>p</i> value
F1 Alphabet skill and letter sounds by...				
Letters uppercase	0.630	0.077	8.146	0.000
Letters lowercase	0.672	0.077	8.768	0.000
Consonant sounds	0.455	0.049	9.250	0.000
Vowel sounds	0.344	0.063	5.497	0.000
F2 reading and decoding skills by...				
Variant vowels (J)	0.931	0.008	114.058	0.000
CVC (E)	0.696	0.026	27.226	0.000
Blends (F)	0.818	0.021	39.795	0.000
Diagraphs, tri- (G)	0.868	0.014	63.013	0.000
R-controlled (H)	0.900	0.012	77.804	0.000
Long vowel spell (I)	0.905	0.011	81.433	0.000
Low frequent (K)	0.907	0.010	94.654	0.000
Multisyllable (L)	0.869	0.012	69.982	0.000
F2 with F1	0.673	0.058	11.691	0.000

R-Square

The squared multiple correlations under *R*-square (R^2) provide information on how much variance the common factors account for in the observed variables (see Table 7). Alphabet skills and letter sounds (F1), for example, explains 39.6 % of the total variance in uppercase letter names (A). Reading and decoding (F2) account for 86.8% of the total variance in variant vowels (J). The low R^2 of .119 suggests that vowel sounds (D) can explain only a small portion of variance of alphabet skill and letter sounds (F1). The correlation between the two common factors, (F1) and (F2), is .68 ($p < .000$). An acceptable correlation is .70-1.0 (Cohen, 1988). The two factors appear to fall in to the range of being significantly related each other.

Table 7

R-Square

Observed variable	Estimate	S.E.	Est./S.E.	Two-tailed <i>p</i> value
A. Letters uppercase	0.396	0.097	4.073	0.000
B. Letters lowercase	0.452	0.103	4.384	0.000
C. Consonants	0.207	0.045	4.625	0.000
D. Vowels sounds	0.119	0.043	2.749	0.006
E. CVC	0.484	0.036	13.613	0.000
F. Blends	0.669	0.034	19.898	0.000
G. Digraphs, tri-	0.753	0.024	31.507	0.000
H. R-controlled	0.811	0.021	38.902	0.000
I. Long spellings	0.820	0.020	40.716	0.000
J. Variant vowels	0.868	0.015	57.029	0.000
K. Low frequent spellings	0.822	0.017	47.327	0.000
L. Multisyllable	0.755	0.022	34.991	0.000

*Determining the Reliability of the CORE Phonics Survey**Classical Test Theory*

*Question 1. What is the evidence of test-retest reliability on the CORE Phonics Survey as measured by a Pearson's *r* correlation coefficient?*

Reliability is the consistency of a measurement, or the degree to which an instrument measures the same way each time it is used under the same condition with the same subjects. One way to measure this consistency is to give a test twice and compare the results of each. A measure is considered reliable if a person's score remains stable from one testing occasion to another. A Pearson's *r* was calculated to estimate the correlation between the 50 student scores on two different occasions two weeks apart. Reliability estimates between .70-.80+ are considered acceptable (Reynolds et al., 2009).

The resulting correlation was .92, which shows a high degree of score stability over time, and thus represents a high estimate of test-retest reliability.

Question 2. What is the evidence of interrater reliability on the Core phonics Survey as measured by a Pearson's r correlation coefficient?

Another way to estimate test reliability is to see if two different raters score the same when testing the same students. A measure is considered reliable if a person's score remains stable from one rater to another. Correlation coefficients of .70-.80 are considered adequate to strong. A score of .90+ is considered high (Reynolds et al., 2009). Twenty-five students were administered the Core Phonics Survey with two different raters scoring the results. The resulting correlation of .98 indicates that the consistency of the scores between both raters was very high correlation.

Question 3. What is the evidence of internal consistency reliability on each subtest and the total CORE Phonics Survey as measured by a Cronbach's Alpha coefficient?

Internal-consistency measures the correlation within items of the instrument itself. It estimates the degree of inter-correlation between the items within each part of the test. To determine the internal consistency of the CORE Phonics Survey, a Cronbach's alpha was calculated for each subtest on the test (A-L) independently (see Table 8). Like other measures of reliability, Cronbach's alpha ranges from 0 to 1.0. When the measure is totally inconsistent it is close to 0 and when the items correlate perfectly with one another it is 1.0. A high alpha coefficient indicates that the items within the subtest are highly intercorrelated. An alpha of .70 or higher is considered acceptable for most purposes and

Table 8

Cronbach's Alpha Analysis for CORE Phonics Survey

Survey section	Title of section	Cronbach's alpha
Part A	Alphabet letters uppercase	.464
Part B	alphabet letters lowercase	.488
Part C	Consonant sounds	.634
Part D	Vowel sounds	.811
Part E	CVC words	.819
Part F	Blends	.838
Part G	Digraphs, trigraphs	.887
Part H	R-controlled vowels	.925
Part I	Long vowels spellings	.944
Part J	Variant vowels	.911
Part K	Low frequency spellings	.939
Part L	Multisyllable words	.968

will be the standard for this analysis (Reynolds et al., 2009). The results show a high degree of intercorrelation between the items within each subtest D-L, all falling between .81 and .97. Parts A (.464), Part B (.488) and Part C (.634) have a moderate intercorrelation between test items in each subtest (Reynolds et al.).

Modern Test Theory

Question 1. To what degree do the raters and occasions contribute to variance among scores on the CORE Phonics Survey as measured by a G study?

