Fall 10-30-2017

The predictability of New Mexico's summative reading assessment by two commonly used early literacy assessments, the Dynamic Indicators of Basic Early Literacy Skills Next (DIBELS Next) and the Developmental Reading Assessment – Second Edition (DRA2)

Michelle I. Osowski
University of New Mexico - Main Campus

Michelle I. Osowski

Candidate

Individual, Family & Community Education

Department

This dissertation is approved, and it is acceptable in quality and form for publication:

Approved by the Dissertation Committee:

Jay Parkes, Ph.D., Department Chair and Professor, Department of IFCE, Chairperson

Terri Flowerday, Ph.D., Professor, Department of IFCE

Jan Armstrong, Ph.D., Professor, Department of IFCE

Thomas Dauphinee, Ph.D., Research Professor, Center for Education Policy Research

Cindy Gregory, Ph.D., Chief Statistician, New Mexico Public Education Department

The predictability of New Mexico's summative reading assessment by two commonly used early literacy assessments, the Dynamic Indicators of Basic Early Literacy Skills Next (*DIBELS Next*) and the Developmental Reading Assessment – Second Edition (DRA*2*)

BY

MICHELLE I. OSOWSKI

B.S., Chemistry, Ft. Lewis College, 1982
M.S.T., Science Teaching, New Mexico Institute of Mining and Technology, 1991
M.S., Chemistry, New Mexico Institute of Mining and Technology, 1994

DISSERTATION

Submitted in Partial Fulfillment of the
Requirements for the Degree of
Doctor of Philosophy

Educational Psychology

The University of New Mexico
Albuquerque, New Mexico

December, 2017

DEDICATION

A mi padre maravilloso, Johnnie C. Whetten.

ACKNOWLEDGMENTS

I would like to acknowledge my dissertation chair Dr. Jay Parkes. His example as my dissertation chair has reinforced my belief in the importance of committed and dedicated educators to student achievement and success—at any level. I am very grateful to the continued support and encouragement shown by the members of my Dissertation Committee. Each committee member offered unique perspectives, and have impacted my graduate career in thoughtful, caring ways too numerous to list.

This study could not have been completed without the cooperation and trust extended by the (former) Secretary of Education, Hanna Skandera, of whom I am especially indebted; and the staff within the Office of Policy, Innovation, and Measurement of the New Mexico Public Education Department.

My family has been a source of unwavering, unconditional support and encouragement. I am blessed to have my mother, Mrs. Ida Scott; step mother, Mrs. Lois Whetten; in-laws (Tom and Linda Osowski), sons (Matthew, Nicholas, and Colton); daughter-in-law (Dorinda), three precious grandchildren; and siblings to provide inspiration or focus as necessary. Most importantly, my loving husband, John, has accomplished the almost impossible task of blending the right amount of friendship with love, romance and companionship in our marriage even with “the dissertation” a constant for the past several years.

Finally, my wonderful friends. You are the family I choose, and I thank you for sharing your life with me.

The predictability of New Mexico's summative reading assessment by two commonly used early literacy assessments, the Dynamic Indicators of Basic Early Literacy Skills Next (*DIBELS Next*) and the Developmental Reading Assessment – Second Edition (DRA*2*).

by

Michelle I. Osowski

B.S., CHEMISTRY, FT. LEWIS COLLEGE, 1982
M.S.T., SCIENCE TEACHING, NEW MEXICO INSTITUTE OF MINING AND TECHNOLOGY, 1991
M.S., CHEMISTRY, NEW MEXICO INSTITUTE OF MINING AND TECHNOLOGY, 1994

DOCTOR OF PHILOSOPHY

ABSTRACT

This study will determine the degree to which first grade literacy tests predict third grade reading performance in order to judge their value as "early warning systems" for reading skills. Reading skills are fundamental to many academic outcomes, so having an early sense of how students are reading is critical for schools. The first grade reading tests being compared are the Developmental Reading Assessment-Second Edition (DRA*2*) and the Dynamic Indicators of Basic Early Literacy Skills – Next (*DIBELS Next*). This study will employ two datasets, one with *DIBELS Next* scores (N=5,456) and one with DRA*2* scores (N=2,209). Logistic regression is used to judge the predictability, and all logistic regression models are generated with the Statistical Package for Social Scientists (SPSS) version 20. The dependent variable is operationalized to be scoring proficient or not proficient on the New Mexico third grade English language arts/reading Standards Based Assessment (SBA). The independent variable is the composite score on the early literacy

assessment. Covariates are demographic characteristics (i.e., gender, racial group, English language learner status and economic disadvantaged status). For both models, the beginning-of-year composite score had a significant overall effect in predicting student proficiency on the SBA. The DRA*2* model had higher percentages of sensitivity, and positive and negative predicted values compared to the *DIBELS Next.* Conversely, the *DIBELS Next* had higher false positive and negative rates than the DRA*2*.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1

Introduction

First grade students that fail to make progress in acquiring essential reading skills and fall behind their peers could be referred to remedial or special education, and may not catch up to their peers by the third grade (Gunn, Smolkowski, Biglan, & Black, 2002). Lesnick, Goerge, Smithgall & Gwynne (2010) report that three-fourths of third grade students struggling with reading did not catch up by the ninth grade, and were four times less likely to graduate from high school compared to a student who is on grade level in reading (Hernandez, 2010). If the same struggling third grade student is identified as economically disadvantaged, i.e., receives free or reduced price lunch, the student is 13 times less likely to graduate from high school by the age of 19 (Sparks, 2011).

Students with poor reading ability experience "…substantial decreases in their self-esteem, self-concept, and motivation to learn to read" (Lesnick, Goerge, Smithgall, & Gwynne, 2010, p. 6). Fiester (2010), author of "Early Warnings: Why Reading by the End of Third Grade Matters", summarizes the concern of the decline in reading ability:

> The bottom line is that if we don't get dramatically more children on track as proficient readers, the United States will lose a growing and essential proportion of its human capital to poverty, and the price will be paid not only by individual children and families, but by the entire country (p. 7).

According to Cunningham and Stanovich (1997):

> Students who master essential reading skills in the primary grades are able to fully benefit from instruction, self-teach, and advance exponentially. Conversely, students who fall behind experience progressively more difficulty bridging the

> gap between them and their classmates, and may ultimately develop a reading disorder (p. 939).

In 2003 and augmented in 2007, the New Mexico legislature recognized the alarming decline in reading proficiency occuring in the public schools, and enacted several key pieces of reform legislation meant to "enhance and upgrade the delivery of quality education…" with a "renewed emphasis on the primary grades, recognizing especially the importance of the first grade to a child's future educational career" (22-1-1.1 A, B, E NMSA 1978). One example of the 2003 or 2007 education reform effort is the requirement for every public school district to implement early literacy assessments in kindergarten through second grade to screen and monitor progress in reading and writing (22-13-1 NMSA 1978; 6.29.1.11 B (2) NMAC). As a result, superintendents had an urgent need for a high quality, instructionally sensitive, resource-conscious assessment that accurately describes a student's progress in reading beginning with kindergarten, but most especially in first grade (22-2C-6 – 11 NMSA 1978, 6.29.1.9.E(2) NMAC, 6.75.2.8. E & L NMAC). Another provision of the reform legislation was that each superintendent is to implement a proactive system for early intervention for students who demonstrate a need for educational support (6.29.1.9 E NMAC), and the identification and improvement programs must be aligned to the results from the assessment implemented by the school district (22-2C-6 NMSA 1978). The success of any intervention model is the correct identification of children that are at risk of reading difficulty; and then providing an appropriate, targeted intervention (Compton, et al., 2010). Previous research conducted by Nelson, Benner and Gonzales (2005)

indicated that pre-reading intervention as early as kindergarten is effective for students. Coyne, Kame'enui, Simmons, and Harn (2004) found that of the 80 kindergarten students that received pre-reading intervention, 74% had notable progress through first grade attaining grade-appropriate skill levels. The New Mexico legislature has targeted first grade as a crucial year for future educational attainment. Third grade is the first academic year that all students enrolled in public schools that receive federal money must be assessed in reading with a standardized, summative assessment that meets federal guidelines (No Child Left Behind Act of 2002). Determining which early literacy assessment administered at the beginning of first grade is the better indicator of reading performance at the end of the third grade has the potential to better identify students at-risk of being poor readers. Using logistic regression as a means of predicting reading proficiency status of students has the potential to answer questions regarding what variables predict proficient reading, as well as those that do not, in an effort to closely monitor individual student progress.

This effort will probe the relationship between first and third grade reading proficiency for students in New Mexico public schools. Specifically, I will investigate the predictive ability of two different early literacy assessments that was commonly administered in New Mexico at the beginning of first grade in the 2011-2012 school year, to reading performance at the end of third grade in New Mexico during the 2013-2014 school year. The early literacy assessments are the Dynamic Indicators of Basic Early Literacy Skills – Next (*DIBELS Next*) and the Developmental Reading Assessment – Second Edition (DRA*2*). Third grade reading proficiency is measured by New Mexico's

Standards Based Assessment (SBA), a standardized, standards-based English language arts/reading assessment administered in the seventh month of a nine-month academic year.

For this study, the composite score from the beginning of year (BOY) early literacy assessment and four demographic covariates (e.g., gender, racial group, English language learner status (ELL), and poverty (e.g., economic disadvantaged status)) are used to predict third grade end-of-year (EOY) reading performance employing logistic regression as the statistical method. The New Mexico legislature requires efforts be made to close the achievement gap in order to ensure an educational system that positively impacts the principles of democracy, fairness and justice for all citizens (6.60.9 NMAC); and statute requires districts to report assessment results by racial group, gender, poverty, and English language proficiency (22-2C-5 NMSA, 1978; 22-2C-11 NMSA, 1978). Logistic regression analyses provide strong support for using composite scores from the first grade BOY early literacy assessments to predict EOY third grade reading performance. The classification rates for the predictive model improve when the BOY composite score and four demographic covariates are included in the model compared to the predictive model with only the intercept.

In the *Literature Review* section, I will begin with a summary of current early literacy predictive studies, followed by a description of the two most common early literacy assessments administered in New Mexico from 2011 through 2014. In the *Methods* section I will provide a summary of academic performance on the third grade English language arts/reading SBA by demographic characteristic for each early literacy assessment. Then, a description of how assessment data are collected, and the

psychometric properties for each early literacy assessment is summarized. Finally, I will detail the outcome and independent variables for the development of predictive models. In the *Results* section I will discuss the modeling strategy and subsequent analyses that I used to determine if the predictive models meet technical standards, and determine which of the two early literacy assessments best predicts student performance in English language arts/reading at the end of the third grade.

CHAPTER 2

Review of Related Literature

The National Institute for Early Literacy (2009) found that "…early literacy skills have a clear and consistently strong relationship with later conventional literacy skills…" (p. 5). Early literacy skills are phonemic awareness, alphabetic principal, basic phonics, word attack, accurate and fluent reading, comprehension, vocabulary and language skills (The National Early Literacy Panel, 2009; National Reading Panel, 2000) . *DIBELS Next* and DRA*2* purport to assess these skills in unique ways and provide a mechanism for teachers to identify students that are not meeting benchmark goals and then monitor the student's progress (Good, et al., 2011; Pearson Education, Inc., 2009). The characteristics of a quality early reading assessment capable of informing instruction include being psychometrically sound (i.e., demonstrated reliability and validity evidence), and able to evaluate specific skills that are amenable to intervention (i.e., phonological awareness and alphabetic skills) (Rouse & Fantuzzo, 2006). To evaluate the effectiveness of curricula, an early reading assessment must be sensitive to change over time (Rouse & Fantuzzo, 2006). Given the demands on a teacher and limited instructional time, it is imperative that an early reading assessment be administered on a large scale and in a cost effective manner (National Reading Council, 1998).

Description of the Two Early Literacy Assessments

The two early literacy assessments used in this investigation fall into a broad category of formative literacy assessments also known as "informal reading and writing inventories" (IRIs) (Burgin & Hughes, 2009). The performance data generated by IRIs

are credible if the scores are reliable and if the literacy assessments look like good instruction.

- *Developmental Reading Assessment–Second Edition (DRA2):*

The DRA*2* measures reading engagement, oral reading fluency and comprehension, and is administered twice per year i.e., beginning-of-year (BOY) and end-of-year (EOY), with the option of a middle-of-year (MOY) administration (Beaver, 2006). DRA*2* covers grade spans K-3 and 4-8. The Spanish version, EDL, (*Evaluación del desarrollo de la lectura)* was revised to assess Spanish-speaking students in kindergarten through sixth grades. The overall score on the DRA2 is an instructional reading level which is defined as the level a student can engage in teacher-instructed text (Rouse & Fantuzzo, 2006). The largest district in New Mexico implemented the DRA*2* as part of the district's assessment framework sometime before 2005. A description of the DRA*2* subtests and administration procedure is summarized in Appendix B.

Criticism of the DRA*2*, as with other IRIs, are interrater agreement concerns that stem from administration issues (e.g., choosing a book at the student's level, the teacher's familiarity with the reading passages, time and effort to master the miscue scoring system, and appropriate prompting of students by the teacher during the testing). The authors of the DRA*2* attempt to mitigate interrater and intra-rater reliability issues by utilizing scoring rubrics to score reading engagement, oral reading, and printed language concepts (Levels A, 1, 2 and 3) or comprehension (Levels 4 through 80). Another criticism is the time per student that it can take to administer and score the DRA*2*, whereas a Running Record—a formative assessment strategy from IRI-based programs–can take less than five minutes (Burgin & Hughes, 2009).

- *Dynamic Indicators of Basic Early Literacy Skills (DIBELS Next):*

DIBELS Next measures are brief, reliable measures used to assess early literacy skills (Good & Kaminski, 2011). The measures are "dynamic" in the sense that pre-reading skills are frequently evaluated by "indicators" of the basic literacy skills (Goffreda, et al., 2009). The *DIBELS Next* are standardized benchmark assessments administered three times per year with grade-level material (Good, et al., 2011).

More than 28,000 schools have used the *DIBELS* assessments (Good & Kaminski, 2016) often as part of the Reading First initiative, which may explain the popular adoption and use (Riedel, 2007). The Reading First program promoted the *DIBELS* over other assessments as the common formative reading assessment required for all federally sponsored Reading First programs. The rationale for promoting one formative reading assessment over states or districts selecting a formative reading assessment was to ensure all students were being assessed in a standardized manner with clear objectives, is sensitive to change, and could be administered in a large scale in a cost effective manner (Rouse & Fantuzzo, 2006). The results would be uniform and could then be used in program evaluation and measure the effectiveness of curricula across school districts (Olson, 2007). New Mexico adopted the *DIBELS Next* (English version) and IDEL (Spanish version) as the common-formative assessment for the Reading First program (2002-2008) (US Department of Education), the New Mexico Reads to Lead initiative of 2012 (and ongoing), and the K-3 Plus literacy project (2007 – present) (Public Education Department, 2015).

