ABSTRACT

DYNAMIC INDICATORS OF BASIC EARLY LITERACY SKILLS

This study evaluated the validity and reliability of the Dynamic Indicators of Basic Early Literacy Skills (DIBELS) using a sample of 30 special education students, fourth through sixth grade. Results support the reliability and validity of DIBELS when compared to the Woodcock Johnson III (WJ III). The DIBELS consisted of four subtests: Letter Naming Fluency (LNF), Phoneme Segmentation Fluency (PSF), Initial Sound Fluency (ISF), and Nonsense Word Fluency (NWF). These subtests were compared to two subtests of the Woodcock Johnson III: Letter Word Identification and Word Attack. DIBELS produced significant correlations for LNF, PSF, NWF, and WJ III Word Attack and between DIBELS LNF, PSF, and WJ III Word Identification. ISF exhibited no significant correlation when compared to either subtest of the WJ III.

Alicia Sandman
August 2010
DYNAMIC INDICATORS OF BASIC EARLY LITERACY SKILLS

by

Alicia Sandman

A thesis
submitted in partial
fulfillment of the requirements for the degree of
Master of Arts in Special Education
in the Kremen School of Education and Human Development
California State University, Fresno
August 2010
APPROVED

For the Department of Counseling, Special Education, and Rehabilitation:

We, the undersigned, certify that the thesis of the following student meets the required standards of scholarship, format, and style of the university and the student's graduate degree program for the awarding of the master's degree.

____________________________________
Alicia Sandman
Thesis Author

____________________________________
Colleen Torgerson (Chair) Counseling, Special Education, and Rehabilitation

____________________________________
Paul Beare Dean, Kremen School of Education and Human Development

____________________________________
Susan Tracz Educational Research and Administration

For the University Graduate Committee:

____________________________________
Dean, Division of Graduate Studies
AUTHORIZATION FOR REPRODUCTION
OF MASTER’S THESIS

____ X ____  I grant permission for the reproduction of this thesis in part or in its entirety without further authorization from me, on the condition that the person or agency requesting reproduction absorbs the cost and provides proper acknowledgment of authorship.

__________  Permission to reproduce this thesis in part or in its entirety must be obtained from me.

Signature of thesis author:______________________________
ACKNOWLEDGMENTS

I would like to say a special thank you to my committee members. I am forever in debt to you for your help and encouragement. You have all been such amazing resources throughout this process. Thank you all for your hard work and commitment to helping me create the very best product possible.
TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>Chapter</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>LIST OF TABLES</td>
<td>vii</td>
</tr>
<tr>
<td>1. INTRODUCTION AND OVERVIEW</td>
<td>1</td>
</tr>
<tr>
<td>Background</td>
<td>1</td>
</tr>
<tr>
<td>Goal</td>
<td>1</td>
</tr>
<tr>
<td>Importance of the Topic</td>
<td>2</td>
</tr>
<tr>
<td>Limitations</td>
<td>3</td>
</tr>
<tr>
<td>State of the Field.</td>
<td>3</td>
</tr>
<tr>
<td>Thesis Outline</td>
<td>3</td>
</tr>
<tr>
<td>Summary</td>
<td>4</td>
</tr>
<tr>
<td>2. DYNAMIC INDICATORS OF BASIC EARLY LITERACY SKILLS</td>
<td>5</td>
</tr>
<tr>
<td>DIBELS Measures</td>
<td>6</td>
</tr>
<tr>
<td>Conclusion</td>
<td>16</td>
</tr>
<tr>
<td>3. METHODS</td>
<td>19</td>
</tr>
<tr>
<td>Introduction</td>
<td>19</td>
</tr>
<tr>
<td>Research Questions</td>
<td>19</td>
</tr>
<tr>
<td>Research Objectives.</td>
<td>19</td>
</tr>
<tr>
<td>Participants</td>
<td>19</td>
</tr>
<tr>
<td>Instruments</td>
<td>20</td>
</tr>
<tr>
<td>Procedures</td>
<td>21</td>
</tr>
<tr>
<td>Data Collection</td>
<td>21</td>
</tr>
<tr>
<td>Chapter</td>
<td>Title</td>
</tr>
<tr>
<td>---------</td>
<td>------------------------------</td>
</tr>
<tr>
<td>4</td>
<td>RESULTS</td>
</tr>
<tr>
<td></td>
<td>Data Analysis</td>
</tr>
<tr>
<td>5</td>
<td>CONCLUSION</td>
</tr>
<tr>
<td></td>
<td>Discussion</td>
</tr>
<tr>
<td></td>
<td>Limitations and Future Research</td>
</tr>
<tr>
<td></td>
<td>Practice Implications</td>
</tr>
<tr>
<td></td>
<td>REFERENCES</td>
</tr>
</tbody>
</table>
LIST OF TABLES

<table>
<thead>
<tr>
<th>Table</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Frequencies and Percentages for Demographic Variables</td>
<td>24</td>
</tr>
<tr>
<td>2. Means and Standard Deviations for the DIBELS and the Woodcock Johnson III (N=30)</td>
<td>24</td>
</tr>
<tr>
<td>3. Means and Standard Deviations for the DIBELS by Rater 1 and Rater 2</td>
<td>25</td>
</tr>
<tr>
<td>4. Correlations and Probability Values for DIBELS Subscales by Rater 1 and Rater 2</td>
<td>25</td>
</tr>
<tr>
<td>5. Correlations and Probabilities Among DIBELS Subscales</td>
<td>26</td>
</tr>
<tr>
<td>6. Correlations and Probability Values Between DIBELS Measures And the Woodcock Johnson III</td>
<td>27</td>
</tr>
</tbody>
</table>
Chapter 1
INTRODUCTION AND OVERVIEW

Background
As a focus to identify children who are at risk of failing to learn to read, the United States federal government approved a Reading First grant in August 2002 designed to enhance students’ reading ability in low income schools. As part of the Reading First grant, the California Department of Education (CDE) approved funds for general education K-3 and special education K-12 to be trained in the use of State Board of Education (SBE) adopted Reading Language Arts instructional programs.