A generalizability study (*G Study*) identified the interactions among students, raters, and occasions on the CORE Phonics Survey. A fully crossed, two-facet design permitted a partitioning of the observed score variance into seven separate variance components. The amount of these seven variance components is reported in Tables 9 and 10. Three of the variance components are large relative to the others. The three largest

Table 9

Estimated Variance Components and Standard Errors for Part 1 (Sections A-D)

Source of variability	<i>df</i>	Variance component	<i>MS</i>	Total % of variability
Student	23	3.028	6.456	49.8
Rater	1	.019	.844	.3
Occasion	1	.000	.094	0
Student x rater	23	.000	1.387	0
Student x occasion	23	1.048	4.507	24.3
Rate x occasion	1	.000	.094	0
Student x rater x occasion	23	1.550	1.550	25.5
Total	95			100.0

Table 10

Estimated Variance Components and Standard Errors for Part 2 (Sections E-L)

Source of variability	<i>df</i>	Variance component	<i>MS</i>	Total % of variability
Student	23	339.420	1419.99	89.8
Rater	1	.000	.667	0
Occasion	1	2.634	187.042	.7
Student x rater	23	.000	7.710	0
Student x occasion	23	27.303	63.172	7.2
Rate x occasion	1	.000	6.000	0
Student x rater x occasion	23	8.565	8.565	2.3
Total	95			100.0

variance components for section one of the CORE Phonic Survey are (a) students (49.8%), (b) student by occasion (24.3%) and (c) student by rater by occasion (25.5%).

The three largest variance components for section two are (a) students (89.8%), (b) student x occasion (7.2%) and (c) student by rater by occasion (2.3%).

Variation Between Students

Students' performance on the CORE Phonics Survey was the object of measurement and represents the target population for making inferences. The mean rating for each student was calculated by averaging the student's scores across two occasions and across two raters. The resulting rating for each student was an estimate of his/her universe or true score.

The variance component indicates how much the universe scores varied from one student to another for the 25 students, providing an estimate of how much the students varied in phonics ability. Ideally, the total variability for students should be larger than other sources of variance since these scores should reflect the real differences in students' phonics ability, or true variance. Any other variance would be sources outside of student ability and, therefore, be sources one would want to minimize or eliminate. The variance component between students both in Part 1 and Part 2 of the CORE Phonics Survey was large relative to the other variance components which indicates that this assessment could indeed reliably detect differences in the decoding abilities of the individual students. Nearly half of the total variance in Part 1 (.489) and 90% (.898) in Part 2 in the scores is due to differences in students' decoding ability.

The Student-by-Occasion Interaction

Since the variance component for students is considered true variance, the largest source of error variability was student-by-occasion. The outcome meant that a student's ranked decoding ability in relation to the other students was affected by differences in the testing occasion. If there was no interaction between student and occasions, each student

would have received the same ranking on both occasions. In Part 1 of the test, 24.3% of the variability in student's ranking on decoding ability was due to the occasion the test was given. In Part 2 of the test, it was only 7.20%, a smaller percentage.

The Residual Variance

The residual variance component included the three-way student-by-rater-by-occasion interaction plus any variation in the rating resulting from other unidentified sources of variance. Since there is no way to remove the three-way interaction from the other unidentified sources, it cannot be determined how much of the residual variance was caused by the three-way interaction or how much is due to other unidentified sources. 25.5% of the total variance in Part 1 was due to either the three-way interaction or other unidentified sources. A much lower percentage (2.30%) of the total variance in Part 2 was due to the three-way interaction or other unidentified sources.

Variance Due to Rater

The variance component for raters was very small (Part 1 = .3%; Part 2 = .00%). This variance is an estimate of the amount of variability in the mean ratings of the two raters averaged across 25 students and both occasions. Less than 1% of the total variability in the ratings was due to the differences in rating and indicates that the raters were essentially equal in their scoring of each student.

The Student-by-Rater Interaction

There was no indication of variance in the overall student means measured by the two raters on two different occasions (0.00) on either part of the CORE Phonics Survey.

This indicates that the students tended to be ranked in the same order by both raters and were equal in their score markings.

Variance Due To Occasion

The variance component for occasions was also negligible, 0.00 and .07, indicating that the variability of student means as measured by the two raters on the two occasions were equal for the group of students as a whole.

The Rater-By-Occasion Interaction

The variance component reported for rater-by-occasion was 0, indicating that the raters scores ranked the students' performance on the CORE Phonics Survey consistently on both occasions. This is true for both section 1 and section 2 of the test.

Generalizability Coefficients

There are two different generalizability coefficients. The *G* coefficient indicates the reliability of using a student's mean rating (averaged across all raters and all occasions) as a basis for comparing that student's relative standing or rank order of decoding ability to that of the other students' in the group. The *G* coefficient then is helpful for making decisions about which students are better or poorer decoders compared to the mean of a the group as a whole. Decisions of this kind are called *relative decisions* in generalizability theory. The *G* coefficient for the first section (parts A-D) of CORE Phonics Survey is .729, an acceptable reliability coefficient (Cohen, 1988). The *G* coefficient for the second section (parts E-L) of the CORE Phonics Survey

is a strong coefficient of .956 (Cohen).

The phi coefficient is also a generalizability coefficient, which describes the reliability of using a student's mean rating (averaged across all rater and all occasions) as a basis for comparing that student's decoding ability to a pre-established criterion or score. These are called *absolute decisions* because they describe the student's performance in comparison to an absolute standard with no consideration to the other students' performance. The phi coefficient is usually smaller than the corresponding *G* coefficients since the phi coefficient accounts for all sources of error in the ratings, whereas the *G* coefficients account only for the error sources, which contribute to the differences in the students' relative standing within a group. The phi coefficient for part 1 (A-D) is .727. The phi coefficient for part 2 (E-L) is .952. These coefficients are reliable to strong (Cohen, 1988).

Question 2. What is optimal number of occasions and raters when administering the CORE Phonics Survey to minimize error variance and optimize the reliability of the resulting rating as measured by a D study?

A *D* study is used for making decisions about the way to reduce or filter out error or variance. It addresses the question of how to minimize error and optimize reliability when using the CORE Phonics Survey by predicting what effect increasing or decreasing the number of levels of each facet will have on each of the test reliability. This helps to identify optimal conditions for reducing passage and rater variance.

Table 11 indicates the differences in the mean scores when raters and occasions are adjusted between raters and occasions. It shows that as the number of raters and

Table 11

D Study Phi Coefficients

Rater	Occasion 1	Occasion 2
Section 1 of CORE Phonic Survey		
Rater 1	.498	.664
Rater 2	.572	.727
Section 2 of CORE Phonic Survey		
Rater 1	.898	.946
Rater 2	.908	.952

occasions are increased, the reliability coefficient is improved. Also noted is the fact that increasing the number of occasions produces a greater effect than increasing the number of raters.