The *DIBELS Next* administration guidelines are different for schools in the K-3 Plus program compared to schools that utilize the *DIBELS Next* as part of their district's

assessment framework. Schools in the K-3 Plus program are directed to administer the BOY assessment within the first ten days of the school year, the MOY on two specific dates in January 2012, and the EOY within a 12-day window that began the end of April 2012 (New Mexico K-3 Plus, 2011). Schools using the *DIBELS Next* as part of their district's assessment framework are to administer the BOY assessment in the first through third months of the school year, the MOY in the fourth to the sixth months of the school year, and the EOY in the seventh to the ninth months of the school year (Good R. , et al., 2011). All schools in the *DIBELS Next* dataset administered the BOY assessment within the first 30 school days, whereas the MOY and EOY administration ranged up to 60 school days.

There are four criticisms of *DIBELS Next.* The first criticism is that the indicators (e.g., subtests) may not be adequate indicators of reading comprehension. If the *DIBELS* subtests are not closely related to reading comprehension, students with high *DIBELS* scores and poor comprehension could be excluded from useful interventions (Riedel, 2007). The second criticism is that some students in reading programs that use the *DIBELS Next* follow a 'stepping-stone model', e.g., mastery of certain skills is required before moving to the next skill. In the case of *DIBELS,* the stepping-stone model is reading nonsense words or pronouncing individual phonemes within words before fluency or comprehension is introduced (Riedel, 2007). The third criticism is that *DIBELS* will become a *de facto* curriculum because of the widespread use. Finally, the fourth criticism that students will believe they need to read fast because the *DIBELS* subtests are timed, which places emphasis on speed and not on comprehension or fluency

(Samuels, 2007). Appendix C describes the *DIBELS Next* measures and the administration process.

Comparing the Utility of the Two Early Literacy Assessments

Implementing an early literacy assessment as a screening tool for pre-reading skills is one of the most popular additions to many school districts' assessment plans (Betts, et al., 2008). According to Betts et al (2008),

> Screening for early literacy deficits is useful to the extent the measures are accurate, sensitive to instructional needs, responsive to the effects of interventions, valid as predictors of later reading outcomes, and fair to all groups for whom inferences will be made (p. 554).

Many decisions will be made with the early literacy data, and those decisions, along with the purpose of the assessment, are key considerations for selecting assessments (Compton, et al., 2010). When adopting instructional materials for all public schools in the state, the Instructional Materials Bureau of the PED must ensure (among other criteria) that all statewide adopted materials align to the state standards, are designed with effective pedagogy utilizing scientifically-based research, and supports accountability (22-15-8 NMSA 1978). Once the PED adopts instructional materials, then superintendents must ensure that adopting any instructional material is cost effective and a smart use of limited resources, (i.e., professional development considerations, alignment to existing core programs, and use of instructional time). It is imperative that the information gained by administering either early literacy assessment justifies the instructional time required to administer and score; and is cost effective for the state (6.75.2 NMAC).

The DRA*2* is a comprehensive package that includes one Teacher's Guide, and one set of leveled books for student assessment (23 books). Although common practice was to share a comprehensive package between two teachers, this was not a practical solution given the time requirements to administer an assessment and the PED's established assessment windows. The form to record the student's oral reading and comprehension skills that a teacher uses is (on average) three pages per leveled book, and must be printed out by the classroom teacher prior to the administration of the assessment. The total estimated administration time per student only accounts for the one-on-one teacher-student time where the student's oral reading is assessed (estimated to be a maximum of 15 minutes per student), and the teacher-student conference (estimated to be a maximum of 7 minutes per student). The total estimated administration time does not include additional student independent work or teacher analysis. The one-on-one teacher-student time is dependent upon each student's independent reading level. For instance, students in levels A-12 require no more than 10 minutes, levels 14-24 require 20 minutes, and levels 28-38 requires a maximum of 12 minutes (Pearson Education, Inc., 2009).

The *DIBELS Next* can be obtained as a printed book from Sophris, or the book can be printed at the school site. Each student book is 32 pages, and includes all necessary assessments for each assessment window for one academic year. Each component of the *DIBELS Next* assessment has a one-minute time limit. The estimated maximum time per first grade student for the BOY assessment is 5 minutes, and the MOY and EOY assessments are 8 minutes each. The total estimated administration time

does not include additional student independent work or teacher analysis (Good R. , et al., 2011).

Table 1 estimates, by early literacy assessment, the total cost of administration and time required of one teacher to administer the three assessment windows to 20 first grade students. What is assumed in the calculations is that the required pre-training and time to become familiar with the teacher and student materials is the same for either the DRA*2* or the *DIBELS Next.*

Table 1

Cost and Time Estimation to Administer the Early Literacy Assessments

Early literacy assessment	Cost		Time to administer
DRA*2*	Comprehensive kit:	$422.97	Average of 15 minutes per student for oral reading and 7 minutes for student conference, 20 students, 3 assessment windows
	Copying costs 3 pages/student, 3 assessment windows, $.05/page, 20 students	$ 9.00	
	Total cost per classroom:	**$431.97**	**22 hours**
DIBELS Next	Copying costs 32 pages/student that includes all assessment windows;, $.05/page, 20 students	**$32.00**	Average of 7 minutes per student, 20 students, 3 assessment windows **7 hours**

At first review, it appears the cost to administer the DRA*2* is approximately 14 times **more** expensive than to administer the *DIBELS Next,* and requires three times the amount of instructional time. The purchase of the comprehensive kit can be used for several years and would not be considered a classroom consumable. If comparing the consumable cost (e.g., copying) of the DRA*2* to the *DIBELS Next*, the DRA*2* appears the most cost effective, yet the most time intensive, early literacy assessment.

Current Literature Regarding Use of DIBELS Next or DRA2 to Predict Reading Performance

No published studies were located that utilized the first grade BOY *DIBELS Next* or DRA*2* instructional reading level in combination with any other measure or by itself as predictor of third grade reading ability. This could be that the BOY first grade administration of the two early literacy assessments are usually considered a benchmark, and not necessarily a summative measure. There are numerous studies utilizing the first grade EOY composite score. In one study the EOY *DIBELS Next* composite score for first grade students and responsiveness criteria collected within a Response-to-Intervention (RtI) framework were accurate predictors of reading disabilities in the beginning of third grade. The authors report the *DIBELS Next* composite score, along with other RtI measures, was a statistically significant, accurate predictor of a student being correctly identified as experiencing reading disabilities (Beach & O'Connor, n.d.). The Dynamic Measurement Group have conducted numerous studies of the *DIBELS* composite score predicting the Group Reading Assessment and Diagnostic Evaluation (GRADE) or California Standards Test (CST) total scale score. The studies were of students in the same grade, and range from 71% to 78% prediction accuracy (Powell-Smith, Good, Habedank Stewart, & Dewey, 2011).

There are considerable studies within the reading disabilities or RtI fields that demonstrate that one or more of the *DIBELS Next* subtests predicts either same grade or later grade reading comprehension. The most common *DIBELS Next* subtest used in predictive studies is oral reading fluency (ORF). *DIBELS Next* ORF is introduced at the middle-of-year (MOY) window of the first grade, and is consistently used at each

administration window through the sixth grade EOY assessment (Good R. , et al., 2011). The *DIBELS Next* ORF was significantly correlated with curriculum-based measurement ORF (Fuchs, Fuchs, Hosp, & Jenkins, 2001), and *DIBELS Next* ORF scores administered in third grade was significantly correlated to comprehension skills measured in the third grade (ranging from .73 to .80) (Barger, 2003; Shaw & Shaw, 2002; Wilson, 2005).

Other *DIBELS* subtests have mixed results when looking for a predictive relationship with summative EOY reading assessments. For instance, two studies of first grade students found no significant relationship between the *DIBELS Next* phonolgical awareness (e.g., Phoneme Segmentation Fluency (PSF)), and alphabetic principal (e.g., Nonsense Word Fluency (NWF)) to first grade reading comprehension measured by the Stanford Diagnostic Reading Test. Another study found a statistically significant correlation between the Stanford Comprehension Cluster administered in the first grade to PSF (r=.38) and NWF (r=.61) also administered in the first grade (Riedel, 2007).

The predictive ability of the DRA*2* to several states' EOY standardized summative assessments has occurred. The first study examined second and third grade DRA*2* scores to the third grade Reading Ohio Achievement Assessment (R-OAA). The second and third grade DRA*2* scores were strong predictors of third grade reading raw scores, and differentially predicted students scoring below or at or above grade level. The DRA*2* was found to be a better predictor for students scoring below grade level than students at or above grade level (Hickey, 2012). A second study looked at the second grade BOY DRA*2* reading level, comprehension and fluency scores and the relationship to the third grade Texas Assessment of Knowledge and Skills (TAKS) reading test. The three DRA*2* scores accounted for 22% of the third grade TAKS Reading test scores, and

the reading level score was found to be a significant predictor of the TAKS Reading test score (p<.0001) (Lewin, 2011).

Predictors of Early Elementary Reading Achievement

The differences in reading achievement within the groups of the four covariates is the achievement gap. The gender gap begins in kindergarten if boys' reading skills are below girls' reading skills and are not addressed. The gender gap compounds by second grade with more than twice the number of boys repeating second grade compared to girls (Entwisle, Alexander, & Olson, 2007). Under achieving boys are more likely to be identified for remedial services and expulsion, experience more difficulty transitioning to high school, and have higher dropout rates than underachieving girls (Entwisle, Alexander, & Olson, 2007). If the boys are economically disadvantaged and underachieving, the research suggests higher placement rates in the juvenile justice system than girls (Kingdon, Serbin, & Stack, 2017). To examine the impact of gender on oral reading fluency, the *DIBELS* ORF was administered in the BOY, MOY and EOY to 5,796 second grade students in a large urban public school district in North Carolina. Statistically significant differences were found between girls and boys in the BOY $t(5795) = 9.71$, d=.26; MOY $t(5795) = 10.19$, d=.27; and EOY $t(5795) = 8.89$, d = .23, suggesting that a student's gender is a reliable predictor of oral reading fluency, as measured by the *DIBELS* ORF (Wang & Algozzine, 2011).

Socio-economic status (SES), i.e., economic disadvantaged status, is a well-known predictor of a student's academic achievement (Thomson, 2010), and the correlation between SES and literacy is well-established beginning with studies conducted in the 1960s and 1970s (Buckingham J. W.). SES has been found to be a

significant mediating factor in persistently low reading scores. Feinstein and Bynner (2004) found 67% of low SES children who were in the lowest test quartile at age five remained in the lowest quartile at age 10, compared with 34% of high SES children. SES is pertinent to the Matthew effect, i.e., the 'spiral of causality' theory, as SES influences the development of emergent literacy skills (Buckinham, 2013). Phonics instruction has been shown to be beneficial to all students with a stronger effect for students from low SES, or students who begin school with low levels of phonological awareness and pre-literacy skills (Buckingham J. W.-W., 2013, p. 203). Neither *DIBELS Next* or DRA*2* assess phonologic skills. In New Mexico, the true SES of school children is not known. SES commonly includes information about a parent or guardian's education level and current employment status and pay. The proxy for SES status in this study is the economic disadvantaged status. Economic disadvantaged status is determined by participation in the free or reduced lunch program, and is reported by the Districts to the PED.

The achievement gap between English language learners (ELL) and all students on the third grade English language arts/reading SBA has increased from the 2007 to the 2014 administration. In 2007, the reading achievement gap between the number of students scoring proficient or advanced was 13.4%. In 2014, the reading achievement gap between the number of students scoring proficient or advanced was 18.1%, an increase in the reading achievement gap of 4.7% (See Appendix A). Deficits in reading achievement worsen as ELL students progress through school, especially if appropriate instructional and assessment methods do not differentiate between learning a second language or a learning disability (Gilbertson & Bluck, 2006). Research clearly shows

that for ELL students to be successful in school, they must achieve English language proficiency (Chung, 2012). Vocabulary acquisition (i.e., vocabulary size and depth of vocabulary knowledge) is the greatest contributor to learning English, significantly impacting reading and oral comprehension (Chung, 2012; Gottardo, Collins, Baciu, & Gebotys, 2008). Predictors of vocabulary knowledge in ELLs are phonological processing. There are three components of phonological processing: phonological awareness, phonological access (or rapid lexical access) and phonological working memory. Phonological access can be measured by rapid naming, and phonological working memory can be measured by pseudoword repetition (Gottardo, Collins, Baciu, & Gebotys, 2008). The first grade BOY administration of *DIBELS Next* includes two subtests that assess for phonological access (Letter Naming Fluency (LNF)) and phonological working memory, i.e., pseudoword repition (NWF). The DRA2 does not specifically assess for phonological access or phonological working memory. One study examining theoretical predictors of first grade ELL students' potential for success in second grade (operationalized as 'consistently average in second grade') found that pseudoword reading in English correctly classified 77%, and vocabulary knowledge correctly classified 88.6% (Gottardo, Collins, Baciu, & Gebotys, 2008).

Conclusion

The almost 10% decline in the number of students' scoring proficient on the third grade English language arts/reading from 2009 through 2014 strongly suggests a disconnect in what is being taught and measured in first grade reading to what is being assessed at the end of third grade (See Appendix A). The decline in the number of students scoring proficient or advanced on the third grade English language arts/reading

SBA implies that the results of the early literacy assessments are not used to effectively identify at-risk students, the intervention system is not effective, or the results (e.g., composite score) are not good predictors of the third grade summative scores.

Research Question

The following question will guide the purpose and direction of this study:

> Given a student's demographic characteristics, which early literacy assessment administered at the beginning of a student's first grade year has a better predictive ability of third grade reading performance, the Dynamic Indicators of Basic Early Literacy Skills *Next* (DIBELS *Next*) or the Developmental Reading Assessment – Second Edition (DRA*2*)?

CHAPTER 3

Methodology

The objective of this effort is to determine which of the two early reading assessments most commonly used in New Mexico, the *DIBELS Next* or the DRA*2*, best predicts student performance at the end of the third grade. The dependent variable (DV) is operationalized to be scoring proficient or not proficient on the New Mexico English language arts/reading Standards Based Assessment (SBA). The predictor is the BOY composite score of either the *DIBELS Next* or the DRA*2*. The covariates are demographic characteristics (e.g., gender, racial group, English language learner status, and economic disadvantaged status).