In May 2008, California published a recommended list of assessments to be used as preventative measures as well as to measure the successfulness of the SBE adopted Reading Language Arts instructional programs. This list was approved for both special and general educators. Among the list was the Woodcock Johnson III (WJ-III), Test of Phonological Awareness (TOPA), and the Dynamic Indicators of Basic Early Literacy Skills (DIBELS). The screening instrument at hand, Dynamic Indicators of Basic Early Literacy Skills (DIBELS), according to Elliott (2001), provides constructive feedback for use in early identification of children who are at-risk for reading failure.

Goal
This study focuses on the Dynamic Indicators of Basic Early Literacy Skills (DIBELS), a formative assessment, and its validity and reliability in the field of special education. This thesis compares DIBELS to a highly accredited standardized assessment, the Woodcock Johnson III (WJ III). This study takes a
contrary position, arguing that although DIBELS is a nonstandardized formative assessment, the scores are correlated to those of the WJ III, in turn, proving that DIBELS is a valid assessment. This study will also look at the test re-test reliability to assess the reliability of the DIBELS measures when measuring reading ability in special education students. These statements will be validated through a review of the literature and confirmed through the results of this study. The purpose of this study is to prove that the DIBELS test is a valid and reliable assessment when given to special education students.

Importance of the Topic

The study proposes to examine the DIBELS measures and how those DIBELS measures relate to the standardized assessment Woodcock Johnson III. The research questions are:

Question 1: What is the reliability of DIBELS for special education students?

Question 2: Are DIBELS test scores correlated to those of the standardized test WJ III?

The formative assessment DIBELS will be compared to a highly accredited standardized assessment, the WJ III. The primary purpose of this thesis is to prove that there are strong correlations between the two assessments, thus proving that DIBELS is a valid and reliable resource to assess students in special education. As Klecker (2007) explains, teachers who use weekly formative assessment have higher summative assessment scores. However, the assessments used in special education classrooms must be deemed valid and reliable.

Researchers define validity as an indication of how sound research is and applies to both the design and the methods of research. Validity in data collection
means that findings truly represent the phenomenon the study is claiming to measure. Reliability, on the other hand, means “repeatability” or “consistency.” A measure is considered reliable if it would provide the same result over and over again.

**Limitations**

The limitations of the current study must be taken into consideration when interpreting the present results. First, generalizability should be considered when looking at the demographic characteristics at hand. Over 90% of the participants were Hispanic, and all were drawn from special education. However, both the tester and the second rater were trained to administer the WJ III as well as DIBELS. Both the tester and the second rater were trained to administer the WJ III as well as DIBELS, and both persons followed test administration guidelines to ensure validity.

**State of the Field**

The majority of studies on the validity and reliability of DIBELS have focused on students in general education settings. There exists, however, a small amount of literature which focuses on the validity and reliability of DIBELS in classrooms of students with special needs. This thesis will contribute to the representation of special education and DIBELS.

**Thesis Outline**

The thesis will be comprised of five chapters. Following the introductory chapter, which outlines the purpose of the study, chapter 2 will highlight related literature on formative assessment as well as the Dynamic Indicators of Basic Early Literacy Skills. Chapter 3 will focus on the selected methodology. Chapter
4 will outline the findings of the research while chapter 5 will conclude the study through a discussion of the research findings, implications for further research, and limitations of the study.

**Summary**

As the current chapter has expressed, the thesis will focus on the validity and reliability of the formative assessment DIBELS. This study will be conducted through a quantitative approach, focusing on formative assessment DIBELS and how those scores are correlated with those of the WJ III, furthermore proving that DIBELS is a valid and reliable formative assessment to use in the special education classrooms. Through the following review of literature this argument will be supported.
A current issue concerning Dynamic Indicators of Basic Early Literacy Skills (DIBELS) is that many educators do not perceive it as a valid and reliable resource to assess literacy in education. This paper will focus on formative assessment as a whole, as well as the validity and reliability of the formative assessment DIBELS. In doing so, it will address only what the research has produced to validate the use of formative assessment DIBELS in special education classrooms.

In order to prevent reading failure in children, the United States federal government created a Reading First grant in 2003 designed to enhance students’ reading ability in low income schools. Reading First was adopted as a significant piece of the No Child Left Behind Act (California Department of Education, n.d.). As part of the Reading First grant, the California Department of Education (CDE) approved funds for general education K-3 and special education K-12 to be trained in the use of State Board of Education (SBE) adopted Reading Language arts instructional programs.

In May 2008, California published a recommended list of assessments to be used to identify students who need preventative measures as well as to measure the success of the SBE adopted Reading Language arts instructional programs. This list was approved for both special and general educators. Among the list was the Woodcock Johnson III (WJ-III), Test of Phonological Awareness (TOPA), and the Dynamic Indicators of Basic Early Literacy Skills (DIBELS). The screening instrument at hand, Dynamic Indicators of Basic Early Literacy Skills (Good,
Kaminski, Simmons, Kame'enui, & Oregon School Study Council, 2001), according to (Elliott, 2001) provides constructive feedback for use in early identification of children who are at-risk for reading failure. DIBELS has been supported and endorsed by the federal government as part of a Reading First program designed to teach and assess early literacy skills in kindergarten through third grade and special education children. The assessment may also be used in some cases as part of a Response to Intervention program.