Figure 2 depicts the possible outcome results in Section 1 of the CORE Phonics Survey when the raters and occasions are interchanged. The solid line represents the scores when testing occurs on just one occasion. The dotted line represents the scores when testing occurs on two occasions. In section 1, the difference between raters on one or two occasions is $(.498 - .572 = .074)$ and $(.664 - .727 = .063)$, respectively. However, when the occasions are considered, the differences are greater. The difference between occasions with one or two raters is $(.498 - .664 = .166)$ and $(.572 - .727 = .155)$. Increasing the number of occasions reduces a greater amount of error variance than increasing the number of raters.

Figure 3 depicts the possible outcome results in Section 2 of the CORE Phonics Survey when the raters and occasions are interchanged. The solid line represents the scores when testing occurs on just one occasion. The dotted line represents the scores

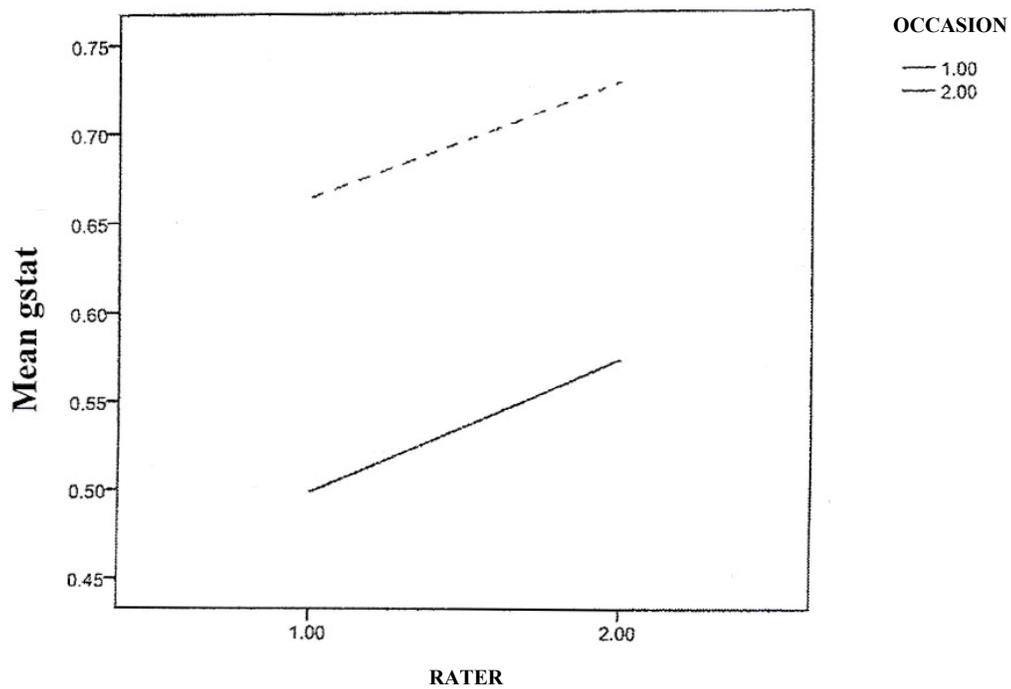


Figure 2. G-stat for alphabet skills and letter sounds: Part 1.

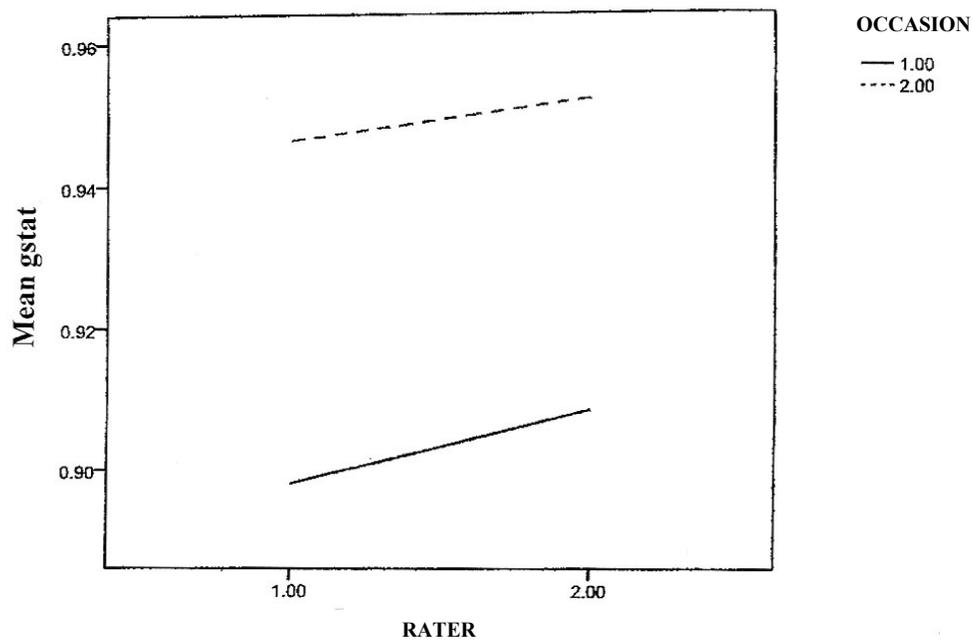


Figure 3. G-Stat for reading and decoding skills: Part 2.

when testing occurs on two occasions. In section 2, the difference between raters on one or two occasions is $(.898 - .908 = .010)$ and $(.946 - .952 = .006)$, respectively. However, when the occasions are considered, the differences are greater. The difference between occasions with one or two raters is $.898 - .946 = .048$ and $(.908 - .952 = .044)$. Although the differences of increasing raters or occasions is minimal in section 2, increasing the number of occasions does reduce a greater amount of error variance than increasing the number of raters.

Standard Error of Measurement

The standard error of measurement (SEM) is a statistic in generalizability that tells how closely the test scores given by the raters on that occasion are to the students' universe score and is an estimate of how much a student's score will likely vary from one occasion to another. Smaller values of this statistic result in more dependable ratings, which allows for more precise generalizations from the student's estimated score to his or her universe score. A SEM score of 0 would indicate that the score would not vary from one test occasion to another and that the score is equivalent to the student's true score. This is important to consider when making relative decisions based on a student's score. The SEM is figured by taking the square root of the relative error variance.