Participants:

The PED cooperated with this project by allowing access to the *DIBELS Next* and DRA*2* data for first graders in the 2011-2012 school year, and SBA English language arts/reading scale scores for these students as third graders in the 2013-2014 school year. The third grade SBA English language arts/reading data provides each student's demographic characteristics, i.e., gender, racial group, economic disadvantage or English language learner status. The PED-assigned nine-digit student identification number is the common variable in the *DIBELS Next,* DRA*2,* and SBA data; and used to match records between the *DIBELS Next* and SBA or the DRA*2* and SBA. The final *DIBELS Next* dataset contains 2,209 unduplicated records of students with complete SBA and *DIBELS Next* data. The final DRA*2* dataset contains 5,456 unduplicated students with complete SBA and DRA*2* data.

DRA2: Of the 7,666 first graders who were administered the DRA*2* in the 2011-2012 school year, 6,358 (82.9%) were tested in the same language as the student's home language during each testing window. Of the 6,358 students with consistent testing language, 896 (13%) student records did not match to a reading record in the 2014 SBA dataset. Each of the 5,462 remaining records had composite scores for BOY, MOY, EOY and a scale score in third grade English language arts/reading from the SBA. Six of the records reported the students as fourth graders instead of third graders. These records were eliminated from the dataset. The dataset was checked for duplicate records, and none existed. The resulting dataset is 5,456 unique student records.

The DRA*2* data represents 5,456 first grade students enrolled in a majority of first grade classes in the largest urban school district in New Mexico during the 2011-2012 school year, and 21.1% of all first grade students enrolled in New Mexico (National Alliance for Public Charter Schools, 2017). The vast majority of the students remained at the same elementary school for all three testing windows during the first grade year (N=5,275, 96.7%), and remained in the same district from first through third grade (N=5,198, 95.3%).

As third graders in 2014, the students enrolled in 192 different public schools throughout 37 school districts and 13 state charter schools. Almost 71% (N=3,855) of the students were identified as economically disadvantaged, and 21.2% (N=1,156) were identified as English language learners. Table 2 summarizes student academic performance on the third grade English language arts/reading SBA as proficient or not proficient for the students in the *DIBELS Next* or DRA*2* datasets by demographic characteristic.

DIBELS Next: During the 2011-2012 school year, 3,116 first graders across New Mexico were administered the *DIBELS Next*. Of those, 2,725 (87.5%) had composite scores for the BOY, MOY and EOY assessments. Of the 2,725 records, 2,325 (85.3%) unique records contained scores for all subscales. Each of the 2,209 remaining records had composite scores for BOY, MOY, EOY and a scale score in English language arts/reading from the third grade SBA. The dataset was checked for duplicate records, and none existed.

The *DIBELS Next* data represents 2,209 unique first grade students enrolled in 69 schools in 25 districts, and 8.6% of all first grade students enrolled in New Mexico (N=25,823) (National Alliance for Public Charter Schools, 2017). Thirty-eight (55%) of the schools voluntarily participated in the K-3 Plus reading initiative during the 2011-2012 school year (New Mexico Public Education Department, 2012).

As third graders in 2014, the students enrolled in 190 different public schools throughout 24 school districts, and 1 state charter school. The majority of the students (92.4%) enrolled in the same district at least one academic year. Almost 86% (N=1,893) of the students were identified as economically disadvantaged, and 23% (N=499) were identified as English language learners.

Table 2

Student Academic Performance on the Third Grade English Language Arts/reading SBA Described as Proficient or Not Proficient for the Students in the DIBELS Next or DRA2 Datasets by Demographic Characteristic

		Proficiency status on SBA 2014					
		Not Proficient		Proficient		Total	
Demographic Characteristic	Subvariable	N	%	N	%	N	%
Economically disadvantaged (ED) status	DRA*2*[a] – ED	2256	41	1599	29	3855	70.7
	DRA*2* – Not ED	366	7	1235	23	1601	29.3
	DIBELS Next[b] – ED	1089	49	804	36	1893	85.7
	DIBELS Next – Not ED	115	5	201	9	316	14.3
English language learner (ELL) status	DRA*2*[a] ELL	783	14	373	7	1156	21.2
	DRA*2* – Not ELL	1839	34	2461	45	4300	78.8
	DIBELS Next[b] – ELL	362	16	137	6	499	22.6
	DIBELS Next – Not ELL	842	38	868	39	1710	77.4
Racial Group	DRA*2*[a] – Asian	35	1	88	2	123	2.3
	DRA*2* - Black	89	2	66	1	155	28
	DRA*2* - Caucasian	367	7	916	17	128	23.5
	DRA*2* - Hispanic	1976	36	1658	0	363	66.6
	DRA*2* - Native American	155	3	106	2	261	4.8
	DIBELS Next[b] – Asian	2	0	8	0	10	0.5
	DIBELS Next - Black	13	1	12	1	25	1.1
	DIBELS Next - Caucasian	180	8	272	12	452	20.5
	DIBELS Next - Hispanic	712	32	558	25	1270	57.5
	DIBELS Next – Nat. American	297	13	155	7	452	20.5
Gender	DRA*2*[a] – Female	1202	22	1526	28	2728	50.0
	DRA*2* - Male	1420	26	1308	24	2728	50.0
	DIBELS Next[b] – Female	565	26	534	24	1099	49.8
	DIBELS Next - Male	639	29	471	21	1110	50.2
Percent by Proficiency status	DRA*2*	48				52	
	DIBELS Next	55				45	

[a] DRA*2* Total N = 5,456
[b] *DIBELS Next* Total N = 2,209

Table 3

The Number, Percent, Minimum, Maximum, Mean, and Standard Deviation of the Composite Score for the BOY Administration of the DIBELS Next or DRA2 Datasets by Demographic Characteristic

Demographic Characteristic	Subvariable	BOY Composite Score					
		N	%	Minimum	Maximum	Mean	Standard Deviation
Economically disadvantaged (ED) status	DRA*2*[a] – ED	3855	71	8	735	153.87	107.91
	DRA*2* – Not ED	1601	29	10	780	243.29	140.31
	DIBELS[b] – ED	1893	86	7	271	113.94	38.32
	DIBELS – Not ED	316	14	17	246	125.22	37.62
English language learner (ELL) status	DRA*2*[a] ELL	1156	21	8	520	127.06	89.33
	DRA*2* – Not ELL	4300	79	8	780	194.37	129.48
	DIBELS[b] – ELL	499	23	0	223	99.68	36.77
	DIBELS – Not ELL	1710	77	6	271	120.19	37.65
Racial Group	DRA*2*[a] – Asian	123	2	19	375	254.65	163.32
	DRA*2* - Black	155	3	10	560	161.61	109.00
	DRA*2* - Cauc.	1283	24	8	780	225.15	142.38
	DRA*2* - Hisp.	3634	67	8	735	164.11	114.08
	DRA*2* - Nat. Am	261	5	9	552	157.26	101.11
	DIBELS[b] – Asian	10	0	78	218	143.10	44.38
	DIBELS - Black	25	1	79	224	120.56	32.15
	DIBELS - Cauc.	452	20	6	246	121.54	35.53
	DIBELS - Hisp.	1270	57	0	271	114.31	39.04
	DIBELS – Nat. Am	452	20	15	259	112.21	38.836
Gender	DRA*2*[a] – Female	2728	50	9	780	189.51	128.18
	DRA*2* - Male	2728	50	8	780	170.71	121.31
	DIBELS[b] – Fem	1099	50	6	266	120.01	36.76
	DIBELS - Male	1110	50	0	271	111.15	39.52

[a] DRA*2* Total N = 5,456

[b] *DIBELS=DIBELS Next*; Total N = 2,209

Measures and Covariates

Training to Administer Early Literacy Assessments A description of the administration procedures and measures of each early literacy assessment is provided in Appendix B (DRA*2*) and C (*DIBELS Next*).

Data Collection: The data collection procedures are unique to each early literacy assessment. For schools in the K-3 Plus program, the *DIBELS Next* is accessed using handheld 'wireless' devices via Amplify's mCLASS:DIBELS Next mobile version (New Mexico K-3 Plus, 2011). Student scores are immediately calculated by measure (e.g., 'subtest'), and stored. mCLASS:DIBELS Next contains data provided by each participating district, i.e., state student identification number, full name of the student, date of birth, grade, district and school name, and name of the student's teacher. Schools not in the K-3 Plus program could use handheld devices or paper copies, and upload the data to the mCLASS:DIBELS database. The *DIBELS Next* scores are exported to the PED via an ftp protocol, and the PED cross-matched the data to the PED student database to ensure accuracy. Approximately 97% of the data in the mCLASS:DIBELS database matched to existing records in the PED student information system.

The DRA*2* requires a paper copy of the text a student will read, and is used to capture student mistakes and then calculate a score for the student. The district developed a DRA*2* database that included each teacher's roster of students. The teacher enters the subtest scores for each student as soon as possible following the administration. The district transferred the DRA*2* scores to the PED via an ftp protocol, and the PED cross-matched the data to the PED student database to ensure accuracy. Approximately 94% of the data in the district database matched to existing records in the PED student

information system. ***Psychometric Properties of the Early Literacy Assessments and SBA****:*

All assessments used in this study met technical standards for reliability and validity.

- *Reliability:* Three common forms of reliability for the *DIBELS Next* are reviewed: test-retest, alternate-form, and inter-rater reliability (Goffreda & DiPerna, 2010). Table D1 in Appendix D is a replication of Table 5.17 (Summary Table of Reliability Estimates for *DIBELS Measures*). The Dynamic Measurement Group used the standards for reliability established by Salvia, Ysseldyke, & Bolt (2007). The minimum standard reported for groups of individuals is .60. All reliability coefficients exceeded the .60 standard (Good, et al., 2011). Four common forms of reliability evidence were reviewed for the DRA*2*: internal consistency, parallel equivalency, test-retest, and inter-rater (Pearson Education, Inc., 2009). Table E1 of Appendix E is an excerpt of Tables 3, 5, 6, and 7 from the DRA*2* K-8 Technical Manual (Second Edition), and summarizes the test-retest reliability (correlation coefficient), inter-rater (likelihood that two randomly selected raters were in exact agreement on the DRA*2* score), and rater-expert (percent agreement between the expert and the non-expert scores). The test-retest correlation coefficients indicate high reliabilities. The inter-rater reliabilities of two randomly selected raters for fluency were in exact agreement 66% of the time, and 72% of the time for comprehension. The percent agreement between the expert and non-expert scores were 79% of the time for fluency and 89% of the time for comprehension. Table E2 of Appendix E is a replication of Table 3 from the DRA*2* K-8 Technical Manual (Second Edition) and summarizes the internal consistency reliability for oral fluency

and comprehension at each level of book. The measures show high-moderate to high reliabilities (Pearson Education, Inc., 2009).

The SBA relied on Cronbach's *a* (alpha) statistic to assess reliability of the 2012-2013 SBA. The reliability of the third grade reading was determined from 23,635 student records and is considered high (α=.85, SEM=3.05) (Measured Progress, 2014).

- *Validity:* A correlation of the *DIBELS Next* first grade BOY and EOY subscale and composite scores to the total test raw scores of the first grade GRADE provides predictive criterion-related validity evidence. Discriminant validity was established by comparing the means of the *DIBELS Next* composite scores for students that scored in two performance levels on the GRADE total test (e.g., below the 40th percentile or at or above the 40th percentile on the GRADE's national norms). A between-groups *t*-test of the difference in means for each grade yielded significant results, and Cohen's *d* is considered large (Good R. H., et al., 2011). Table D2 in Appendix D is an excerpt of Table 6.3 and summarizes the statistics of the predictive validity evidence. Table D3 in Appendix D is an excerpt of Table 6.19 and summarizes the statistics of the discriminate validity evidence.

Three types of validity are reviewed for the DRA*2*: content-related, criterion-related, and construct validity. Pearson reports that content validity is "built into the DRA*2* assessment during the development process" (Pearson Education, Inc., 2009, p. 35), and is confirmed through 66 teacher ratings on the extent the DRA*2* measures different aspects of student reading performance (Pearson Education, Inc., 2009).

The DRA*2* comprehension and fluency scores were correlated with four well-

known tests of reading comprehension and fluency, and demonstrated a large to very large relationship. Additionally, teachers were asked to rate each student's reading ability on a 5-point scale. Those scores were correlated to the student's DRA*2* composite scores, and demonstrated a moderate to large relationship (Pearson Education, Inc., 2009). The two constructs of comprehension and fluency emerged through Principal Components Analysis. The two factors cumulatively accounted for 51 to 62% of the variance (depending upon level of DRA*2*). An exploratory factor analysis estimated with maximum likelihood confirmed the two constructs. Table E3 in Appendix E summarizes the measures used to establish criterion validity for the DRA*2*.

Measured Progress relied on multiple aspects of validity to describe the overall validity of the SBA. The aspects include content, response process, internal structure (i.e., classical item statistics, differential item functioning analysis, dimensionality analysis, and item response theory (IRT) parameters and procedures), and relationship to other variables (Measured Progress, 2014).

More detailed information about the reliability and validity for the *DIBELS Next* is summarized in the *DIBELS Next Technical Adequacy Brief* (Dewey, Powell-Smith, Good, & Kaminski, 2015), the DRA*2* in the *DRA2 Technical Manual* (Pearson Education, Inc., 2009), and for the SBA in the *New Mexico Standards Based Assessment 2013-14 Technical Report* (Measured Progress, 2014).

Primary Outcome Variables and Covariates: The outcome variable, SBA_Prof, is dichotomous and indicates if the student scored proficient ('1') or not proficient ('0') on the third grade English language arts/reading SBA. The predictor variable is the BOY composite score and is a continuous variable. The four covariates are categorical variables (e.g., gender, racial group, economic disadvantaged status, and English language learner status):

- *Composite score*. The composite score is a continuous variable determined from the BOY administration. DRA*2* range from 8 to 780. The *DIBELS Next* composite score is a summation of multiple *DIBELS Next* measures and provides the best overall estimate of a student's skills (Good R. & Kaminski, R.A., 2011). A derived composite score for the DRA*2* is a continuous, non-negative whole number that utilizes all subscale scores from the BOY administration. The level of book was recoded from a nominal to an ordinal scale to eliminate the alpha designation as well as the inconsistent ranges (e.g., A, 1, 2, 3, 4, 6, 8, 10, 12, 14, 16, 18, 20, 24, 28, 30, 34, 38, 40, 50, 60, 70, 80 recoded to 1, 2, 3,….23). The composite score is the sum of the subscale scores (e.g., reading engagement, oral reading fluency, and comprehension) multiplied by the recoded book level (see equation 2):

Composite Score =

*[(Reading Engagement + Oral Reading Fluency + Comprehension Total)]*Book level*

(Equation 2)

- *Racial group.* Racial group is parent or guardian reported and is represented with five categories (1=Asian, 2=Black, 3=Hispanic, 4=Native American, 5=Caucasian). Caucasian is the reference group.