**DIBELS Measures**

The DIBELS measures created at the University of Oregon is a formative assessment used to evaluate early literacy skills, which are directly related to later reading competence (University of Oregon, 2001). The assessment was specifically contrived to assess phonological awareness, alphabetic principle, and fluency. The DIBELS measures assessment and includes seven subsets, each designed to assess one or more of the basic early literacy skills (phonological awareness, alphabetic principle, and fluency). The seven subsets include Initial Sound Fluency (ISF), Letter Naming Fluency (LNF), and Phoneme Segmentation Fluency (PSF), which all test for phonological awareness. The Nonsense Word Fluency (NWF) tests children’s alphabetic principle. Oral Reading Fluency (ORF) and Retelling Fluency (RTF) are both used to assess a child’s reading fluency and comprehension. And, lastly, Word Use Fluency (WUF) is a subset designed to assess for vocabulary and oral language skills.

**DIBELS Use in Formative Assessment**

The first focus of this paper is one of much importance, formative assessments. When used correctly, formative assessments can show what a student knows and doesn’t know, and in turn, can help educators set appropriate
learning goals. When asked, Lee (personal communication, October 15th, 2009) shared her insight about formative assessment: “When using formative assessment, the most important idea is to use the data in order to set new goals and guide instruction.” Once those goals are in place they can be used to track progress and set future goals. In general, formative assessments concentrate on improving students’ learning (Yin et al., 2008).

Formative assessments are used when deciding what instruction should come next. Stiggins (2009) provides the picture that classroom and formative assessment provides knowledge about where a student is located in their learning process. This is a continuous process that is done bi-weekly or monthly.

One should also consider the purpose of a formative assessment. Clear learning goals must first be set into place. These goals should be within the developmental reach of the students and manageable within the time allotted to teach the standards. A commitment to standards must also be made. Without the commitment to standards based instruction assessments alone are only used for grading and ranking students. High quality assessments must also be chosen in order to maintain valid and reliable results. The assessment should be research based and appropriate for the learning being measured. Lastly, teachers should communicate assessment results quickly and effectively. Graphs and charts help administrators and children alike to see the results in a comprehensible way (Stiggins, 2009).

The majority of professional educators haven’t been trained to use a standardized formative assessment system (Donovan & Cross, 2002). For many years informal nonresearch-based approaches have been used to assess students who were distinguished as having some learning difficulty. The problem with this model is discussed below. According to Donovan and Cross, a substantial number
of students are failing to learn to read and teachers are not being held accountable. Previously, there have been no measures taken to gauge students’ learning, and therefore teachers have not been held accountable for the yearly growth of a child’s education in kindergarten through third grade and special education. Nor did teachers know what should be taught next if they were unaware of what students know and don’t know.

In a study to measure formative assessment and its impact on student motivation, Yin et al. (2008) conducted a small research experiment involving two groups. One group used a curriculum with embedded formative assessment. The other group used a curriculum without embedded formative assessment. Twelve teachers and their students were assigned to either the experimental group or the control group. Both groups were given the same pre- and posttest. Neither group was informed of the study to protect against the Hawthorne and John Henry effects. Finally, the results of those tests were examined for effects of motivation and/or achievement due to formative assessment.

To measure achievement, Yin et al. (2008) developed four assessments: a multiple choice test, a performance test, a short answer assessment, and a predict-observe-explain assessment. Altogether, the four tests consisted of 38 items on both the pretest and posttest. The four assessments were examined for interrater reliability.

Yin and his colleagues (2008) found that both the descriptive statistics and HLM analyses showed that the assessments embedded in the curriculum used by the experimental group didn’t have a significant influence compared with those of the students in the control group. It did, however, prove that the way a teacher may implement formative assessment is crucial to the effectiveness of motivation in the classroom setting. Teachers must not only administer the formative assessment
correctly, they must also use the results to guide future instruction. It is not enough to simply give the assessment. Educators must also use the data to reflect upon and provide positive feedback.

Frey (2009) engaged in a study involving elementary school teachers at an urban elementary school servicing over 1,200 students, grades kindergarten through fifth. The subjects were involved in a series of professional development classes to gauge the effect of teacher use and student success on formative assessments. The professional development classes focused on three main topics. The first was to develop a pacing calendar that worked for each grade level. The teachers came together to agree upon appropriate standards, vocabulary, and instructional strategies for each unit of study. Next, teachers designed common assessment items based on content standards and the unit of study. Finally, teachers met to analyze and discuss test scores and plan for future instruction.

Frey (2009) showed that by collaborating weekly and working together towards a common goal, a system was created which helped to develop, analyze, and plan appropriate learning goals. Hence teachers were able to design and implement effective instruction and interventions. Between the baseline in 2001 and 2005, Frey found a significant increase in student test scores. The study concluded that increased test scores may result from fostering teachers who have been trained to give formative assessments and are able to link classroom instruction with formative assessments.

Klecker (2007) addressed the impact of formative assessment on student learning through an online course at a university level. This study used two Internet classes, one control class with 33 students and the other was the experimental group with 34 students. Students randomly registered for either of two classes, either the control class or the experimental class. The classes were
the same except for one variable; the experimental group was required to complete weekly 20-item formative assessments. At the end of the semester, students in both classes were given a summative assessment containing 60 multiple-choice questions.

The results revealed that the use of weekly formative assessments led to numerically higher summative assessment scores at the end of the semester. The mean for the experimental group was 54.45, and was higher than the control group 51.56, but not a statistical difference.

Similar to the Klecker study, Fisher (2007) conducted research that concentrated on using formative assessment schoolwide at an urban high school in San Diego, California. The study was based on 2,300 high school students. Most of these students qualified for free and reduced lunch, and many of them were also English learners. A goal of this high school was to develop a school wide expectation of making and using common assessments. The product of this study was a four-step process to using formative assessments school wide.