Table 12 shows how the size of the standard error of measurement varies as a function of the number of raters and occasions used. As the numbers of occasions and raters increases, the standard error of measurement decreases, however, increasing the number of occasions has a greater effect on decreasing the error than increasing the

Table 12

Standard Error of Measurement

Rater	Occasion 1	Occasion 2
Section 1 of CORE Phonic Survey		
Rater 1	1.749	1.238
Rater 2	1.504	1.066
Section 2 of CORE Phonic Survey		
Rater 1	6.20	4.39
Rater 2	5.85	4.12

number of raters. This is true for both sections of the CORE Phonics Survey. For example in section 1, one rater on one occasion produces a SEM of 1.749, where as one rater on two occasions produces a SEM of 1.238.

Summary

The CORE Phonics Survey is a widely used tool of assessment that needed statistical evidence to support its use. Several statistical tests were analyzed to determine the degree of validity and reliability of it.

Classical Validity Testing

Content validity was strengthened by comments from phonics experts John A. Smith and J. Lloyd Eldredge. In criterion validity there was a 90%+ overlap in concepts when comparing the CORE Phonic Survey and the Utah State Language Arts Core and 94% overlap when comparing concepts with the Scott Foresman Core Reader, 2008.

Modern Validity Testing

Construct validity was shown with a CFA, which indicated that the two-factor construct was supported having a correlation .68 for factors 1 and 2.

Classical Reliability Testing

Classical reliability testing of test/retest, interrater and internal consistency was done using Pearson's r and Cronbach's alpha. Test/retest results showed a .92 correlation and a .98 interrater correlation. Inter-item correlations on Part 1 of the test were moderate and on Part 2 were high.

Modern Reliability Testing

Approximately 50% of the variance in scores in part 1 was due to differences in students' phonics knowledge. About 90% of the variance in part 2 was due to differences in students' phonics knowledge. The phi coefficient in the *D* Study indicated that increasing the number of testing occasions reduces error variance in testing.

CHAPTER V

DISCUSSION

As discussed in the introduction, research shows evidence that phonic knowledge and acquisition is necessary for beginning readers and therefore, appropriate assessment is essential in identifying specific student needs in this area. Assessments need to be valid and reliable in order to insure that students are progressing in the desired direction. The purpose of this study was to investigate the psychometric properties of the CORE Phonics Survey, a widely used decoding assessment, and to determine whether this test is a valid and reliable assessment tool for phonics knowledge. Seven analyses were done to examine the validity and reliability of the CORE Phonics Survey. The following are the conclusions derived from the analyses and interpretation of the results, their limitations and recommendations for future research.

Discussion of the Validity Analysis

Classical Test Theory

Content Validity

The comments from John A. Smith supported many of the differences discussed above in the criterion validity analysis. The items left out of the CORE Phonics Survey (i.e., compound words, contractions, prefixes, and suffixes) may be items to be classified under other headings. Sight words and structural analysis items may need to be in categories separate from phonics. Other than those concerns, John Smith thought the

CORE Phonics Survey was a strong assessment of phonics knowledge.

J. Lloyd Eldredge claimed that, overall, the CORE Phonics Survey has content validity. Specifically it has content validity for both implicit and explicit phonics advocates. It also has content validity for phonics advocates who recognize the need to break multi-syllable words into appropriate syllables. Additionally, he agreed that this assessment represents the most important phonics elements. Finally, the CORE Phonics Survey includes pseudo words that represent the structure of real words.

Criterion Validity

There was a 90% overlap in the skills identified in the Utah State Language Arts Core Curriculum on phonics and in those addressed in the CORE Phonics Survey. Most every skill required by the state of Utah in phonics was represented to some degree on the CORE Phonics Survey. The only skill at the kindergarten level that was not included on the CORE Phonics Survey was that of matching a letter to an isolated sound. The CORE Phonics Survey does require the students to identify the letter and produce the letter sound, but it does not require the student to identify the letter when given an isolated sound.

On the first- and second-grade level, contractions, compound words, prefixes, and suffixes were skills not included on the CORE Phonics Survey that were on the Utah State Core. Although these are certainly commonly taught skills in reading instruction, there is often a discrepancy as to whether these are categorized as decoding skills or structural analysis skills. Many authors of reading texts would classify them as structural analysis skills outside of the construct of decoding. This rationale was confirmed by

John A. Smith's critique. His opinion was that these skills were part of structural analysis skills rather than decoding skills.

The only other skill that was not addressed in the CORE Phonics Survey that was included on the Utah State Core Curriculum was that of word families or onset and rime. e.g., -ack, -ail, -ake. These word chunks were included within the word lists on the word list, but the test does not require the student to use the word chunk as a means of decoding other unknown words. However, the purpose of teaching word families is to provide a strategy for young readers to identify unknown words. Because the CORE Phonic Survey is interested only in assessing if the students can identify unknown words, the authors of the CORE Phonics Survey may have thought it unnecessary to include this skill on an assessment.

Moreover, the CORE Phonics Survey includes additional skills of phonics not mentioned in the Utah State Core Curriculum on phonics. Variant vowel combinations (J) such as ew, ow, ue, ou, oo, ew, ou, aw are also included as important skills. Low-frequency vowel and consonant spelling (K) included the phonics concepts of silent letters (kn, gh, wr, gn, b, wr,), soft c/g sounds, and y as a vowel. This seems particularly noteworthy since variant vowels (J) shows to have the strongest relationship (.93) in Decoding and Spelling skills than of any other factor. Low-frequency vowel and consonant spelling (K) also has a very strong relationship (.91) in decoding and spelling skills. Therefore, including these additional skills strengthens the validity of the CORE Phonics Survey.

Modern Test Theory

Finally, an important area of validity is construct validity. Confirmatory factor analysis is a multivariate statistical method that seeks to confirm or refute a hypothesized structure in the data. The authors of the CORE Phonics Survey had hypothesized that phonics basically comprised of two-factors: (a) alphabet skill and letter sounds, and (b) reading and decoding skills. Results of the CFA supported this conceptualization. The correlation between factor one (alphabet skill and letter knowledge) and factor 2 (reading and decoding) was .68.