- *Gender.* Gender is parent or guardian reported. Gender is represented with two categories (0=male, 1=female). Male is the reference group.
- *Economic disadvantaged (ED) status.* The proxy for ED status is participation in the free or reduced lunch program, and is reported by the Districts to the PED. It is listed in the SBA dataset. ED status is represented with two categories (0=not ED, 1=ED). Not ED is the reference group.
- *English language learner (ELL) status.* ELL status is represented with two categories (0=not identified as an ELL, 1= identified as ELL). ELL status is reported by the Districts to the PED, and is listed in the SBA dataset. ELL status does not distinguish between students that are bilingual and not fluent enough in English to pass the language screening assessment; or students that do not speak, read, or write in English at any level. Not ELL is the reference group.

This effort will not use several collected measures. Specifically, the composite score for the MOY and EOY administrations, and numerous demographic characteristics listed in the SBA file. The MOY and EOY administrations can vary by as much as 60 days, which has the potential to influence the composite score and skew the results. Additional demographic characteristics includes, but are not limited to, if the student received Title I services, was classified as a migrant or homeless, or identified as gifted.

CHAPTER 4

Results

There are three goals for this project:

1. Develop a technically defensible predictive model for each of the early literacy assessments;
2. determine which early literacy assessment has better predictability of third grade reading performance; and based on the predictive model,
3. compare and contrast the relationships between the predictors and dependent variables.

The prediction models of the *DIBELS Next* or DRA*2* to the NM third grade English language arts/reading SBA incorporates three premises:

1. After identifying and removing records that perform as an outlier, exhibit high leverage, or strong influence in either early literacy dataset, the *DIBELS Next* and DRA*2* datasets are clean, and
2. The data in the SBA data set are accurate; and
3. The test administration protocols were followed for both the *DIBELS Next* and DRA*2* assessments.

All logistic regression modeling and statistical methods are generated with the Statistical Package for Social Scientists (SPSS) version 20, and will utilize unconditional maximum likelihood for model estimation (UCLA: Statistical Consulting Group, n.d.). Two immediate concerns were overfitting a model and theorizing a model that will successfully converge. Overfitting the model is not of concern for either early literacy assessment as the ratio of outcomes per IV for the *DIBELS Next* is approximately 11 to 1,

the ratio for the DRA*2* is 27 to 1, and there are at least 50 cases per parameter (Walsh, 1987). Convergence of the model occurred within five iterations.

Assumptions

Preliminary analyses were conducted to confirm that logistic regression assumptions were met as is required (Stoltzfus, 2011). Specifically:

- *Cases are independent.* Independence of records (and errors) requires no duplicate records exist within each dataset to avoid repeated measures or other correlated outcomes (Stoltzfus, 2011). This assumption was met in the construction of the datasets (described in the *Methods* section).
- *The IVs are measured without error.* Both early literacy assessment datasets have normally distributed standardized and deviance residuals with 95% of cases between ± 2.00, and 99% of the cases between ± 2.50 (Menard, 2001). Table 4 summarizes by dataset the number and percent of cases that have standardized and deviance residuals within 95% and 99% of total cases.

Table 4

The Number of Cases with Standardized and Deviance Residuals Within 95% and 99% of the DIBELS Next and DRA2 Datasets

		Standardized Residual		Deviance Residual	
Dataset	N (Total)	N (%) within 95%	N (%) within 99%	N (%) within 95%	N (%) within 99%
DIBELS Next	2209	2151 (97.4%)	2183 (98.8%)	2183 (98.8%)	2206 (99.9%
DRA*2*	5456	5268 (96.6%)	5384 (98.7%)	5386 (98.7%)	5445 (99.8%)

The datasets do not contain strong influential outliers. Appendices F (DRA*2*) and G (*DIBELS Next*) contain four graphs: Figure 1: The predicted probabilities vs.

standardized residuals, Figure 2: The predicted probabilties vs. deviance residuals, Figure 3: The predicted probabilities vs. delta Chi-squared, and Figure 4: The predicted probabilities vs. delta deviance. For the DRA*2* dataset, three records consistently stood out and were eliminated: DRA_IDs 2842, 3277, and 2867 leaving a final DRA*2* dataset of N=5,453.

For the *DIBELS Next* dataset, twelve records consistently stood out but only three records had conflicting data: DIBNext_IDs 96, 1739, and 22, and were eliminated from the dataset resulting in a final *DIBELS Next* dataset of N=2,206.

- *The IVs are not linear combinations of each other:*
    - *Continuous IVs:* Testing for linearity of the continuous IV (e.g., composite score) to its logit was accomplished with the Box-Tidwell test. The interaction between the BOY DRA*2* composite score and the composite score logit is not significant (β = -.002, SE = .001, p = .347), and the change in the Likelihood Ratio statistic with the interaction is significant ((χ^2 (df=8, N=5,453) =1808.048, p =.000), both indicating the assumption for absence of multicollinearity in the continuous IV has been met. The interaction between the BOY *DIBELS Next* composite score and the composite score logit is not significant (β = .008, SE = .006, p = 204), and the change in the Likelihood Ratio statistic with the interaction is significant ((χ^2 (df=9, N=2,206) =428.580, p =.000), both indicating the assumption for absence of multicollinearity in the continuous IV has been met.

- *Categorical IVs:* Multicollinearity among the categorical IVs is tested by determining the tolerance and Variance Inflation Factor (VIF). Appendix H Figure H1 summarizes the VIF and tolerance for all IVs as DVs against other IVs for the DRA*2* dataset, and Appendix H Figure H2 summarizes the VIF and tolerance for all IVs as DVs against other IVs for the *DIBELS Next* dataset. No VIF values met or exceed the maximum acceptable value of 10 in either the DRA*2* or *DIBELS Next* (Cohen, Cohen, West, & Aiken).

- *The model is correctly specified*: Overdispersion and underdispersion are statements of the variance, and can indicate an omission of an important predictor. Dispersion is presented in Tables 5 and 6 as a component to establishing the internal validity of both models derived from each dataset. The data indicate that the proposed models for both early literacy assessments are correctly specified, and an important predictor has not been omitted.

Internal Validity

Internal validity was determined via the 'holdout method' (e.g., splitting the dataset into two equivalent samples). The first half of the sample was the training sample used to create the logistic regression model, and the second half of the sample is the test sample and used to confirm the model (Stoltzfus, 2011). Each dataset was split into two separate and equivalent samples after the variables have been sorted as follows:

- Composite score (descending)
- District number (ascending)
- School number (ascending)
- Gender (ascending)
- Racial group (ascending)

- Economic disadvantaged status (ascending)
- English language learner status (ascending)

The training sample used step-wise (forward) with backward elimination and likelihood statistics for model building. The test sample utilized the block method (Stoltzfus, 2011). A summary of the composition of each IV in the DRA*2* training and testing datasets can be found in Appendix I Table I1, and for the *DIBELS Next* training and testing datasets in Appendix J Table J1.

Goodness-of-Fit

Goodness-of-fit measures indicate how well the model provides an explanation or prediction and accounts for the variations within the DV (Menard, 2001; Pampel, 2000). This effort compared results from two inferential (e.g., Likelihood Ratio Test and the Hosmer-Lemeshow), and four descriptive tests (e.g., McFadden, Cox and Snell, Nagelkerke, and correlation). Table 5 summarizes the percentage correct, inferential, and descriptive model fitting statistics (degrees of freedom and significance in parenthesis) for the DRA*2* training and test datasets. All IVs are included in the internal validity checks.

Table 5

The Percentage Correct, Dispersion, Inferential and Descriptive Model Fitting Statistics for the DRA2 Training and Test Datasets

DRA*2*	Training	Test
Number in Subsample	2,726	2,727
Percentage Correct	73.0%	73%
Dispersion	.980	.988
Inferential		
Likelihood Ratio	880.349 (*df*=8, sig.=.000)	944.984 (*df*=8, sig.=.000)
Hosmer-Lemeshow	5.681 (*df*=8, sig.=.683)	12.334 (*df*=8, sig.=.137)
Descriptive		
McFadden[a]	.276	.281
Cox and Snell	.276	.293
Nagelkerke	.368	.391
Correlation	.540 (sig. =.000)	.541 (sig. = .000)

The Likelihood Ratio inferential test for the DRA*2* training dataset (χ^2 (*df*=8, *N*=2,726) =880.349, *p* =.000) and test dataset (χ^2 (*df*=8, *N*=2,727) =944.984, *p* =.000) indicate the full model is a good fit of the data. The Hosmer-Lemeshow tests the null hypothesis that predictions made by the model fit group membership, and a nonsignificant Chi-square indicates the test dataset (χ^2 (*df*=8, *N*=2,726) =5.681, *p* = .683) and training dataset (χ^2 (*df*=8, *N*=2,727) =12.334, *p* = .137) fit the model well. The Cox and Snell for the training dataset (.276) is within 10% of the test dataset (.293), and the same holds true for the Nagelkerke for the training dataset (.368) and test dataset (.391) indicating similar goodness of fit conclusions. The McFadden metric indicates an association between the IVs and DV, and is within .005 between the training (.276) and test (.281) datasets. The correlation between the predicted probability and proficiency on the SBA explains 54% of the variation.

Both internal validity models overwhelmingly suggest that all covariates are statistically significant contributors to the overall model.

Table 6 summarize the percentage correct, inferential and descriptive model fitting statistics (degrees of freedom and significance in parenthesis) for the *DIBELS Next* training and test datasets. All IVs are included in the internal validity models.

Table 6

The Percentage Correct, Dispersion, Inferential and Descriptive Model Fitting Information for the DIBELS Next Training and Test Datasets

DIBELS Next	Training	Test
Percentage Correct	67.4	68.8
Dispersion	1.15	1.13
Inferential		
Likelihood Ratio	428.170 (*df*=8, sig.=.000)	222.322 (*df*=8, sig.=.000)
Hosmer-Lemeshow	4.584 (*df*=8, sig.=.801)	5.839 (*df*=8, sig. = .665)
Descriptive		
McFadden[a]	.244	.239
Cox and Snell	.171	.183
Nagelkerke	.229	.244
Correlation	.428 (sig=.000)	.428 (sig.=.000)

The Likelihood Ratio inferential test for the *DIBELS Next* training dataset (χ^2 (*df*=8, *N*=1,104) =428.170, *p* =.000) and test dataset (χ^2 (*df*=8, *N*=1,102) =222.322, *p* =.000) indicate the full model is a good fit of the data. The Hosmer-Lemeshow test is nonsignificant indicating the test dataset (χ^2 (*df*=8, *N*=1,102) =4.584, *p* = .801) and training dataset (χ^2 (*df*=8, *N*=1,102) =5.839, *p* = .665) fit the model well. The Cox and Snell for the training dataset (.171) is within 10% of the test dataset (.183), and the same holds true for the Nagelkerke for the training dataset (.229) and test dataset (.244) indicating similar goodness of fit conclusions. The McFadden metric indicates an

association between the IVs and DV, and is within .005 between the training dataset (.244) and test dataset (.239). The correlation between the predicted probability and proficiency on the SBA explains 42.8% of the variation for both the test and training datasets.

The McFadden goodness-of-fit statistic is comparable across models that may have different IVs and the same DV (Peng & So, 2002). The McFadden statistic for the DRA*2* training (.276) and test (.281) are higher than the *DIBELS Next* training (.244) and test (.239) indicating more variance is explained in the DRA*2* model than the *DIBELS Next* model.

The internal validity models differ in what covariates are statistically significant contributors to the model. The test validity model with stepwise (forward) entry suggests that gender and racial group are *not* statistically significant contributors to the overall model. The training validity model with block entry suggests that gender is not a statistically significant contributor, as well two categories of the racial group covariate (e.g., Asian and Black).*Overall Model Assessment*

The overall model will utilize the block method (Stoltzfus, 2011), and all covariates were included in the models for both early literacy datasets.

- *DRA2:* The DRA*2* overall, final model with just the constant has a 52% overall success rate in classification. The Wald test is significant (χ^2 (8.473, *df*=1, sig.=.004)) indicating the constant is not zero. Adding the IVs significantly improved the model (χ^2 (*df*=8, *N*=5,453) = 1822.895, *p* = .000) suggesting the full model explains more of the variance in the outcome than the null model. The overall success rate in classification improved to 73.1%, and the non-significant Hosmer-Lemeshow test indicate the data fit the model

well (χ^2 (*df*=8, *N*=5,453) =11.446, *p* =.178). The parameter coefficient for the Asian racial group was not statistically significant and the confidence interval included zero. This finding was re-confirmed using two different logistic regression routines available in SPSS version 20 (e.g., NOMREG and generalized linear regression). Therefore, a parameter coefficient for Asian will not be evaluated, and will *not* be included in the final logistic regression prediction equation (see the discussion in the subsequent section, *Examining the Contribution of Individual Predictors*). The logistic regression prediction equation for the DRA*2*, where *p* is the probability of being proficient on the third grade English language arts/reading NM SBA, is:

$$Log\left(\frac{p}{1-p}\right) = -.695 + .010 * CompScore - .899 * EconDis - .333 * ELL - .641 * Black - .427 * Hispanic - .706 * Nat.American + .274 * Gender \quad \text{(Eq. 3)}$$

- The *DIBELS Next* overall, final model with just the constant has a 54.4% overall success rate in classification. The Wald test is significant (χ^2 (17.368, *df*=1, sig.=.000)), indicating the constant is not zero. Adding the IVs significantly improved the model (χ^2 (*df*=8, *N*=2,206) = 428.170, *p* = .000) suggesting the full model explains more of the variance in the outcome than the null model. The overall success rate in classification improved to 67.7%, and the non-significant Hosmer-Lemeshow test indicate the data fit the model well (χ^2 (*df*=8, *N*=2,206) =12.767, *p* =.120). The parameter coefficients for gender, Asian and Black racial group identities were not statistically significant and the confidence intervals included zero. This finding was re-confirmed using two different logistic regression routines available in SPSS

version 20 (e.g., NOMREG and generalized linear regression). Therefore, parameter coefficients for Asian, Black and gender (overall) will *not* be included in the final logistic regression prediction equation for *DIBELS Next* (see the discussion in the subsequent section, *Examining the Contribution of Individual Predictors*). The logistic regression prediction equation for the *DIBELS Next*, where p is the probability of being proficient on the third grade English language arts/reading NM SBA, is:

$$Log\left(\frac{p}{1-p}\right) = -1.934 + .022 * CompScore - .514 * EconDis - .518 * ELL - .356 * Hispanic - .712 * Nat.American \quad \text{(Eq. 4)}$$

Best Predictability of Third Grade Reading Proficiency

The second goal of this study is to determine which early literacy assessment has better predictability of third grade reading performance. Correct predictions, error rates, and the overall odds ratio for the overall, final DRA*2* and *DIBELS Next* are summarized in Table 7. Correct predictions include (Peng & So, 2002):

- Sensitivity AKA 'true positives' (i.e., percentage of records that were observed to be proficient which were correctly predicted to be proficient by the model),
- specificity AKA 'true negatives' (i.e., percentage of records that were not observed to be proficient which were correctly predicted as not proficient by the model),
- positive predictive values (the percentage of correctly predicted cases of being proficient compared to the total number of cases predicted as having the characteristic), and

- negative predictive values (the percentage of correctly predicted cases of not being proficient compared to the total number of cases predicted as not being proficient).