The four steps that Fisher (2007) found useful were developing pacing guides, designing common assessments, conducting item analyses, and engaging in instructional conversation. During the first step of the process teachers of like subjects came together to identify key vocabulary and create similar formative assessments. Teachers also thought of ways to accommodate students with disabilities, English language learners, and gifted and talented students. Secondly, teachers were asked to design common assessments for like subjects. For example, all biology teachers were to design common formative assessments, and all history teachers were to design formative assessments. Teachers attended professional development classes to learn about assessment design. More importantly, they also learned about determining what appropriate instructional
measures to take after giving a formative assessment. Once the teachers gave the formative assessments, item analysis began. A program, Edusoft (software for analyzing data), was used to discriminate between responses to decide where students still lacked knowledge. Finally, teachers were asked to engage in instructional conversation. Each week time was allotted at staff meetings for colleagues to converse with each other to determine what was working and what was not working in the classroom.

Fisher (2007) found that this high school achieved major gains. Students who took biology in 2005 scored 51% basic or above, which was a gain from 28% in 2003. Similarly, students who took history in 2005 scored 49% basic or above, a gain from 30% in 2003.

For an assessment to be formative, the data must be analyzed and feedback must be given. In 1998, Black and Wiliam conducted a study which found evidence that formative assessments, paired with frequent feedback, brought forth substantial gains in academic achievement. It is also clear that formative assessment alone will not suffice.

One weakness of formative assessment found in the literature review done by Black and Wiliam (1998) was that teachers do not implement formative assessment to their full potential. Generally, teachers do not use formative assessments to discuss or review when planning new instruction. Formative assessments were intended to guide instruction, therefore, to be used appropriately, (Black & Wiliam) teachers should review test scores and give immediate feedback to students. Another potential flaw of formative assessment implementation was that teachers have a tendency to foster competition between classmates rather than personal improvement. In many instances, student feedback was shared classwide and even schoolwide, creating a sense of failure for some
students. Black and Wiliam stated that it is important when using formative assessments to give feedback that fosters and celebrates personal improvement by using individual graphs and charts.

**Reliability and Validity of DIBELS**

DIBELS is a formative assessment used to help drive instruction and identify students who may be at risk for reading failure. The reliability and validity of any assessment is important. A study done on 75 kindergarten students from a midwestern city evaluated the reliability and validity of the DIBELS measures (Elliott, 2001). There were 41 male subjects and 34 female subjects in four classes in three elementary schools. Twenty-seven students were receiving free and reduced lunch, and a small percentage were receiving special services. Thirty-seven percent of these students were of non-White ethnicity, while 63% of participants were of White ethnicity.

Elliott (2001) used the following assessments when measuring the reliability and validly of DIBELS: the DIBELS battery, the Woodcock-Johnson Psychoeducational Achievement Battery-Revised [WJ-R] (Woodcock, 1998; Woodcock & Johnson, 1989), the Test of Phonological Awareness-kindergarten form [TOPA] (Torgesen & Bryant, 1994), the Kaufman Brief Intelligence Test [KBIT] (Kaufman & Kaufman, 1990), and lastly Developing Skills Checklist [DSC] (CTB Macmillan/McGraw-Hill, 1990). The DIBELS measures were used as the criterion measures and the WJ-R, TOPA, DSC, and informal teacher questionnaire were used as the predictor measures. The K-BIT scores were used to control for differences in ability in the regression analyses. The regression analysis from this study found that when compared to other assessments, the DIBELS measures were highly correlated. Correlations between the WJ-R and
DIBELS ranged from .75 for the LNF, .72 for SNF, .64 for ISF, and .60 for PSF. SNF and LNF represented the highest correlation. ISF and PSF showed less magnitude.

Implications from this study confirm DIBELS measures to be reliable and valid (Elliott, 2001). However, initial phoneme ability and phonemic segmentation ability still need fine-tuning. In addition, improved training and administration procedures of the Dynamic Indicators of Basic Early Literacy Skills should be refined.

Nelson (2008) looked into the DIBELS’ predictive power of screening at-risk students. This study included 177 general education students in kindergarten from a midwestern state. There were 10 classrooms in two different schools participating. The demographics of these students were predominantly White males. The three levels of risk status—at risk, some risks, and low risk—were designated by test developers (University of Oregon, 2001). The guidelines for at risk students at the middle of kindergarten are scores below 10 on Initial Sound Fluency, 15 on Letter Naming Fluency, 7 on Phoneme Segmentation Fluency, and 5 on the Nonsense Word Fluency. Performance from 10 to 25 on ISF, 15 to 26 on LNF, 7 to 17 on PSF, and 5 to 12 on NWF places students at some risk. Scores above the high end of those ranges are considered at low risk for reading hardship.

After comparing DIBELS scores to those of the WJ-III, findings show that all correlations were significant. DIBELS and WJ-III produced correlations of .38 for ISF, .61 for LNF, .51 for PSF, and .74 for NWF. The DIBELS yielded low correlations as well as positive predictive power and high negative predictive power. This indicates that DIBELS is far better at predicting students who are proficient readers rather than students who are suffering from inadequate reading skills (Nelson, 2008).
Burke (2009) examined the validity of four DIBELS subtests: ISF, PSF, LNF, and NWF. The sample consisted of 218 kindergarten students in general education. The students were tested on the Test of Word Reading Efficiency (TOWRE) (Torgesen, Wagner, & Rashotte, 1999) sight word efficiency and the Woodcock Reading Mastery Test-Revised (WRMT-R) passage comprehension subtest (Woodcock, 1998), as well as the DIBELS measures. Assessments revealed significant correlations ranging from .32 to .89 (Burke). From the two phonologic measures taken PSF was just higher than ISF, but neither were more correlated than the alphabetic measures. NWF was the strongest correlate at .56. Furthermore, Burke found strong evidence of predictive validity especially when considering early literacy skills.

Gonzales (2008) conducted a study to substantiate the validity of DIBELS across student population other than general education elementary students. This study contributes to the field of education by differentiating between low average, average, and above average students with emotional and behavioral disturbance (EBD) based on the Woodcock Johnson Reading Mastery Test-Revised (WRMT-R). Participants included 145 students each identified as at risk of EBD. The students were selected from seven schools within an urban school district in the Midwest. The students in this case were both first and second graders. The mean age for students in kindergarten was 5.6 while first grade had a mean age of 6.2. The demographics of these students were predominantly white males. Data collection began by testing all students individually with both the DIBELS and the WRMT-R. All kindergartners were tested on ISF, LNF, NSW, and PSF. First grade students were tested on LNF, NSW, PSF, and ORF.