In the CFA, the correlations in factor two are stronger than those in factor one. Examining the descriptive analysis table (Table 3), may explain the low correlations in factor one. There is very little variance in the student scores in parts A, B, C, and D. The skills in these sections are ones that are often very easy even for many first graders (i.e., naming alphabet letters), let alone second graders. Therefore, most students got these questions correct. Table 2 showed that the mean was only one point (or less) than the total possible. The column entitled *3rd quartile* shows that 75% of the students are scoring the total possible. With such a small range of variation in scores, the correlations will remain low even if in reality those parts of the test are highly correlated.

Although none of the tests of goodness of fit hit the exact cut-off values, Hu and Bentler (1999) provided a rule of thumb for deciding which statistics to report and choose cut-off values for declaring significance. When RMSEA values are close to .08 or below and CFI and TLI are close to .95 or greater, the model may have a reasonably good fit. Therefore, although goodness-of-fit was not established with the χ^2 , the RMSEA, CFI

and TLI together support an adequate goodness-of-fit for the predicted model and the data results.

Discussion of Reliability Analysis

Classical Test Theory

Test/Retest and Interrater Reliability

These forms of reliability testing show how consistent test results are when given on more than one occasion or by different raters. If scores of individual tests vary markedly from one testing to another, the change may be attributable to problems with test reliability rather than to changes in the people being tested. The results of the Pearson's r indicated that the reliability of the CORE Phonics Survey when measuring the effects of the occasion was .916. The results when measuring the effect of the rater was .979. This high reliability coefficient indicates that the results on different testing occasions and among different raters would be very close to the same.

Internal Consistency Reliability

Results show a high degree of inter-item correlation within each subtest D-L, all falling between .811 and .968. Part C (consonant sounds) has a moderate inter-item correlation (.634). Part A, letter names uppercase (.464) and Part B, letter names lowercase (.488) show a low inter-item correlation. This indicates that knowing letter A, for example, does not have a high correlation for knowing letter Q. Letter names should each be tested individually to assess students overall knowledge on letter names.

Modern Test Theory

Generalizability testing was done for each section of the test separately.

Section One

The interpretation of the data collected in the generalizability study for the CORE Phonics Survey indicates that Section 1 of the survey had a moderately strong positive coefficient (.73) when crossed with raters and occasions. Nearly 50% of the variance in the student scores was due to true variance in the students' knowledge of alphabet skills and letter sounds. The facet that contributed most to the error variance in the students' scores was the number of occasions the test was administered. The data provided in chapter four shows that the 24% variance was due to the student by occasion interaction. This degree of variance would mean that students' scores in this section of the test would not necessarily reflect a students true score in alphabet skill and letter knowledge. However, increasing the number of occasions significantly decreases this amount of variance.

The residual variance obtained in section one is also quite high (25.5%) and indicates that perhaps a three-facet design would be preferable to the two-facet design. The residual variance reflects the interactions between student by rater by occasion, *plus* any other unexplained interactions that may have had effect in the outcomes. Adding a third facet to the design would permit the researcher to estimate one new variance component, which would help explain more of the 25.5% variance reported in the analysis.

A possible third facet to include is *task*. On sections A-D, the sections have different tasks. Two of the tasks (A and B) are to name the letters and two of the tasks are to give the sounds (C and D). Completing a generalizability study on these sections individually would shed light on whether the nature of the task was part of the variance.

Section Two

In the second section of the Core Phonic Survey, entitled Reading and Decoding skills, there was a very strong positive generalizability (95.6%) in student scores. Nearly 90% of the variance in the students' scores was due true variance, or the differences in student knowledge on reading and decoding skills. Students crossed with occasions contributed only 7.2 % of the error variance. Although this is not a large portion of the variance, in the D study it shows that increasing the number of raters offers an even more reliable basis for making absolute decisions. Oosterhof (1996) advised, "Because the usefulness of assessments is significantly reduced if our observations fail to generalize beyond what we observe, it is important to be aware of the conditions that reduce generalizability" (p. 45). Raters did not contribute any error variance and suggests students' score would be essentially the same regardless of the rater.

From the phi coefficient, we learn that increasing the testing occasions to two would decrease the error variance and improve the generalizability of the test. In section one of the CORE Phonics Survey, the addition of one more testing occasion improved the generalizability significantly from .572 to .729, a substantial difference. It also had a positive effect in section two, but it was not as substantial due to the fact that the true variance was already very high. In section two, the correlations were above .90

regardless of the number of raters or occasions in this section.

It is interesting that although the generalizability coefficient for section two (.956) is significantly larger than that for section one (.729), the SEM for section two is also larger. The SEM indicates how a student's test score would likely vary from one testing situation to another if that student were repeatedly tested. A standard error measurement (SEM) of 0 for a set of test scores, for instance, would indicate that a student's score would not vary from one administration to another and thus be a true reflection of the true score. The fact that the SEM for scores in section 2 with one rater and one occasion would have an error variance of plus or minus 12.4 (6.2×2 SEMs), should cue teachers in taking thoughtful care in making relative decisions about students standing within the group based on individual scores. When students' scores border cut-off percentages for ranking placements, teachers should especially take into consideration the twelve points that may change that student's ranking status within the group.

When considering the rater as a factor of error variance, that did not seem to be much of a contributing factor. Therefore, it can be assumed that the amount of training provided the raters was sufficient. All raters were reported as having approximately 30 minutes to 1 hour training on administering this assessment. The small amount of needed training for this test contributes to the usefulness and practicality for classroom settings.

Give the popularity of the CORE Phonics Survey as a decoding assessment, the findings of this study are particularly important. Statistically, this assessment holds up to the standards of validity and reliability on every measure addressed in this study. Each area showed an adequate to strong result in the testing.

Implications for Instruction

Give the popularity of the CORE Phonics Survey as a decoding assessment, the findings of this study are particularly important. Statistically, this assessment holds up to the standards of validity and reliability on every measure addressed in this study. Each area showed an adequate to strong result in the testing.

In the area of validity, this assessment tool appropriately represents a two-factor construct of decoding with appropriate concepts that support that construct. Teachers can feel confident that approximately two thirds of the variance of students scores is do to phonics knowledge. It is also helpful for teachers to understand that phonics is a two-factor construct so that they can address both of these factors in their instruction. If a teacher taught kindergarten, letter names and sounds would be the focus. As the students progressed, concepts listed under reading and decoding would then be addressed.