Error rates include:

- False positive rate (i.e., predicting that a record would be proficient when it actually was not), or
- false negative rate (i.e., predicting that a record would not be proficient when it actually was).

Table 7

The Summary of the Predicted Group Discrimination Metrics for the DRA2 and DIBELS Next Models

Predicted group discrimination	DRA*2*	*DIBELS Next*
Correct Predictions		
Sensitivity	72.7%	59.4%
Specificity	73.3%	74.6%
Positive predicted values	74.7%	66.2%
Negative predicted values	71.3%	68.7%
Error Rates		
False positives	25.4%	33.8%
False negatives	28.7%	31.8%
Odds Ratio	7.3	4.3

The DRA*2* model had higher percentages of sensitivity, and positive and negative predicted values compared to the *DIBELS Next.* Conversely, the *DIBELS Next* had higher false positive and negative rates than the DRA*2*.

Table 8 summarizes the percent of accurately predicted group membership (e.g., proficient or not proficient) by SBA proficiency status (e.g., not proficient or proficient) by covariate for the DRA*2* or *DIBELS Next*.

Table 8

The summary of the percent of accurately predicted group membership (e.g., proficient or not proficient) by SBA proficiency status (e.g., not proficient or proficient) by categorical covariate variable for the DRA2 or DIBELS Next.

Covariate		SBA Proficiency Status	Correctly Predicted (%)	
			DRA2	*DIBELS Next*
Economic Disadvantaged (ED) Status	Not ED	Not proficient	34.2	45.6
		Proficient	94	87.1
	ED	Not proficient	81.1	78.6
		Proficient	54.4	51.1
	Total	Not proficient	74.6	75.4
		Proficient	71.7	58.3
English language learner (ELL) Status	Not ELL	Not proficient	68.6	67
		Proficient	77	65.1
	ELL	Not proficient	88.5	95
		Proficient	36.2	15.3
	Total	Not proficient	74.6	75.4
		Proficient	71.7	58.3
Racial Group	Asian	Not proficient	48.6	0
		Proficient	85.2	100
	Black	Not proficient	85.4	84.6
		Proficient	59.1	58.3
	Hispanic	Not proficient	78.7	77.6
		Proficient	61.8	52.2
	Nat. American	Not proficient	84.5	88.9
		Proficient	60.4	32.9
	Caucasian	Not proficient	47.9	45
		Proficient	90.4	84.2
	Total	Not proficient	74.6	75.4
		Proficient	71.7	58.3
Gender	Female	Not proficient	68.4	71.2
		Proficient	75.6	64.6
	Male	Not proficient	79.8	79.2
		Proficient	67	51.2
	Total	Not proficient	74.6	75.4
		Proficient	71.7	58.3

The DRA*2* model was markedly more accurate in classifying proficient (82%) and not proficient (64%) groups of students. In comparison, The *DIBELS Next* model classified proficient (18%) and not proficient (36%) groups of students much less frequently.

Measures of predictive efficiency: Two statistics guide the determination of a model's predictive efficiency: lambda-p (λ_p) and tau-p (τ_p) (Menard, 2001). Table 9 summarizes the λ_p, τ_p , binomial statistic (*d*), and significance for *d.*

Table 9

Indices of Predictive Efficiency for the DRA2 and DIBELS Next Models

Early Literacy Assessment	*Lambda-p* (λ_p)	*Binomial statistic d*	*Tau-p* (τ_p)	*Binomial statistic d*
DRA*2*[a]	.482	36.889 (p<.0001)[c]	.999	36.889 (p<.0001)[c]
DIBELS Next[b]	.291	12.483 (p<.0001)[c]	.999	12.483 (p<.0001)[c]

[a] *N*=5,453
[b] *N*=2,206
[c] http://stattrek.com/online-calculator/binomial.aspx

The λ_p for the DRA*2* indicates a strong reduction in the error of prediction (.482), and the τ_p indicates the model almost completely reduces the error of classification of cases (.999). The binomial *d* is the same for both λ_p and τ_p (50% error expected for both): *d*=36.889 with statistical significance p<.0001. The λ_p for the *DIBELS Next* indicates no more than a weak relationship between the observed and predicted classification of cases (.291), and the τ_p indicates the model almost completely reduces the error of classification of cases (.999). The binomial *d* is the same for both λ_p and τ_p (50% error expected for both): *d*=12.483 with statistical significance $p < .0001$.

The ROC curve for the DRA*2* dataset suggests that in order for the DRA*2* model to correctly classify 80% of first grade students that were administered the DRA*2* at the BOY, about 38% will be misclassified. The DRA*2* overlay plot of sensitivity and specificity versus predicted probability, the optimal cutoff of .49 yields approximately 72% correct classification for both groups.

The ROC curve for the *DIBELS Next* dataset suggests that in order for the *DIBELS Next* model to correctly classify 78% of first grade students that were administered the *DIBELS Next* at the beginning of the year, about 42% will be misclassified. The *DIBELS Next* overlay plot of sensitivity and specificity versus predicted probability. the optimal cutoff of .599 yields approximately 41% correct classification for both groups.

Odds Ratios

The third goal of this study is to obtain valid estimates for the IV-DV relationship for each of the early literacy assessments. The very similar odds ratios for the BOY composite scores (OR_{DRA2} = 1.010, $OR_{DIBELS\ Next}$ = 1.023) indicate the DRA*2* contrived composite score was a good aggregate of all dimensions contained within the scoring rubric. The odds ratios illustrating the effect of the BOY composite score, gender, racial group, economic disadvantaged or English language learner status on the performance of the third grade English language arts/reading SBA for the DRA*2* and *DIBELS Next* models are summarized in Table 10.

Table 10

The Odds Ratios of the Independent Variables in Predicting Proficiency on the SBA for the DRA2 and DIBELS Next Logistic Regression Models

Independent Variable	DRA*2*	*DIBELS Next*
Constant	.499	.145
Composite score	1.010	1.023
Gender (Male is reference)	1.315	Not significant
Racial group (Caucasian is reference)		
Asian	Not significant	Not significant
Black	.527	Not significant
Hispanic	.652	.701
Nat. Am	.493	.491
Economic Disadvantaged Status (not Economic Disadvantaged is reference)	.411	.598
English Language Learner (ELL) Status (not ELL is reference)	.717	.596

DRA*2*: Looking first at the log odds of the BOY composite score, there is a significant overall effect (*Wald*=749.603, *df*=1, *p*=.000) indicating a higher BOY composite score increases the odds of being proficient on the English language arts/reading SBA. The effect of gender is also significant and positive; girls are 32% more likely to be proficient than boys on the English language arts/reading SBA. The beta coefficients for all ethnicities are significant and negative: Specifically, Black, Hispanic, and Native American students are 47%, 35%, and 50% (respectively) *less* likely than Caucasian students to be proficient on the English language arts/reading SBA. If a student is identified as an English language learner, that student is 28% *less* likely to be proficient than a non-English language learner on the English language arts/reading SBA. The logistic regression equation predicts that a student scoring a BOY composite score of 131 or higher on the DRA*2* would be proficient. Approximately 78 per cent (N=1721) of the students scored a BOY composite score of 131 or higher.

Approximately 49 per cent (N=2693) of the students scored 131 or higher on the BOY administration of the DRA*2*. A composite score of 131 can be achieved with a level 3 book, usually considered a kindergarten level book.

DIBELS Next: The odds ratio for the BOY composite score has a significant overall effect (*Wald*=223.716, *df*=1, *p*=.000) indicating a higher BOY composite score increases the odds of being proficient on the English language arts/reading SBA. The log odds for Asian or Black students are not statistically different than Caucasian students and cannot be interpreted. The beta coefficients for Hispanic and Native American students are significant and negative: Specifically, Hispanic, and Native American students are 30%, and 51% (respectively) *less* likely than Caucasian students to be proficient on the English language arts/reading SBA. If a student is identified as an English language learner, that student is 40% *less* likely to be proficient than a non-English language learner on the English language arts/reading SBA. The logistic regression equation predicts that a student scoring a BOY composite score of 88 would be proficient. The recommended composite score cut point for students considered to be at risk and likely to need strategic support is 97 (Good R. , et al., 2011). A cut score of 97 accounts for 70 per cent (N=1542), a difference of 179 students.

CHAPTER 5

Discussion

The purpose of this study was to determine which early literacy assessment commonly administered in New Mexico at the beginning of the first grade year has better predictive ability of the end-of-year third grade English language arts/reading assessment. The major findings are:

1. The BOY composite score for either early literacy assessment proved to be one of the strongest predictors of a student's proficiency on the summative third grade English language arts/reading assessment. The *DIBELS Next* has incorporated a composite score, and has used the composite score in several studies predicting end-of-grade summative performance. The DRA*2* does not have a composite score as part of the data analysis for each student, and one recommendation is for the publishers of the DRA*2*, or any state or local educational agency using the DRA*2*, to derive one. Composite scores have great utility when communicating results to all stakeholders within a school, are sensitive to individual changes, and are relatively cost effective to calculate
2. Both the DRA*2* and *DIBELS Next* are screening tools that claim to reliably detect deficits in reading skills or strategies, yet the *DIBELS Next* was unable to discern gender differences of the students that comprised the *DIBELS Next* datasets. This is alarming, as the longitudinal data of the percent of boys scoring proficient or advanced on the third grade English language arts/reading assessment is markedly lower than the percent of girls.

3. Both DRA2 and *DIBELS Next* have additional progress monitoring activities delivered in supplemental instructional periods to address gaps in skills or strategies to move the student from 'at risk' to 'on track'. Assuming that the respective progress monitoring protocol was effective, the false negative rate (i.e., predicting that a record would not be proficient when it actually was) should be higher than the false positive rate (i.e., predicting that a record would be proficient when it actually was not). However, for a screening measure to be effective, the cut-points to demarcate at risk or no risk must yield a high percentage of true positives with a balance of false positives (Compton, et al., 2010). The ratio of true positives to false positives for the DRA*2* is 2.86: about 3 students being correctly identified as proficient to every 1 student predicted to scoring proficient but actually scoring not proficient. This is compared to 1.76 for the *DIBELS Next* (less than 2 students being correctly identified as proficient to every 1 student predicted to scoring proficient but actually scoring not proficient). Neither the *DIBELS Next* nor the DRA*2* meet the sensitivity (.90) or specificity (.80) threshold (Compton, et al., 2010).

Importance of the Findings

Both the *DIBELS Next* and the DRA*2* are early literacy batteries comprised of multiple measures generally recognized as being accurate enough to identify individual children who are at risk of being a poor reader (Snow, Burns, & Griffin). The *DIBELS Next* was adopted by the PED and the DRA*2* by a district as the required universal screening measure to determine the academic levels of proficiency for each early elementary student in accordance with administrative code (6.29.1.9.E (1)). In this study,

both the DRA*2* and the *DIBELS Next* exhibited false positive rates above the acceptable range of less than 20% (Compton, et al., 2010). High numbers of false positives actually burden schools and districts to provide early intervention to more students than necessary (Compton, et al., 2010). Employing logistic regression as a means of studying the predictive ability of an early literacy assessment to a subsequent summative assessment is a reasonable, efficient technique with many benefits. The first benefit is to accurately identify at-risk students, and the second is to establish accurate demarcation points based on student performance in New Mexico.

Threats to Internal Validity

The student records contained in the DRA*2* (21.1%) and *DIBELS Next* (8.6%) datasets accounted for almost 30% of all first grade students enrolled in New Mexico public schools during the 2011-2012 school year (National Alliance for Public Charter Schools, 2017). Although the ELL and non-ELL populations are very similar between the two datasets, there are notable differences that could be a result of the selection criteria established by the PED for districts or schools to participate in any of the state funded reading initiatives. For instance,

1. The *DIBELS Next* dataset has considerably more economically disadvantaged students than the DRA*2* dataset.
2. More Asian and Black students are represented in the DRA*2* dataset than the *DIBELS Next* dataset.
3. Native Americans are over-sampled in the *DIBELS* Next dataset and under sampled in the DRA2 dataset when compared to the population metrics of all first graders enrolled during the 2011-2012 school year.

Threats to External Validity

The data selected for this study were information-rich cases that allowed for an in-depth investigation into one aspect of early literacy assessments. The data contained within the two early literacy datasets are representative of schools and districts across New Mexico, and accurately reflects the unique enrollment and population of each district's community. Generalizability is limited to samples that are similar in terms of demographics and assessment performance. The findings are also limited to students that had BOY, MOY, and EOY composite scores as well as third grade SBA reading scores. A strength of the study is that students that transferred schools from first grade to the third grade SBA are included, but students that entered mid-year of the first grade were not included, so interpretations of an entire school or district population are not appropriate. It is possible that a school or district offered more than one reading intervention or program, other than those associated with the progress monitoring strategies embedded in each of the two early literacy assessments, and it is possible the effects of other interventions or programs interacted. It is also possible that ELL students, and students demonstrating early cognitive impairment, may not have responded as well as non-ELL or non-cognitively impaired students because common assessment accommodations are not permitted with the *DIBELS Next*, as discussed further below.

Imprecision of Measures

Imprecision of measures can impact the collected data, which will impact the findings. Variables included in this study that have potential imprecision in the reported data are:

English language learner status: The data associated with identifying ELL students could be subject to four possible anomalies. The first problem is that ELL students are treated as a homogeneous group. In reality, ELL students are a heterogeneous population characterized by differences in current English proficiency, native language explicitly linked to country of origin, native language literacy when entering the US schools, and amount of formal education prior to entering the US (Lakin & Young, Fall 2013). The second problem is the assumption that every ELL is accurately identified across all districts. The third problem is that the first grade BOY administration of the *DIBELS Next* assessment does not include word identification. Longitudinal studies with early elementary monolingual Spanish-speaking students found that word-level reading skills predicted a considerable amount of variance in later reading comprehension, and language comprehension is critical to reading comprehension (Farnia & Geva, 2013). A fourth problem is the *DIBELS Next* does not allow typical assessment accommodations, i.e., stating the directions in the student's home language, or allowing the student to have extended time.