Results from the Gonzales (2008) study revealed positive outcomes. For kindergarten only, LNF was significantly correlated with the WRMT-R.
However, all of the first grade DIBELS tests were significantly correlated with the WRMT-R standard scores. These results showed consistency with Elliott (2001) study. DIBELS measures were also compared to the WRMT-R to determine predictive power of group membership: low average, average, and high average. The kindergarten ORF and LNF had the greatest contribution to predicting group membership. The first grade test indicated a high contribution from all DIBELS measures when predicting group membership.

Gonzales (2008) also offers an explanation for the difference between kindergarten results in this study and others similar to this. One reason might be the homogenous group of mostly white males with EBD, which contributed to the low correlation between tests. Another explanation might be that this study tested students only one time where many other studies have tested students several times to obtain a mean score for the DIBELS measures.

One study measures the effectiveness of the DIBELS phoneme segmentation fluency (PSF) in a Response to Intervention model. Hagens (2008) identifies three components of an RTI model used in the study: (a) validated curriculum must be used schoolwide; (b) student progress must be measured consistently and repeatedly; and (c) instruction is modified to meet students’ needs. The identified RTI model is used in conjunction with DIBELS’ PSF to inform instruction, progress monitor consistently, and check for increased student outcomes.

The participants chosen for this study were 75 first grade students from three schools in the Pacific Northwest (Hagens, 2008). The first school chosen to participate was identified as the low socio-economic-school (SES) used as the intervention and control site. The second and third schools were identified as the high SES, used as comparison groups. Dependent measures in this study were
DIBELS PSF and NWF. The independent variables used were SES of students’ families and an instruction program. The instructional program consisted of both English language instruction (ELI) and Math instruction (MI). Both programs were taught for 20-25 minutes per day, 4 days a week. The study was conducted over a 12-week time span. Before week 1 all participants were tested on PSF and NWF tasks. Thereafter, weekly PSF data were gathered and used for guiding instruction. At the end of week 12, all participants were again given the PSF and NWF tests.

The results of this study support the conclusion that using PSF to consistently and repeatedly measure student knowledge and inform instruction resulted in increased phoneme segmentation skills (Hagans, 2008). Ninety-six percent of students receiving ELI along with progress monitoring attained goal outcome levels on the PSF tests. Only 64% of children in the low SES who did not receive ELI attained the goal outcome levels on the PSF tests a difference of 32%. Furthermore, based on the DIBELS PSF measure is effective when monitoring student progress and increasing student outcomes.

Conclusion

The DIBELS assessment includes seven subsets each designed to assess one or more of the basic early literacy skills (phonological awareness, alphabetic principal, and fluency) (University of Oregon, 2001). The seven subsets include Initial Sound Fluency (ISF), Letter Naming Fluency (LNF), and Phoneme Segmentation Fluency (PSF), which all test for phonological awareness. The Nonsense Word Fluency (NWF) tests children’s alphabetic principal. Oral Reading Fluency (ORF) and Retelling Fluency (RTF) are used both to assess a child’s reading fluency and comprehension. And, lastly, Word Use Fluency
(WUF) is a subset designed to assess for vocabulary and oral language skills. When a student has been identified as at risk for reading difficulty, he or she is placed into the RTI program whose main focus is reading intervention. DIBELS is used in, but not limited to, RTI and Reading First programs (California Department of Education, n.d.).

Stiggins (2009) and Fisher (2009) both suggest important steps for using formative assessments. Both of their studies identified using rich quality assessments that commit to assessing standards. Fisher used a common assessment for easier generalization of scores across staff. Selecting appropriate learning goals and setting pacing guides were also important measures to take when using formative assessments. However, the most important was the ability to communicate results. Teachers must collaborate with students, staff, and families to examine results and designate implications for future instruction (Fisher; Stiggins; Yin et al., 2008).

Klecker (2007) found that weekly formative assessments lead to higher summative assessment results. Nonetheless, in order to use formative assessment to its full potential the results of the assessment must be analyzed and used to plan future instruction (Black & Wiliam, 1998). Providing opportunities for teachers to collaborate weekly and work together towards a common goal will create a system which helps develop, analyze, and plan appropriate learning goals (Frey, 2009).

By looking at the data for formative assessment DIBELS, one can see that there is strong evidence of its validity. Burke (2009) and Hagens (2008) both found strong evidence of predictive validity when considering early literacy skills. Students who were tested on DIBELS PSF demonstrated the strongest correlation with the WRMT-R. However, all subtests showed significant predicative validity.
Elliott (2002) and Gonzales (2008) measured DIBELS against the Woodcock Johnson (Woodcock 1998; Woodcock & Johnson, 1989). Both studies came to similar conclusions. Kindergarten tests indicate that only LNF was significantly correlated with the WRMT-R. However, all of the first grade DIBELS tests were significantly correlated with the WRMT-R standard scores. Each of these studies offers explanations for the low correlations.

The research shows empirical evidence that the formative assessment DIBELS is an accurate assessment when measuring reading ability. For many years informal nonresearch-based approaches have been used to assess students who were distinguished as having some learning difficulty (Donovan & Cross, 2002). Formative assessment, though around for many years, is evolving and the research on this topic must not be taken lightly. When incorporating formative assessment into the classroom effectively students’ academic ability improves. Therefore, it is imperative that educators use formative assessment to inform future instruction.