The content included in the test closely parallels the concepts outlined in a national basal reading program and then goes beyond what is included in the Utah State Core Curriculum. In the areas of variant vowels and long vowel spellings the CORE Phonics Survey includes additional phonic concepts. This is noteworthy because these two skills showed to have the highest relationships to reading multisyllable words. Teacher can be aware to include these concepts in their instruction as a way of helping students better read multisyllable words.

In the area of reliability, one particular area of interest to practitioners would be the benefit of including two testing occasions when testing alphabet skills and letter sounds. Increasing the number of testings may increase the reliability of identifying

students' real score.

Overall, this study shows that the CORE Phonics Survey is a moderate to strong tool for identifying strengths and weaknesses in phonics knowledge. Educators can feel assured that the data acquired from this assessment can appropriately identify areas of need for students and can provide information that indicates if reading goals are being met. This is an important finding for educators since phonics has been identified as one of the crucial early literacy skills that make a difference in reading success (Cunningham & Cunningham, 1992; NRP, 2000).

Limitations

One limitation of this study was that the schools used in this study were not randomly selected and thus the results cannot be truly generalizable to all populations. These were accessible schools that had already begun using the new edition of the CORE Phonics Survey. Because this edition was so new at the time that this study was conducted, very few schools were using it and therefore, selection was dependent on those schools that did. However, though the schools were not randomly selected, there was consideration in selecting schools that were varied in SES and student achievement scores in reading.

Another limitation was the number of expert opinions that were reported in the content validity section. Only two of the three experts on reading and decoding instruction responded to the invitation to review the CORE Phonics Survey. Although the feedback from John A. Smith and Lloyd Eldredge was very valuable, it would

strengthen the content validity of this assessment if other expert opinions were also included in this report.

An additional limitation was the possibility of test anxiety. Students in School D were administered the test by university students, they may have performed better or worse on the testing depending on their relationship with the individual(s) giving the test. In a typical classroom situation, classroom teachers rather than unknown adults administered the CORE Phonics Survey.

Finally, in the generalizability testing, only 25 students and 2 raters participated on 2 occasions. Small sample sizes cannot be generalized to all other students and raters and occasions. Therefore, replicating this study with a bigger population would increase the generalizability of the findings.

Recommendations for Future Research

An obvious recommendation that would strengthen the content validity of the CORE Phonics Survey would be to further solicit expert opinions concerning the completeness and accuracy of the items on this assessment.

Additionally it was pointed out in the discussion of results that the error variance in the generalizability testing of section 1 of the CORE Phonic Survey (Alphabet Skills and Letter Sounds) could be diminished by increasing the number of possible sources of error variance and repeating the analysis. Including another facet, namely task, may explain the amount of unexplained variance. Section A and B have a similar task of naming the alphabet letters. Sections C and D require an additional task of producing the

sound of the letters. This could help to explain the sources of error variance found in that section. It is recommended that a generalizability study be done on each of these sections individually. This may help teachers eliminate other sources of error variance when testing.

Finally, it would be interesting to investigate validity further and conduct a predictive validity research to see if students' knowledge in decoding is telling of their future success in reading.

REFERENCES

- Adams, M. G. (1990). *Beginning to read: Thinking and learning about print*. Cambridge, MA: MIT Press.
- Anderson, R. C., Heibert, E. H., Scott, J. A., & Wilkinson, I. A. G. (1985). *Becoming a nation of readers: The report of the commission on reading*. Washington, DC: National Institute of Education.
- Bailey, M. H. (1967). The utility of phonic generalizations in grades one through six. *The Reading Teacher*, 20, 413-418.
- Balmuth, M. (1982). *The roots of phonics: A historical introduction*. New York: McGraw-Hill.
- Bentler, P. M. (1990). Comparative fit indexes in structural models. *Psychological Bulletin*, 10(2), 238-246.
- Camilli, G., Vargas, S., & Yurecko, M. (2003). Teaching children to read: The fragile link between science and federal policy. *Education Policy Analysis Archives*, 11, 1-52.
- Carmines, E.G., & Zeller, P.A. (1979). *Reliability and Validity Assessment*. Beverly Hills, CA: Sage.
- Chall J. (1967). *Learning to read: The great debate*. New York: McGraw-Hill.
- Chall, J. (1983). *Stages of reading development*. New York: McGraw-Hill.
- Clymer, T. (1996). The utility of phonic generalizations in the primary grades. *The Reading Teacher*, 50(3), 182-187. (Original work published 1963)
- Cohen, J. (1988). *Statistical power analysis of the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.
- Consortium on Reading Excellence. (2008). *Assessing reading: Multiple measures, K-8*. Novato, CA: Arena.
- Cunningham P. M., & Cunningham, J. W. (1992). Making words: Enhancing the invented spelling-decoding connection. *The Reading Teacher*, 46, 10-107.
- Dykstra, R. (1968). The effectiveness of code- and meaning -emphasis on beginning reading programs. *The Reading Teacher*, 22, 17-23.

- Ehri, L. C. (1998). Grapheme-phoneme knowledge is essential for learning to read words in English. In J. L. Metsala & L. C. Ehri (Eds.), *Word recognition in beginning literacy* (pp. 3-40). Mahwah, NJ: Erlbaum.
- Eldredge, J. L. (1993). *Decoding strategies*. Dubuque, IA: Kendall Hunt.
- Eldredge, J. L. (1995). *Teaching decoding in holistic classrooms*. Columbus, OH: Prentice Hall.
- Eldredge, J. L. (2003). *Phonics for teachers: Self-instruction methods activities*. Columbus, OH: Prentice Hall.
- Eldredge, J. L., & Bader, L.A. (2004). (2005). *Teaching decoding: Why and how*. Upper Saddle River, NJ: Merrill.
- Flesch, R. (1955). *Why Jonny can't read*. New York: HarperCollins.
- Foreman, B. R., & Moats L. C. (2004). Conditions for sustaining research-based practices in early reading instruction. *Remedial and Special Education, 25*(1), 551-560.
- Gall, M. D., Gall, J. P., & Borg, W. R. (2007). *Educational research: An introduction* (8th ed.). Boston: Pearson/Allyn and Bacon.
- Gay, L. R. (1985). *Educational evaluation and measurement: Competencies for analysis and application*. Columbus, OH: Merrill.
- Good, R. H., & Kaminski, R. A. (Eds.). (2002). Dynamic indicators of basic early literacy skills (6th ed.). Eugene, OR: Institute for Development of Education Achievement.
- Grimm, L. G., & Yarnold, P. R. (2000). *Reading and understanding more multivariate statistics*. Washington, DC: American Psychological Association.
- Goodman, K. S. (1980). Reading: A psycholinguistic guessing game. *Journal of the Reading Specialist, 6*, 126-135.
- Honing, B., Diamond, L., & Nathan, R. (2008). *Assessing reading: Multiple measures for kindergarten through eighth grade*. Novato, CA: Arena.
- Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling, 6*, 1-55.