Racial group: Racial group is a parent or guardian-reported variable, and limited to five federally defined classifications. Students from mixed racial group parents cannot be identified in two racial group categories, and students from the Middle East are regularly misclassified as 'Asian" or "Native American". Both situations have the potential to muddle the findings.

Economic disadvantaged status: SES is a composite variable that is based on household income, parent occupation or employment status, and household size. A complication in using free and reduced-price lunch eligibility as a proxy for socio

economic status (SES) is the lack of discernment between situational and generational poverty, which is critical in truly understanding the background of a six-year-old student and his or her distinct academic strengths and weaknesses. Specifically, free and reduced-price lunch data does not distinguish between a student whose parents are in advanced degree programs and temporarily below the poverty line (i.e., a child in situational poverty); versus a child of a single parent who has no high school diploma and is the third generation in poverty (i.e., a child in generational poverty).

Limitations

This project has several limitations:

1. The DRA*2* data are limited to one district, and not all first graders from all elementary schools within the district are represented in the dataset.
2. The research is limited to students who were in attendance for the *DIBELS-Next* or DRA*2* assessment administrations in first grade in 2011, and the third grade English language arts/reading section of the SBA in 2014.
3. Demographic characteristics of participating students, and the crucial elements of the educational environments, are hopelessly confounded across the districts and programs.
4. This study is limited to the utility of each early literacy assessment and does not address any aspect of reading theory related to the development, administration or interpretation of results of either of the two early literacy assessments.

Recommendations

The ultimate goal of every teacher, school or district administrator, or state-level policymaker is to foster student success both in school and in preparation for being

finished with school. This encompasses the objective of students becoming competent and confident readers. Existing statute and policy are inadequate in supporting that ambition. Based on the findings of this study, I offer the following four recommendations:

How a student is identified, and ultimately classified, as a good or struggling reader depends upon the measures and criteria used to classify students. Either the New Mexico legislature or the Secretary of Education must define the minimum standards for selecting and implementing any early literacy assessment in public schools in New Mexico. The minimum standards should include the eight "desirable" criteria for any assessment (Betts, et al., 2008; Compton, et al., 2010; Rouse & Fantuzzo, 2006):

1. Evidence of meeting traditional psychometric standards of reliability and validity,
2. capacity to predict later reading outcomes and model growth,
3. sensitivity to change, instructional needs including effects of interventions,
4. independence from a specific curriculum or program of instruction,
5. the capacity of the results to inform instruction,
6. feasibility including cost to implement and time to administer, and
7. classification accuracy with sensitivities above .90 and specificities above .80, and
8. fair to all groups about which inferences will be made as a result of the administration.

Yet, no language within New Mexico statute or administrative code clearly describes what any assessment adopted or purchased with public funds must be able to demonstrate

other than alignment to the adopted content standards. This vagueness allows the Secretary, PED bureau chiefs, superintendents, or administrators to pick any assessment that claims to be aligned to the content standards with performance benchmarks, purports to measure student achievement in reading skills, and can discern between the subgroups in an effort to monitor the achievement gap. The result is a smorgasbord assessment system and findings.

The New Mexico legislature must have an understanding of how their work to address the pervasive underachievement in reading of New Mexico public schools' students has caused serious tension between the PED and local superintendents or administrators of charter schools, possibly setting the reform efforts back 10 or more years. Specifically, the New Mexico legislature has stated the Secretary is '…the governing authority' with the '…control, management and direction of all public schools…' (22-2-1 NMSA 1978), but the superintendent of a public school district or administrator of a charter school is the chief executive officer of the school district (22-5-14 NMSA 1978). Unfortunately, who is responsible for adopting a scientifically-based reading program and assessment is a quagmire of conflicting or vague language that has caused conflict throughout the state. For instance:

- The *DIBELS Next* was selected as the common formative early literacy assessment for all three statewide reading initiatives (New Mexico Public Education Department, 2012), even though there were considerable existing controversies within the academic and policy research regarding *DIBELS* when it was adopted (Riedel, 2007).

- The PED did not subject the *DIBELS Next* to the scrutiny as required in the summer review process as it was deemed "other instructional materials" and could be adopted with an 'in-house review' (22-15-8 NMSA 1978).
- Later, the *DIBELS* became the only early literacy assessment that could be used in the student achievement measure section of the NMTEACH Educator Effectiveness system. The decision to restrict all districts to using only the *DIBELS Next* as the kindergarten through second grade assessment completely disregarded the autonomy of public and charter schools, and circumvented the adoption and purchasing requirements that superintendents must adhere to (6.17.2 NMAC).

The cost to taxpayers to purchase materials and provide professional development to implement the *DIBELS*, beginning about 2007, has been at least $3 million per year (New Mexico Public Education Department, 2012). The $3 million per year does not include district or school-related costs. The cost to students, however, is immeasurable.

The continuous decline in the numbers of proficient readers at the third grade is a strong indicator that the procedures for determining who is at risk of reading failure are severely lacking. The Secretary and superintendents must be held accountable for *how* students are identified, and the method must be more than what is involved in existing 'data-based decision making' protocols. The Secretary must adopt predictive measures that are accurate measures of later reading outcomes, and apply those measures on an annual basis. The measures must include classification accuracy, sensitivity, specificity, and cut-points that demarcate 'proficient' from 'not proficient'.

There must be changes to policy that stipulates how federal and state funds are expended in public schools to address the omnipresent low reading rates in New Mexico. For instance, public school districts or charter schools with at least the state average of number of students that receive free or reduced lunch can apply for federal Title I monies to address the deficits in reading and math. Districts or schools can choose between a targeted or schoolwide assistance model to expend Title I funds. Targeted assistance will provide reading intervention to students that receive free or reduced lunch *and* are not proficient in reading as measured by the state summative tests beginning with the third grade. Identified students receive supplemental instruction that may be limited to a computer-aided reading intervention program such as Read 180 or Accelerated Reading. A schoolwide assistance model will allow the school or district to apply the monies in such a way that every student in the school or district could benefit, for instance a librarian (if the school or district did not have sufficient budget for a librarian). In the schoolwide assistance model, all students benefit equally, and no preference is given to those students that are not proficient and receiving free or reduced school lunch. The problem with both models becomes the lack of a *timely* and *accurate* identification system of any student at risk of reading difficulty beginning with kindergarten or first grade. The targeted assistance model allows funding to go to *any* student below proficient beginning at the *end* of third grade who receives free or reduced price lunch, which is not timely. The schoolwide assistance model does not target need to specific students, which is neither a timely nor accurate system of identifying students at risk. New Mexico must petition the Office of Education for a waiver to the existing policy and request Title I funds be used for *any student,* beginning in kindergarten, that is below

academic targets for reading skills. The argument could be grounded in the research findings that the SES of a school has more impact on a student's academic achievement than the student's SES (Fergusson, 2008).

APPENDICES

Appendix A

Third Grade English Language Arts/Reading SBA Percent Proficient or Advanced

Table A1: Longitudinal Summary of English Language Arts/Reading SBA Percent Proficient and Advanced for All Students and Subpopulations in Third Grade

Subpopulation	2014	2013	2012	2011	2010	2009	2008	2007
All students	51.8	55.2	52.4	52.9	57.4	61	53.2	54.1
Female	55.9	59.7	57.5	57.9	62.1	66.3	58.6	59.0
Male	47.9	51.0	47.4	48.2	52.8	55.9	48.1	49.4
White	67.4	71.1	68.5	70.5	73.2	76.5	69.3	69.6
Black	48.1	56.6	48.8	50.8	53.8	59.6	49.7	48.9
Hispanic	48.3	51.0	47.8	48.2	52.9	56.5	47.2	48.7
Asian	76.3	75.3	74.3	69.8	74.5	78.3	69.8	69.2
American Indian	32.3	39.2	36.3	35.5	40.3	41.8	38.9	38.6
Economically disadvantaged	44.8	48.4	44.8	45.6	50.5	53.8	44.5	46.4
Not Economically disadvantaged		*	*	*	*	*	69.1	*
English language learners (current)	33.7	35.9	28.1	33.0	39.8	45.3	34.5	40.7
Not English language learner		*	*	*	*	*	59.3	*

*Data not provided by the PED

Appendix B

DRA*2* Subtests (Measures) and Administration Procedures

A district-level assessment coordinator that specialized in early literacy assessments provided the professional development to administer the DRA. The teacher administers the assessment in a one-on-one setting usually in the student's classroom. The assessment manuals for both early literacy assessments provide guidance for the testing environment: the student and teacher should be at a small table in a relatively quiet part of the classroom, and the classroom should have minimal disruptions or noise. Remaining students would be engaged in seatwork or other quiet activity. The time for administration depends on the student's independent reading level. Emergent readers (Levels A-3) will take 5-10 minutes, early readers (Levels 4-12) will take 10 minutes, transitional readers (Levels 14-24) will take 15-20 minutes, and extended readers (Levels 28-38) will take 30 or so minutes.

There are four steps to administering the DRA*2*:

- Assess the student's reading engagement.
- Assess the student's oral reading.
- Evaluate the student's comprehension/printed language concepts.
- Assess the student's performance.

The method of assessing each student begins with a student or teacher selecting a fiction or nonfiction book from the DRA*2* kit. Each book is leveled based on word count, sentence complexity, use of graphics and placement of text. Nonfiction texts are included at Levels 16, 28, 38 and 40 so that a teacher can assess how well students preview, read, and comprehend informational texts.

The DRA*2* measures are:

- Reading Engagement: Levels A through 3 assess Reading Engagement on three indicators (e.g., literacy support, favorite book, and book-handling skills), rated in three performance levels (emerging, developing, or independent). Levels 4 through 24 assess Reading Engagement on two indicators (e.g., book selection and sustained reading) in four performance levels (intervention, instructional, independent, or advanced). Levels 28 through 40 assess Reading Engagement on

two indicators (e.g., wide reading and self-assessment/goal setting) in the same performance levels as Levels 4 through 24.

- Oral Reading Fluency: Levels A through 3 assess Oral Reading Fluency on three indicators (e.g., monitoring/self-corrections, use of cues, and accuracy) rated in three performance levels (emerging, developing, or independent). Levels 4 through 12 assess Oral Reading Fluency on four indicators (e.g., phrasing, monitoring/self-corrections, problem-solving unknown words, and accuracy) in four performance levels (intervention, instructional, independent or advanced). Levels 14 through 40 assess Oral Reading Fluency on four indicators (e.g., expression, phrasing, rate and accuracy) rated in the same performance levels as Levels 4 through 12 (Pearson Education, Inc., 2009). Accuracy is measured with the number of miscues out of a total number of words in a book. The number of words and the complexity of the sentences defines a level of book. Reading rate is not assessed until Level 14, which is typically at the end of the first grade. An independent reader reads at a rate of 40 to 70 words per minute.

Printed Language Concepts/Comprehension assesses if a student understands the text. This is accomplished by asking a student to retell elements of the story such as the main ideas, key facts, characters, events or topics. At Level 4 “Printed Language Concepts” becomes “Comprehension”. The indicators for Comprehension change over the level of book. Unlike Reading Engagement and Oral Reading Fluency, in which the indicators are fairly consistent for a grade band of student reading (e.g., Levels A through 3 usually are beginning kindergarten readers), the indicators for Comprehension are rather perplexing. As an example, the indicator *Retelling: Sequence of events* begins in Level 4 and remains until Level 14, is not an indicator in Level 16, then resumes in Levels 18 – 24 with the exact descriptors found in Levels 4 through 14. The indicator *Retelling: Characters and Details* follows the same pattern as *Retelling: Sequence of events*. The indicator *Reflection*, however, begins at Level 4 through Level 14, is not in Level 16, then resumes in Levels 18 through 24, is not in Levels 28 or 30, then resumes in Levels 34, 38, and 40. The rationale for leaving out some indicators at different levels is not addressed in either the Technical or Administration Manual.

Appendix C

DIBELS Next Subtests (Measures) and Administration Procedures

Professional development to administer the *DIBELS Next* was provided by an Amplify-approved trainer over the summer to those teachers of the K-3 Plus program. The teacher administers the assessment in a one-on-one setting usually in the student's classroom. The assessment manuals for both early literacy assessments provide guidance for the testing environment: the student and teacher should be at a small table in a relatively quiet part of the classroom, and the classroom should have minimal disruptions or noise. Remaining students would be engaged in seatwork or other quiet activity. The time for administration depends on the student's grade level. First grade students at the beginning of year will need approximately a total of 5 minutes. The order of the subtests is not important. The primary concern is to closely follow the script for administration, start a timer for one minute for each of the three subtests, begin promptly after saying "begin", and accurately score each response. Teachers may use reminder procedures and apply a 3-second 'wait rule' for students to respond. Each subtest has specific rules of when to discontinue the test. After totaling the score for each subtest, the teacher transfers the results into the database maintained by Amplify.

- Letter Naming Fluency (LNF) is a direct measure of a student's fluency with naming letters. The Dynamic Measurement Group report that fluency in naming letters is a strong and robust predictor of later reading achievement and an indicator of risk. All letters are included, upper and lower case, and arranged in random order. The total score is the number of correct letter names a student says in 1 minute (Good, 2011).). Goffreda and DiPerna (2010) found the reliability evidence for LNF in seven of the studies: four reported test-retest (range of .83 to .93), two reported alternate form (.80 and .94), and one reported inter-rater (.94). The reliability evidence was fairly robust across multiple indices which indicates that the probes were consistent measures of student performance across time periods, forms, and examiners (p.469).

- Phoneme Segmentation Fluency (PSF) is a brief, direct measure of phonemic awareness: the student's fluency in segmenting a spoken word into its component parts or sound segments. The student hears a word and is asked to say the sounds in the word. A correct sound segment is any different, correct part of the word the student says. The total score

is the number of correct sound segments that the student says in one minute (Good, et. al, p.22, 2011). Goffreda and DiPerna (2010) found the reliability evidence for PSF in three studies: one reported test-retest (.88), and two reported alternate form (.62 and .97).

- Nonsense Word Fluency (NWF) is a brief, direct measure of the alphabetic principal and ability to blend letter sounds into consonant-vowel-consonant (CVC) and vowel-consonant (VC) words (Good, 2011). The words usELL are phonetically regular make-believe (i.e., nonsense or pseudo) words. Two subtests comprise NWF: Correct Letter Sounds (CLS) and Whole Words read (WWR). CLS is the number of letter sounds producELL correctly in one minute (Good, et. al, p. 24, 2011). Goffreda and DiPerna (2010) found the reliability evidence for NWF in three studies: two reported test-retest (.87 and .92), and one reported alternate form (.58).