DIBELS is a valid and reliable formative assessment that could be considered for future use. The DIBELS 1-minute probes result in data that should be used to guide instruction in the classroom. Nevertheless, one must not forget the benefit of using a progress-monitoring tool such as DIBELS lies in the extent to which teachers use the data to inform their classroom practices (Ball, 2009).
Chapter 3

METHODS

Introduction

The purpose of this chapter is to provide an overview of the methods and research design applied to a study on the Dynamic Indicators of Basic Early Literacy Skills. The thesis will investigate the reliability and validity of the formative assessment DIBELS. The study will compare correlations of DIBELS test scores to those of the Woodcock-Johnson Tests of Achievement, Third Edition (WJ III; Woodcock, McGrew, & Mather, 2001). Secondly, the study will use test re-test reliability to measure the reliability of the DIBELS assessment.

Research Questions

Question 1: What is the reliability of DIBELS for special education students?

Question 2: Are DIBELS test scores correlated to those of the standardized test WJ III?

Research Objectives

To study the literature on formative assessment and DIBELS as a valid and reliable formative assessment.

To study the correlations between DIBELS test scores and the WJ III test scores.

To determine, by looking at the correlations, whether DIBELS is a valid resource to assess basic literacy skills in special education.

Participants

Participants were 30 students receiving special education services at Jefferson Elementary School in the Fresno (California) Unified School District.
Their qualifications were that they receive special education services in either a Special Day Class or a Resource Specialist Program for students with mild/moderate disabilities. Each student was selected by age and participation in a program for students with disabilities. These students range between grades 3-6 and are between 8 and 12 years of age. The sample’s ethnicity breakdown is as follows: White 0%, Hispanic 83%, African American .06%, Asian American 1%, and other 0%.

**Instruments**

Four clusters from the Dynamic Indicators of Basic Early Literacy Skills (DIBELS) assessment were administered. The four clusters were: DIBELS Initial Sound Fluency (ISF), Letter Naming Fluency (LNF), Phoneme Segmentation Fluency (PSF), and Nonsense Word Fluency (NWF). These were administered throughout the month of February on each of the 30 students.

The Woodcock Johnson III Letter-Word identification and Word Attack subtests will represent the basic reading cluster of the WJ III. The WJ III was selected as the reference standard because of its reliability and validity in the field of education. The WJ III is designed to measure academic achievement in reading and math for the assessment of student’s strengths and weaknesses for special program planning. (WJ III; Woodcock, McGrew, & Mather, 2001). The LW is used for reading decoding while WA measures reading decoding and phonetic coding in analysis and synthesis. These are the two sub-tests that most resemble the DIBELS measures.

Furthermore, McGrew and Woodcock (2001) reported median validity coefficients of .94 and .87 for the LW and WA subtests. The data also reveals median test-retest reliability coefficients of .95 and .83 for the LW and WA.
subtests. Additionally, McGrew and Woodcock set a standard score of 90 in order to distinguish between adequate and inadequate reading ability.

Procedures

An experimental design was used with approximately 30 special education students. Two teachers from a school in Fresno Unified School District were selected to take part in this study. Each of the students were assessed with the non-standardized assessment DIBELS. The DIBELS assessment includes: The DIBELS Initial Sound Fluency (ISF), Letter Naming Fluency (LNF), and Phoneme Segmentation Fluency (PSF), which all test for phonological awareness and the Nonsense Word Fluency (NWF) tests children’s alphabetic principle. The students were also tested on a standardized assessment the WJ III. The two subtests from the WJ-III that the students were tested on include the Letter-Word Identification (LW) and the Word Attack (WA) probe. Students were tested in a quiet classroom by one of the trained teachers. Stopwatches were used for the one-minute DIBELS probes. Each assessment was double rated for inter-rater reliability.

Data Collection

The DIBELS measures were administered from mid-February through mid-March. Each measure was administered individually over a week’s period of time. The WJ III was administered once over a school year and used during assessment procedures annually for all students with an individualized educational plan for mild/moderate disabilities. Each test session took about 20-30 minutes per student.

The tests were administered by the teacher (researcher), as well as another professional in the field of special education who was trained to give the assessments. Test re-test reliability was used in order to get the most reliable
scores. Once each child was given each assessment, the data was converted from a raw score to a standard score.

**Statistical Analysis**

(Statistical Package for the Social Sciences) will be used to correlate the data at hand. Correlations will be run between test scores of the WJIII as well as the DIBELS measures. Data will be analyzed by looking for correlations between the four subtests of DIBELS as well as the two subtests of the WJIII. Another set of correlations will be run to determine the reliability of DIBELS from the test re-test data collected. This data will be used to determine whether DIBELS is reliable when rated for interrater reliability.
Chapter 4

RESULTS

Data Analysis

The purpose of this study was to understand the DIBELS measures and how those DIBELS measures relate to the standardized assessment Woodcock Johnson III. The following questions will be examined to further understand the relationship between the two assessments: What is the reliability of DIBELS for special education students? Are DIBELS test scores correlated to those of the standardized test WJ III? The primary purpose of this thesis is to prove that there are strong correlations between the two assessments, thus, proving that DIBELS is a valid and reliable resource to assess students in special education.

Table 1 depicts the ratio of male to female students that were tested as well as their ethnicity. Only 20% of the subjects were female as 80% of the students were male. The chart also shows that 80% of the students tested were Hispanic. African American and Asian students only amounted to 20% of the study.

Table 2 shows the mean for DIBELS subtests as well as for the WJ III subtests. The standard deviation was also calculated to show how much variation exists from the average. The WJ III produced a much greater variation between the mean and standard deviation than the scores of the DIBELS.