- Jenkins, J. R., & O'Connor, R. (2002). Early identification and intervention for young children with reading/learning disabilities. In R. Bradley, L. Danielson, & D. Hallahan (Eds.), *Identification of learning disabilities* (pp. 99-149). Hillsdale, NJ: Erlbaum.
- Jöreskog, K. G. (1969). A general approach to confirmatory factor analysis. *Psychometrika*, 34, 183-202.
- Kame'enui, E. J. (2000). *Final report on the analysis of reading assessment instruments for K-3*. Retrieved October 25, 2004, from <http://uoregon.edu/assessment/index.html>
- McCardle, P., & Chhabra, V. (2004). *Voice of evidence*. Baltimore: Brooks.
- National Institute of Child Health and Human Development (NICHD). (2000). *Report of the National Reading Panel. 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* (NIH Publication No. 00-4754). Washington, DC: U.S. Government Printing Office. Also available on-line: <http://www.Nichd.nih.gov/publications/nrp/report/htm>
- 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*. Washington, DC: National Institute of Child Health and Human Development.
- National Research Council. (1999). *Starting out right: A guide for promoting children's reading success*. Washington, DC: National Academy.
- Oosterhof, A. (1996). *Developing and using classroom assessments*. Englewood Cliffs, NJ: Merrill.
- Rasinski, T., & Padak, N. (1996). *Holistic reading strategies: Teaching children who find reading difficult*. Columbus, OH: Merrill/Prentice Hall.
- Rathvon, N. (2004). *Early reading assessment*. New York: Guilford.
- Reynolds, C.R., Livingston, R.B., & Willson, V. (2009). *Measurement and assessment in education* (2nd ed.). Columbus, OH: Merrill/Prentice Hall.
- Scott Foresman Reading Street Series*. (2008). Upper Saddle River, NJ: Pearson Education.
- Smith, F. (1979). *Reading without nonsense*. New York: Teachers College Press.

- Smith, N. B. (2002). *American reading instruction*. Newark, DE: International Reading Association.
- Stanovich, K. E. (1990). A call for an end to paradigm wars in reading research. *Journal of Reading Behavior*, 22(3), 221-231.
- Stevens, J. (1996). *Applied multivariate statistics for the social sciences* (3rd ed.). Mahwah, NJ: Erlbaum.
- Thompson, B. (1998). Statistical significance testing and effect size reporting: Portrait of possible future. *Research in the Schools*, 5(2), 33-38.
- Vellutino, R. R., & Scanlon, D. M. (1991). The preeminence of phonologically based skills in learning to read. In S. A. Brady & D. P. Shankweiler (Eds). *Phonological processes in literacy: A tribute to Isabelle Y. Liberman* (pp. 237-252). Hillsdale, NJ: Erlbaum.
- Wagner, R. K., Torgeson, J. K., & Rashotte, C. A. (1999). *Comprehensive test of phonological processing*. Austin, TX: Pro-Ed.

CURRICULUM VITAE

LORILYNN BASTIAN BRANDT

Contact Information**Office:**

Utah State University
School of Teacher Education and Leadership
6705 Old Main Hill
Logan, Utah 84322-6705
Tel: (435) 797-0895

Home:

29 North 376 East
Smithfield, UT 84335
Tel: (435) 563-1239
Email: Lori.Brandt@usu.edu

Education

Doctor of Philosophy, October 2009. College of Education and Human Services, Utah State University, Logan, Utah. Curriculum & Instruction Specialization, Reading and Literacy Emphasis. Research Interests: Research-based literacy Instruction, assessment, reading motivation.

Dissertation: *Study on Validity and Reliability for the CORE Phonics Survey*.
Chair: Professor Parker Fawson.

Research Interests: Teacher education, literacy-based instructional practices, assessment, motivation

Master of Education, 1995. College of Education, Brigham Young University, Provo, UT. Professional Specialization: Reading and Literacy.
Master's Project: *Inservicing Teachers on Current Reading Practices*.

Bachelor of Science, 1986. College of Education, Brigham Young University, Provo, Utah. Major in elementary education.

Additional Endorsements: Reading Endorsement I & Advanced Reading Endorsement.

Additional Training: Utah State Writing Project, Six Traits of Effective Writing, Cognitive Coaching, Portfolio Assessment, Talents Unlimited, Cooperative Learning, Boys Town Model, Tribes, Math Their Way, Elements of Effective Instruction.

Educational Experiences

Instructor, Department of Teacher Education and Leadership, Utah State University, 1999 to present. Write course syllabus. Provide lectures and guest lectures on selected topics. Evaluate student assignments and practicum teaching experiences. Train pre-service teachers to carry out the assessment- instruction cycle in reading instruction. Write letters of recommendations for student portfolios.

Curriculum and Course Development, Department of Elementary Education, 2001 and 2004. Develop, compile and organize curriculum materials for Independent Study Course (ELED 3100). Manage Independent Study course and help students successfully complete requirements.

Student Practicum Organization, Department of Elementary Education, USU, 2000. Organize practicum experiences for Foundations classes. Work with principals and teachers to establish sites. (ELED 3000).

Supervision, Department Teacher Education and Leadership, USU, 1999-present. Supervise and evaluate pre-service level II and level III education students in local K-8 classrooms.