Appendix D

Reliability and Validity Estimates for *DIBELS* Measures – First Grade

Table D1

Summary of Reliability Estimates for DIBELS Measures – First Grade

	Type of Reliability					
	Alternate Form		Test-Retest		Inter-Rater	
DIBELS Measure	Single-Form	Three-Form	Single-Form	Three-Form	Single-Form	Three-Form
Letter Naming Fluency (LNF)	--	--	--	--	.99	1.00
Phoneme Segmentation Fluency	--	--	--	--	.95	.98
NWF Correct Letter Sounds	.85	.94	.76	.90	.99	1.00
NWF Whole Words Read	.90	.96	.70	.88	.99	1.00
DORF Words Correct	.95	.98[a]	--	.95[a]	--	--
DORF Accuracy	--	.88[a]	--	.84[a]	--	--
DIBELS Composite Score	.95	--	.94	--	.99	--

- [a] Reliability coefficients calculated from the median score of three benchmark passages, and are thus reported as three-form or triad reliability. Three-form reliabilities that are not marked are estimated using the Spearman-Brown Prophecy Formula.

Table D2

First Grade DIBELS Next Predictive Validity Coefficients and Discriminant Validity Statistics for BOY Administration with GRADE Total Test

DIBELS Measure	Predictive Validity
LNF	.54
PSF	.33
NWF Correct Letter	.43
NWF Whole Words Read	.39
Composite Score	.55

Table D3

First Grade DIBELS Next Discriminant Validity Statistics for BOY Administration with GRADE Total Test

Below 40th Percentile			Above 40th Percentile			Difference Stat	
N	Mean	SD	N	Mean	SD	*t*-Stat	Cohen's *d*
54	105.00	29.68	139	145.0	39.54	7.33	1.11

Appendix E

DRA*2* Reliability and Validity Estimates

Table E1

Summary Tables of Reliability Estimates for DRA2

	Type of Reliability		
Grade 1-3	Test-retest	Inter-rater	Rater-expert
Comprehension	.99	.72	.79
Fluency	.97	.66	.89

- NOTE: Test-retest occurred in Spring 2008. Sample size of students grades 1 through 3 was 90.
- Second test administered approximately 14 days following the first test administration with a different reading passage within a level to reduce potential confounding effects associated with student memory.

Table E2

Summary Tables of Internal Consistency Reliability for DRA2

	Internal Consistency (Cronbach's alpha)	
Level	Oral Fluency (N of indicators=4)	Comprehension (N of indicators=6-7)
4	.784	.818
6	.849	.805
8	.680	.778
10	.736	.825
12	.758	.853
14	.542	.779
16	.731	.583
18	.614	.816
20	.725	.739
24	.725	.710
28	.788	.693
30	.778	.717
34	.745	.636
38	.611	.655
40	.762	.722
50	.785	.759
60	.717	.818
70	.621	.728
80	.622	.730

Table E3

Summary Table of the Measures Used to Establish Criterion Validity of the DRA2

Measure	Comprehension		Fluency	
	N	Spearman's Rho	N	Spearman's Rho
GORT-4	66	.66	66	.62
GORT-4	66	.65	66	.69
DORF	66	.70	66	.74
Teacher rating of student reading ability	188	.6	188	.63

Appendix F

DRA2 Outlier Graphical Analyses

Figure F1: DRA*2* Predicted Probability vs. Standardized Residuals

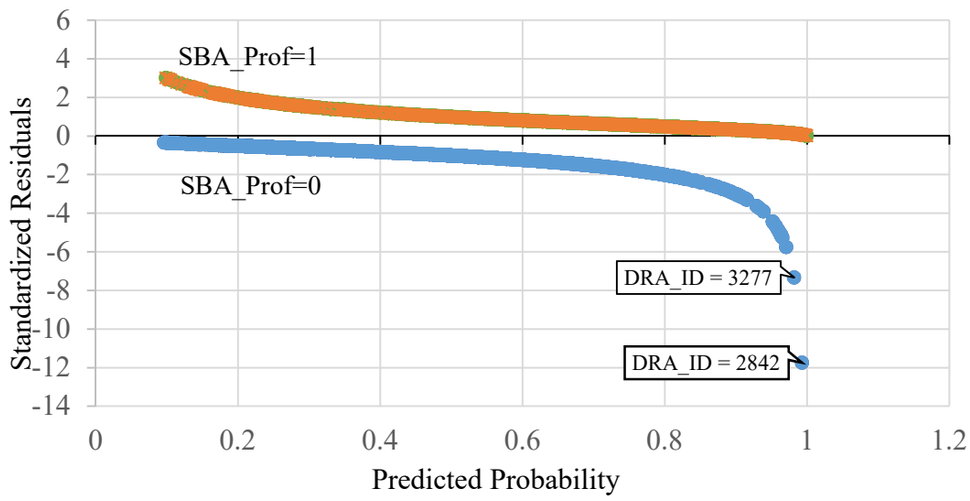


Figure F2: DRA2 Predicted Probability vs. Deviance Residuals

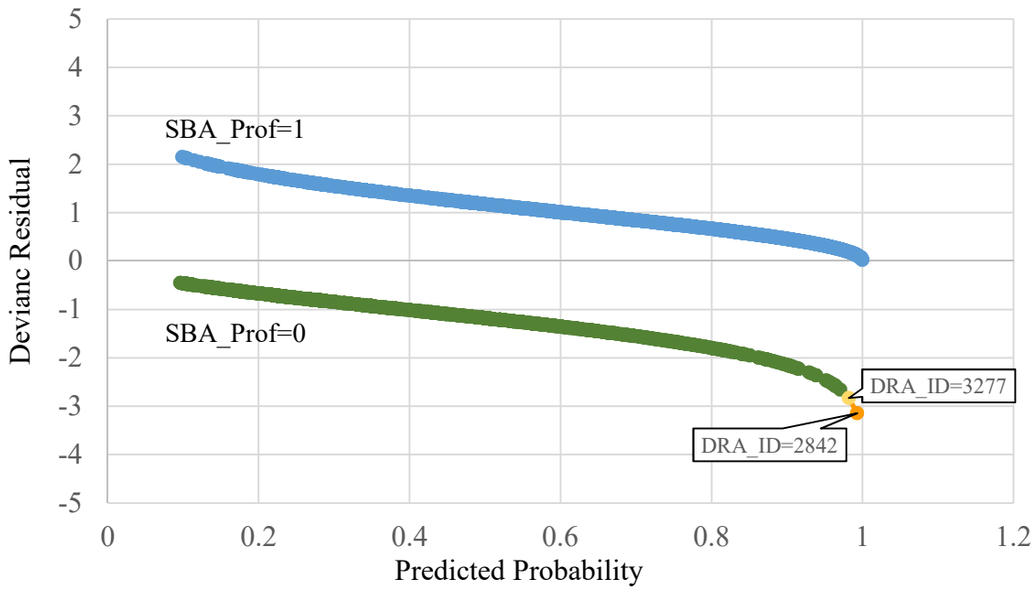


Figure F3: DRA2 –Predicted Probabilities vs. Delta Chi-Squared

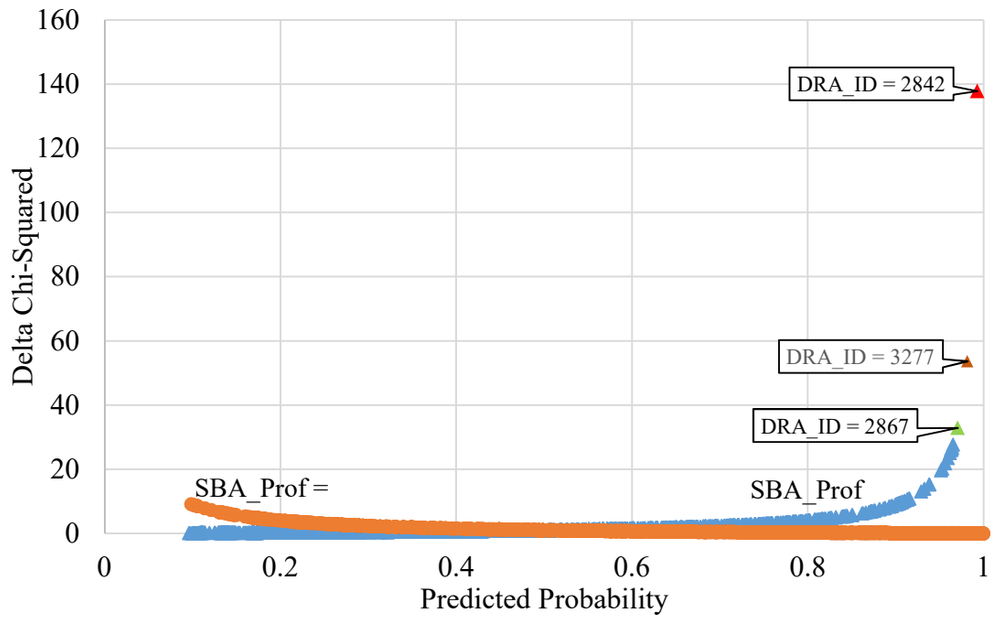
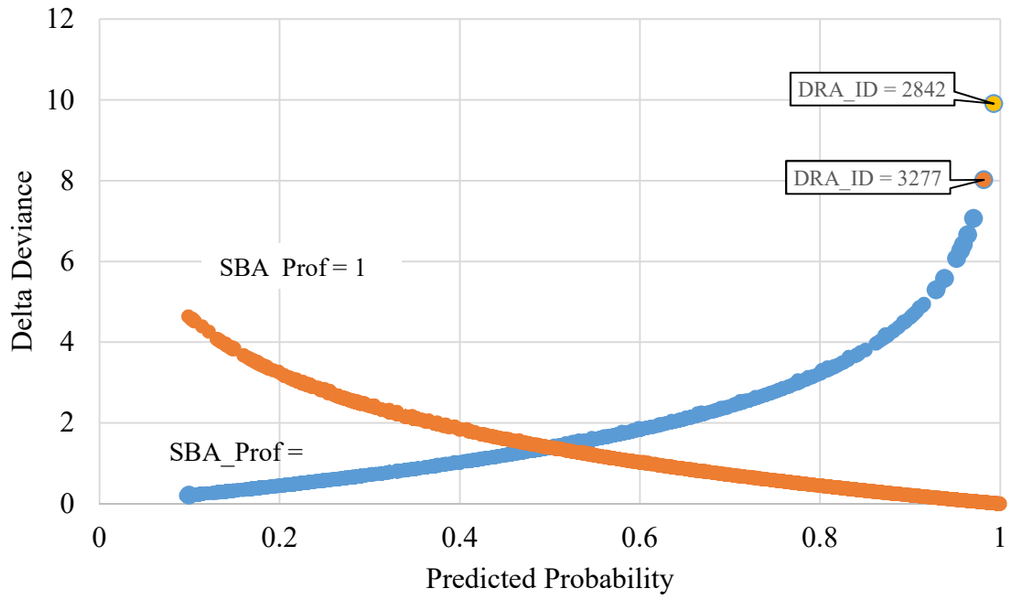