Table 3 indicates the mean and standard deviation for the DIBELS subtests for rater 1 and 2. The mean and standard deviation for both rater 1 and 2 are almost identical which is shown in Table 4 to produce significant correlations between rater 1 and 2.
Table 1

*Frequencies and Percentages for Demographic Variables*

<table>
<thead>
<tr>
<th>Variables</th>
<th>n</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gender</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td>24</td>
<td>80%</td>
</tr>
<tr>
<td>Female</td>
<td>6</td>
<td>20%</td>
</tr>
<tr>
<td>Ethnicity</td>
<td></td>
<td></td>
</tr>
<tr>
<td>African American</td>
<td>3</td>
<td>10%</td>
</tr>
<tr>
<td>Asian</td>
<td>3</td>
<td>10%</td>
</tr>
<tr>
<td>Hispanic</td>
<td>24</td>
<td>80%</td>
</tr>
</tbody>
</table>

Table 2

*Means and Standard Deviations for the DIBELS and the Woodcock Johnson III (N=30)*

<table>
<thead>
<tr>
<th>Variables</th>
<th>M</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>DIBELS</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Initial Sound</td>
<td>36.43</td>
<td>16.55</td>
</tr>
<tr>
<td>Letter Naming</td>
<td>68.27</td>
<td>24.88</td>
</tr>
<tr>
<td>Phoneme Segmentation</td>
<td>44.27</td>
<td>16.11</td>
</tr>
<tr>
<td>Nonsense Word</td>
<td>44.67</td>
<td>13.50</td>
</tr>
<tr>
<td>Woodcock Johnson</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Letter Word</td>
<td>74.30</td>
<td>17.77</td>
</tr>
<tr>
<td>Word Attack</td>
<td>86.87</td>
<td>17.87</td>
</tr>
</tbody>
</table>
Table 3

*Means and Standard Deviations for the DIBELS by Rater 1 and Rater 2*

<table>
<thead>
<tr>
<th>Variables</th>
<th>Rater 1</th>
<th>Rater 2</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>M</td>
<td>SD</td>
</tr>
<tr>
<td>DIBELS</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Initial Sound</td>
<td>36.43</td>
<td>16.55</td>
</tr>
<tr>
<td>Letter Naming</td>
<td>68.27</td>
<td>24.88</td>
</tr>
<tr>
<td>Phoneme Segmentation</td>
<td>44.27</td>
<td>16.11</td>
</tr>
<tr>
<td>Nonsense Word</td>
<td>44.67</td>
<td>13.50</td>
</tr>
</tbody>
</table>

Table 4

*Correlations and Probability Values for DIBELS Subscales by Rater 1 and Rater 2*

<table>
<thead>
<tr>
<th>Subscale</th>
<th>r between Rater 1 and 2</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>DIBELS</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Initial Sound</td>
<td>.821</td>
<td>.045</td>
</tr>
<tr>
<td>Letter Naming</td>
<td>.984</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Phoneme Segmentation</td>
<td>.967</td>
<td>.002</td>
</tr>
<tr>
<td>Nonsense Word</td>
<td>.748</td>
<td>.087</td>
</tr>
</tbody>
</table>
Table 4 presents the interrater reliability for the 4 DIBELS subscales. These ranged from .748 to .984, with one above .80 and 2 above .90. The significance of the DIBELS measures shows one measure significant to <.10, 1 measure significant to <.05, and two measures significant to <.01. These correlations established significant reliability for the data in this study.

Table 5 presents the correlations and probabilities between the DIBELS subscales. These correlations ranged from -.105 to .457. The correlations between letter naming fluency and phoneme segmentation fluency ($r = .457, p = .011$) is significant at the .05 level. The correlations between phoneme segmentation and nonsense word ($r = .309, p = .097$) is significant at the .10 level.

**Table 5**

*Correlations and Probabilities Among DIBELS Subscales*

<table>
<thead>
<tr>
<th>Subscales</th>
<th>Letter Naming</th>
<th>Phoneme Seg.</th>
<th>Nonsense Word</th>
</tr>
</thead>
<tbody>
<tr>
<td>DIBELS</td>
<td>$r (p)$</td>
<td>$r (p)$</td>
<td>$r (p)$</td>
</tr>
<tr>
<td>Initial Sound</td>
<td>.117 (.538)</td>
<td>.239 (.203)</td>
<td>-.105 (.581)</td>
</tr>
<tr>
<td>Letter Naming</td>
<td>-</td>
<td>.457 (.011)</td>
<td>.210 (.265)</td>
</tr>
<tr>
<td>Phoneme Seg.</td>
<td>-</td>
<td>-</td>
<td>.309 (.097)</td>
</tr>
<tr>
<td>Nonsense Word</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

The correlations between initial sound and nonsense word ($r = -.105, p = .581$) is negative, which indicates high scores on one subscale are related to low scores on the other subscale. However, the correlation is close to zero so there really is no relationship between those two subscales.

The Woodcock Johnson letter word subscale in Table 6 shows correlations with DIBELS subscales. DIBELS letter naming fluency (.074) and phoneme segmentation (.065) both correlated with the Woodcock Johnson letter word.
subscale with significance of \( p < .10 \). DIBELS nonsense word subscale had a negative correlation of -0.074 showing no real relationship between the two subscales.

Table 6

*Correlations and Probability Values Between DIBELS Measures and the Woodcock Johnson*

<table>
<thead>
<tr>
<th>DIBELS and the Woodcock Johnson</th>
<th>WJ III Letter Word</th>
<th>WJ III Word Attack</th>
</tr>
</thead>
<tbody>
<tr>
<td>Initial Sound</td>
<td>.071 (.708)</td>
<td>.145 (.443)</td>
</tr>
<tr>
<td>Letter Naming</td>
<td>.331 (.074)</td>
<td>.455 (.011)</td>
</tr>
<tr>
<td>Phoneme Segmentation</td>
<td>.341 (.065)</td>
<td>.482 (.007)</td>
</tr>
<tr>
<td>Nonsense Word</td>
<td>-.074 (.698)</td>
<td>.328 (.077)</td>
</tr>
</tbody>
</table>

The Woodcock Johnson Word Attack subscale shows significant correlations with DIBELS phoneme segmentation .007 \( (p < .01) \). DIBELS Letter Naming Fluency (.011) shows a significant correlation of \( p < .05 \) and DIBELS Nonsense Word Fluency (.077) also revealed a significant correlation of \( p < .10 \). DIBELS Initial Sound Fluency was not strongly correlated with either of the Woodcock Johnson subscales showing no real relationship between the two assessments.
Chapter 5

CONCLUSION

Discussion

This study examined the DIBELS measures and how those DIBELS measures relate to the standardized assessment Woodcock Johnson III. The research questions were:

Question 1: What is the reliability of DIBELS for special education students?