Research Assistant, EEJ Early Childhood Center, Utah State University, 2005-2009. IES Primary Grade Teacher Quality Research Grant. Observe and evaluate primary grade teachers as data collection for research study.

Teaching Assistant, Elementary Education Department, Utah State University, 2005. Assist with course instruction and supervision of practicum students.

Teacher, Wilson Elementary, Logan School District, Logan, UT. 1997-1999. Grade 2. Classroom teacher. Presented inservice for grades 1-2 teachers on reading/writing workshop. Headed a school wide postal system. Served as cooperating teacher for Utah State Student Teachers.

Teacher, Syracuse Elementary, Davis School District, Farmington, UT. 1989-1997. Grades 1 and 2. Classroom teacher. Served as a cooperating teacher for Weber State University student teachers. Taught district and school inservice classes on Reading and Language Arts topics. Headed a school-wide postal system. Developed and provided inservice for a modified Reading/Writing Workshop for primary grades. Organized a *Meet the Author* program for grades 1-2. Headed a school-wide *Young Writers' Fair*. Organized and headed school-wide *Students as Authors* writing fair for writing project. Served as faculty mentor teacher.

University Courses Taught

Teaching Reading, ELED 3100. Utah State University. Fall 1999 – present.
Covers current research-based reading instruction.

Teaching Reading Satellite Class, ELED 3100. Utah State University. Fall 1999-2007– Covers current research-based reading instruction. Classes broadcast to sites throughout Utah.

Teaching Reading Independent Study, ELED 3100. Covers current research-based reading instruction. Students follow independent study program and receive feedback from the instructor.

Methods in Teaching Reading and Practicum, ELED 4040. Utah State University. Fall 2005 –present.
Prepares students to use data from reading assessments to identify areas of need and plan instruction. Special emphasis on explicit, differentiated instruction for struggling students. Extensive practicum.

Language Arts Methods and Practicum, ELED 4030. Utah State University. Fall 1999- Summer 2001. Explores language development in children and its implications and application in instructional methods and assessment. Practicum experience.

Foundations in Education and Practicum, ELED 3000. Utah State University. Fall 2002, Spring 2003. (3 credits). Introduction to historical, philosophical, and social factors shaping current educational practices. Practicum experience.

Improvement of Language Arts Instruction, ELED 6380. Spring 2002.
Graduate course exploring current topics and instructional practices in language arts.

Professional Presentations

Brandt, L. (2008). *Fluency and vocabulary instruction*. Class presented to *Reading First* teachers, Ogden School District teachers. Ogden, UT.

Fawson, P. C., Reutzell, D. R., and Brandt, L. (November, 2007). *"What's in it for Me?" The Impact of Four Incentive Paths on Third Graders' Decisions About Reading"*. Paper presented at the 51st annual meeting of the College Reading Association, Salt Lake City, Utah.

Brandt, L (2007). *Fluency and vocabulary instruction*. Class presented to Reading First teachers, Granite School District, Salt Lake, UT.

- Brandt, L. (2005). *Early reading and literacy practices*. Workshop presented to the members of the LDS Smithfield church community.
- Brandt, L. (2002). *Using books in the Elementary Classroom*. Presented to Utah State Elementary Education Student Organization.
- Brandt, L. (1999). *Early reading instruction*. Workshop presented to Wilson Elementary Faculty, Logan, UT 84321.
- Brandt, L. (1996). *Reading Instruction for early readers*. Workshop presented to Syracuse, Elementary Faculty, Syracuse, UT.
- Brandt, L. (1995). *Current Trends and Issues in Reading instruction*. A series of six presentations for Davis School District.
- Brandt, L. (1994). *Reading and writing workshop for K-2 classrooms*. Workshop presented to Weber State student teachers.

Professional Service

- Wasatch Campus/ School of Education liaison, UVU 2009.
- Benefits and Budgets Committee, UVU 2009.
- Rand and Promotion Committee, UVU 2009.
- Search committee for department lecturer, UVU 2009.
- Fill-in presenter for *Reading First* seminars, 2007-2008.
- Literacy tutor for individuals and small groups in local elementary schools, 2000-2008.
- Presenter for local community women's organizations on early literacy, 2005.
- Evaluator for group assessments for admittance into the USU Elementary Education program, 2001-2003.
- Keynote speaker for USU student organization on literacy themes, 2002.
- Writing assessment scorer for high school students, 2001-2002.
- Presenter for Weber State student teachers on educational themes, 1995.

Professional Organizations

- International Reading Association (IRA)
- National Council for Teachers of Mathematics (*NCTM*)
- Parent Teacher Association (*PTA*)

Honors and Awards

- Graduate Student Stipend, \$4000/year, 2006.

- Davis County School District *Teacher of the Year*, 1995.
- Syracuse Elementary *Teacher of the Year*, 1995.
- KSL Teacher Feature, 1994.
- Standard Examiner, *Apple for the Teacher Award*, 1994-1996.

References

Dr. Parker Fawson, Associate Professor, College of Teacher Education and Leadership, Utah State University, Logan, UT, 84332. Tel: (435)797-0392; Email: Parker.Fawson@usu.edu.

Dr. John Smith, Professor and Chair, Department of Curriculum and Instruction, University of Texas at Arlington, Box 19777, Arlington, TX 76019. Tel: (817) 272-0116; email: smithj@uta.edu

Dr. Ray Reutzel, Professor, Endowed Chair of Emma Eccles Jones Early Childhood Center, Utah State University, Logan, UT, 84332. Tel :(435)797-8631; email: Ray.Reutzel@usu.edu.

Dr. Jim Dorward, former Interim Department Head and Professor, College of Teacher Education and Leadership, Utah State University, Logan, UT 84332. Tel: (797-0385); email: Jim.Dorward@usu.edu.

Dr. Cindy Jones, Assistant Professor, College of Teacher Education and Leadership, Utah State University, Logan, UT, 84332; Tel: (435) 797-7027; email: cindy.jones@usu.edu.

Bernie Hayes, Professor and Department Chair (retired), College of Elementary Education, Utah State, Logan, UT 84321.

Lindsey Dickinson, principal, Mountain View Elementary, North Layton, UT 84040. Tel: (801) 402-3700.

Ross Quist, principal, Bountiful Elementary, Bountiful, UT 84010. Tel: (801) 402-1350.