Question 2: Are DIBELS test scores correlated to those of the standardized test WJ III?

In answer to question 1, the evidence of DIBELS reliability for special education students is clear. Inter-rater reliability has been established with high correlations. The test retest data confirms that any trained professional may come and administer the DIBELS probes and get the same data as another professional might.

However, DIBELS validity has come into question. In answer to question 2, three of the four DIBELS probes revealed significant correlations with scores of the WJ III. The Letter Naming Fluency, Phoneme Segmentation Fluency, as well as the Nonsense Word Fluency all proved to be correlated with the WJ III probes given. However, DIBELS Initial Sound Fluency exhibited inadequate significance when compared to the WJ III.

The current study’s results yield high correlations between the DIBELS subscales and the Woodcock Johnson III reading subscales. These results confirm earlier findings (Elliott, 2001) of the high correlations between the informal assessment DIBELS and the formal assessment Woodcock Johnson III. These
results support this study’s hypothesis arguing that although DIBELS is a nonstandardized formative assessment, the scores are correlated to those of the WJ III, in turn, proving that DIBELS is a valid assessment.

The data from the research provide significant support for both the DIBELS Letter Naming Fluency as well as the Phoneme Segmentation Fluency. Strong correlations for the Letter Naming Fluency and Phoneme Segmentation Fluency also emerged as the strongest DIBELS measures by Good et al. (2001). The DIBELS Initial Sound Fluency emerged as the weakest measure having little correlation with either of the Woodcock Johnson III subscales. The Initial Sound Fluency does not show sufficient validity for assessing students with disabilities. The DIBELS Nonsense Word Fluency showed a significant correlation (.077, p<.10) with the Woodcock Johnson III Word Attack subscale. However, the Nonsense Word Fluency produced a negative correlation -.074 with the Woodcock Johnson III letter word subscale. Furthermore, the Nonsense Word Fluency is correlated only with the Woodcock Johnson III Word Attack subscale and not at all related to the Letter Word Identification.

Nelson (2008) reported similar findings with significance for all subtests, as opposed to this study which found no significance to one subtest, the DIBELS Initial Sound Fluency. Gonzales (2008) conducted a similar study on a large sample of students with Emotional Behavioral Disorder (EBD). Gonzales found all DIBELS first grade correlations were significantly related with those of the WJ III.

As displayed in Table 4 (p. 25) the test retest data produced indicative correlations suggesting strong reliability for the purpose of this study. So question 1 (What is the reliability of DIBELS for special education students?) is clear and precise. Rater 1 and rater 2 for the DIBELS measures had very similar data thus
proving the DIBELS tests to be reliable when testing students with special education needs.

**Limitations and Future Research**

As in many studies, this study contains standard limitations in sampling and generalizability. First, the sample size was very small. As in many cases a larger sample size would provide a better window to the extent of validity and reliability between the two assessments. Secondly, one must look at generalizability within the study. This study was conducted only on students with special needs. It cannot be related to the general population. Furthermore, 80% of the participants were of Hispanic origin, and all were drawn from the same urban area. Finally, because of the limitations in sampling and generalizability, one must consider a larger sample size with a more diverse population for future research between the Woodcock Johnson III and the DIBELS measures.

**Practice Implications**

The DIBELS measures have strong validity and reliability when assessing students with special needs. The measures are both efficient and effective when used as a formative assessment to test students who may be at risk of reading failure. The assessments may be given by teachers or other trained personnel in a brief amount of time. Special educators should embrace these practical one minute probes as a quick and useful way to assess students with disabilities. The data provided by DIBELS is easy to score and can be graphed to monitor progress throughout the year.

Using DIBELS as a formative assessment throughout the year can show what a student knows and doesn’t know. In turn, the DIBELS results can help
educators set appropriate learning goals for students. Learning goals can then guide a teacher when deciding what instruction should come next.
REFERENCES


California State University, Fresno

Non-Exclusive Distribution License
(to make your thesis available electronically via the library’s eCollections database)

By submitting this license, you (the author or copyright holder) grant to CSU, Fresno Digital Scholar the non-exclusive right to reproduce, translate (as defined in the next paragraph), and/or distribute your submission (including the abstract) worldwide in print and electronic format and in any medium, including but not limited to audio or video.

You agree that CSU, Fresno may, without changing the content, translate the submission to any medium or format for the purpose of preservation.

You also agree that the submission is your original work, and that you have the right to grant the rights contained in this license. You also represent that your submission does not, to the best of your knowledge, infringe upon anyone’s copyright.

If the submission reproduces material for which you do not hold copyright and that would not be considered fair use outside the copyright law, you represent that you have obtained the unrestricted permission of the copyright owner to grant CSU, Fresno the rights required by this license, and that such third-party material is clearly identified and acknowledged within the text or content of the submission.

If the submission is based upon work that has been sponsored or supported by an agency or organization other than California State University, Fresno, you represent that you have fulfilled any right of review or other obligations required by such contract or agreement.

California State University, Fresno will clearly identify your name as the author or owner of the submission and will not make any alteration, other than as allowed by this license, to your submission. **By typing your name and date in the fields below, you indicate your agreement to the terms of this distribution license.**

<table>
<thead>
<tr>
<th>Alicia Sandman</th>
</tr>
</thead>
<tbody>
<tr>
<td>Type full name as it appears on submission</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>7/9/2010</th>
</tr>
</thead>
<tbody>
<tr>
<td>Date</td>
</tr>
</tbody>
</table>