Figure F4: DRA2 –Predicted Probabilities vs. Delta Deviance



Appendix G

DIBELS Next Outlier Graphical Analyses

Figure G1: *DIBELS Next* Predicted Probability vs. Standardized Residuals

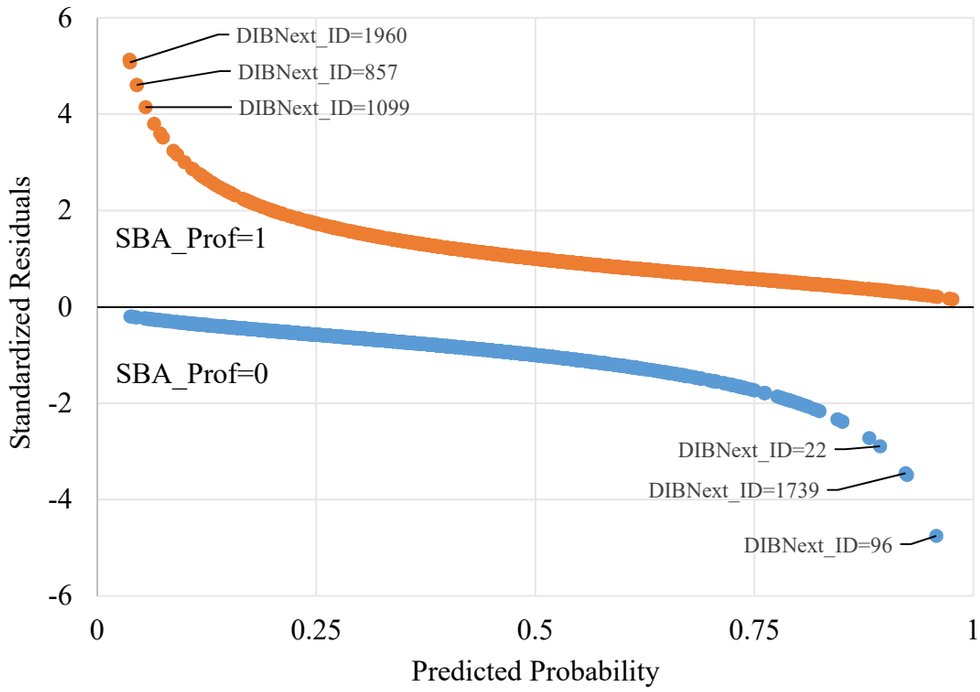


Figure G2: *DIBELS Next* Predicted Probability vs. Deviance Residuals

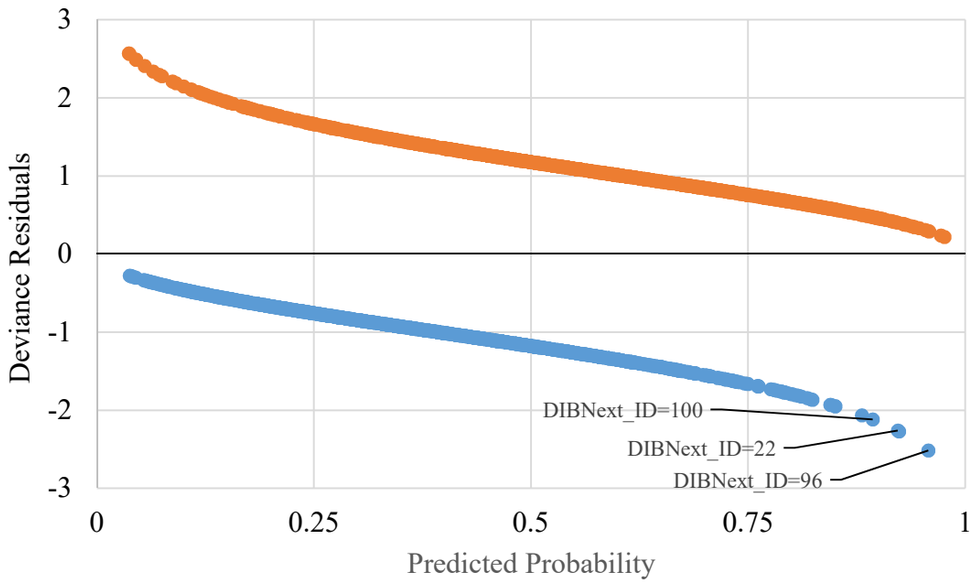


Figure G3: *DIBELS Next* –Predicted Probabilities vs. Delta Chi-Squared

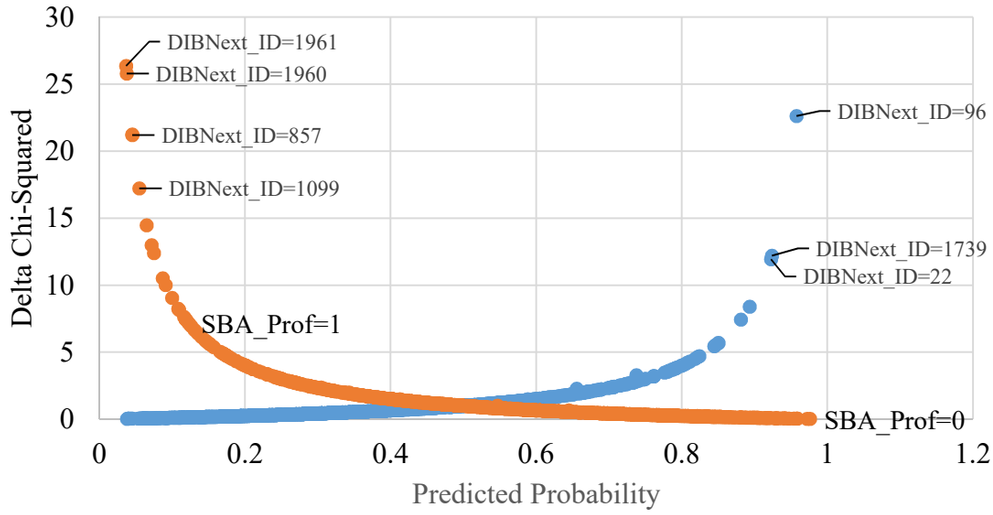
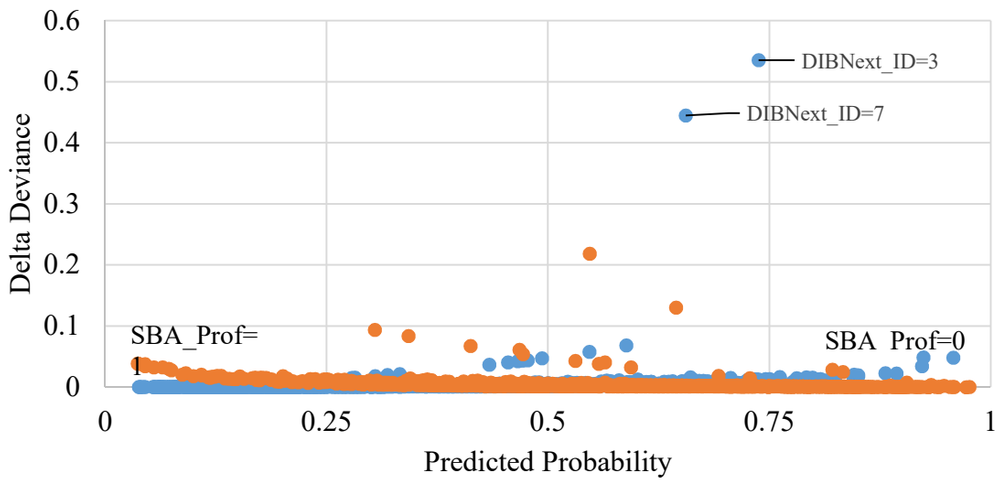


Figure G4: *DIBELS Next* –Predicted Probabilities vs. Delta Deviance



Appendix H

DRA*2* and *DIBELS Next* Multicollinearity Statistics

Table H1. DRA*2* Multicollinearity Statistics

	Statistic	BOY Composite	Asian	Black	Hispanic	Native American	Cauc	Econ Disad.	ELL Status	Gender
BOY Comp	VIF		1.038	1.024	<o	1.023	1.267	1.259	1.138	1.001
	Tol		.963	.976	<0	.977	.789	.794	.879	.999
Asian	VIF	1.165		2.28	10.559	3.069	8.99	1.334	1.161	1.007
	Tol	.859		.439	.095	.326	.111	.75	.861	.859
Black	VIF	1.165	1.821		8.248	7.313	7.313	1.334	1.161	1.007
	Tol	.859	.549		.121	.137	.137	.75	.861	.993
Hispanic	VIF	1.165	1.047	1.024		1.024	1.272	1.334	1.161	1.007
	Tol	.859	.955	.976		.977	.786	.75	.861	.993
Nat. American	VIF	1.165	1.485	1.553	4.995		4.73	1.334	1.161	1.007
	Tol	.859	.674	.644	.2		.211	.75	.861	.2
Caucasian	VIF	1.165	1.103	1.124	1.573	1.199		1.334	1.161	1.007
	Tol	.859	.907	.89	.636	.834		.75	.861	.993
Econ Disad	VIF	1.1	1.031	1.024	<0	1.024	1.137		1.129	1.006
	Tol	.909	.97	.977	<0	.977	.879		.886	.994
ELL Status	VIF	1.141	1.035	1.013	<0	1.021	1.241	1.297	<0	1.007
	Tol	.876	.966	.987	<0		.806	.771	<0	.993
Gender	VIF	1.141	1.035	1.013	<0		1.241	1.297		1.007
	Tol	.876	.966	.987	<0		.806	.771		.993

Table H2. *DIBELS Next* Multicollinearity Statistics

	Statistic	BOY Composite	Asian	Black	Hispanic.	Native American	Cauc.	Econ Disad.	ELL Status	Gender
BOY Comp	VIF		1.008	1.014		1.094	1.207	1.104	1.091	1.002
	Tol		.992	.986		.914	.828	.906	.916	.998
Asian	VIF	1.073		3.476	54.667	37.008	36.833	1.109	1.137	1.015
	Tol	.932		.288	.018	.027	.027	.901	.880	.986
Black	VIF	1.073	1.400		22.114	15.260	15.267	1.109	1.137	1.015
	Tol	.932	.714		.045	.066	.066	.901	.880	.986
Hispanic	VIF	1.073	1.010	1.014		1.094	1.207	1.109	1.137	1.015
	Tol	.932	.991	.986		.914	.828	.901	.880	.986
Nat. American	VIF	1.073	1.027	1.051	1.644		1.820	1.109	1.137	1.015
	Tol	.932	.974	.951	.608		.549	.901	.880	.986
Caucasian	VIF	1.073	1.020	1.050	1.811	1.817		1.109	1.137	1.015
	Tol	.932	.980	.952	.552	.550		.901	.880	.986
Econ Disad	VIF	1.068	1.007	1.013		1.082	1.150		1.133	1.014
	Tol	.936	.993	.987		.924	.870		.882	.986
ELL Status	VIF	1.030	1.020	1.050	1.724	1.735		1.106		1.015
	Tol	.970	.981	.952	.580	.576		.904		.986
Gender	VIF	1.060	1.009	1.014		1.094	1.207	1.109	1.137	
	Tol	.944	.991	.987		.914	.828	.902	.880	

Appendix I

DRA2 Variable Summary for the Training and Test Datasets

Table I1: DRA2 Variable Summary for the Training and Test Datasets

		Training		Test	
		Mean	Count	Mean	Count
BOY Composite Score		181.36		178.46	
SS		39.51		39.42	
SBA Proficiency	Not proficient		1307		1312
	Proficient		1419		1415
Subtotal			*2,726*		*2727*
Home_Language	Albanian		0		1
	American Sign Language		4		3
	Arabic		8		8
	Cantonese		0		1
	Chinese		10		10
	English		1821		1822
	Farsi		3		3
	French		2		2
	German		1		1
	Hebrew		1		1
	Indonesian		1		1
	Japanese		1		1
	Jicarilla Apache		1		0
	Keres		6		7
	Korean		3		3
	Laotian		1		0
	Mandarin		0		1
	Navajo		31		31
	NULL		1		0
	Other		14		15
	Polish		1		0
	Russian		2		3
	Spanish		782		782
	Swahili		1		1
	Tagalog		4		3
	Tewa		0		1
	Thai		1		1
	Tiwa		1		1
	Towa		2		1
	Urdu		1		1
	Vietnamese		18		19
	Zuni		4		3
Subtotal			*2,726*		*2,727*
Racial group	1 Asian		62		61
	2 Black		81		74
	3 Hispanic		1815		1818
	4 Nat.American		136		125
	5 Caucasian		632		649
Subtotal			*2,726*		*2,727*
Economic Disadvantaged Status	Not economically disadvantaged		805		795
	Economically disadvantaged		1921		1932
Subtotal			*2,726*		*2,727*

		Training		Test	
		Mean	Count	Mean	Count
Economic Disadvantaged Status	Not economically disadvantaged		805		795
	Economically disadvantaged		1921		1932
Subtotal			*2,726*		*2,727*
ELL Status	Not ELL		2150		2147
	ELL Current		576		580
Subtotal			*2,726*		*2,727*
Gender	Female		1364		1363
	Male		1362		1364
Subtotal			*2,726*		*2,727*

Appendix J

DIBELS Next Variable Summary for the Training and Test Datasets

Table J1: *DIBELS Next* Variable Summary for the Training and Test Datasets

		Training		Test	
		Mean	Count	Mean	Count
BOY Composite Score		115.00		115.75	
SS		37.69		37.40	
SBA Proficiency	0 Not proficient		596	605	
	1 Proficient		508	497	
Subtotal			*1104*	*1102*	
Racial group	1 Asian		5	5	
	2 Black		13	12	
	3 Hispanic		635	633	
	4 Native American		225	226	
	5 Caucasian		226	226	
Subtotal			*1104*	*1102*	
Economic Disadvantaged Status	0 Not economically disadvantaged		156	159	
	1 Economically disadvantaged		948	943	
Subtotal			*1104*	*1102*	
ELL Status	0 Not ELL		853	854	
	1 ELL Current		251	248	
Subtotal			*1104*	*1102*	
Gender	1 Female		550	547	
	2 Male		554	555	
Subtotal			*1104*	*1102*	

REFERENCES

Allison, P. D. (2014). Measures of Fit for Logistic Regression. *SAS Global Forum*, (pp. 1-12).

Barger, J. (2003). *Comparing the DIBELS Oral Reading Fluency indicator and the North Carolina end of grade reading assessment.* Asheville: NC: North Carolina Teacher Academy.

Bartlett, J. (2014, 02 16). *The Hosmer-Lemeshow goodness-of-fit test for logistic regression.* Retrieved from The Stats Geek: http://thestatsgeekcom/2014/02/16/the-hosmer-lemeshow-goodness-of-fit-test-for-logistic=regression

Beach, K. D., & O'Connor, R. E. (n.d.). Early response-to-intervention measures and criteria as predictors of reading disability in the beginning of third grade. *Journal of Learning Disabilities, 48*(2), 1-76. doi:10.1177/0022219413495451

Beaver, J. (2006). *Developmental Reading Assessment K-3 Teacher Guide.* Parsippany, NJ: Pearson Education, Inc. publishing as Celebration Press.

Betts, J., Reschly, A., Pickart, M., Heistad, D., Sheran, C., & Marston, D. (2008). An examination of predictive bias for second grade reading outcomes from measures of early literacy skills in kindergarten with respect to English language learners and ethnic subgroups. *School Psychology Quarterly, 23*(4), 553-570. doi:10.1037/1045-3830.23.4.553

Binomial Calculator: Online Statistical Table. (2017). Retrieved July 26, 2017, from StatTrek.com: http://stattrek.com/online-calculator/binomial.aspx

Buckingham, J. W.-W. (2013). Why poor children are more likely to become poor readers: The school years. *Austrailian Journal of Education, 57*(3), 190-213. doi:10.1177/0004944113495500

Buckingham, J., Wheldall, K., & Beaman-Wheldall, R. (2013). Why poor children are more likely to become poor readers: The school years. *Austrailian Journal of Education, 57*(3), 190-213. doi:10.1177/0004944113495500

Buckinham, J. B. (2013). Why poor children become poor readers: The early years. *Educational Review*. doi:10.1080/00131911.2013.795129

Buis, M. L. (2016). Logistic regression: When can we do what we think we can do? 1-14.

Canary, J. D., Blizzard, L., Barry, R. P., Hosmer, D. W., & Stephen, J. Q. (2015). A comparison of the Hosmer-Lemeshow, Pigeon-Heyse, and Tsiatis goodness-of-fit tests for binary logistic regression under two grouping methods. *Communications in Statistics - Simulation and Computation*, 1-51. doi:10.1080/03610918.2015.1017583

Chung, S. F. (2012, September). Research-based vocabulary instruction for English language learners. *Reading Matrix: An International Online Journal, 12*(2).

Claessens, A. D. (2009). Kindergarten skills and fifth-grde achievement: Evidence from the ECLS-K. *Economics of Education Review, 28*, 415-427.

Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (n.d.). *Applied multiple regression/correlational analysis for the behavioral sciences* (3rd ed.). Hillsdale, NJ: Earlbaum.

Compton, D. L., Fuchs, D., Fuchs, L. S., Bouton, B., Gilbert, J. K., Barquero, L. A., . . . Crouch, R. C. (2010). Selecting at-risk first-grader readers for early intervention: Eliminating false positive and exploring the promise of a two-state gated screening process. *Journal of Educational Psychology, 102*(2), 327-340. doi:10.1037/a0018448

Coyne, M. D., Kame'enui, E. J., Simmons, D. C., & Harn, B. A. (2004). Beginning reading intervention as inoculation or insulin: First-grade reading performance of strong responders to kindergarten intervention. *Journal of Learning Disabilities, 37*(2), 90-104.

Cunningham, A. E., & Stanovich, K. E. (1997). Early reading acquisition and its relation to reading experience and ability ten years later. *Developmental Psychology, 33*, 934-945.

DeMaris, A. (1991). A framework for the interpretation of first-order interaction in logit modeling. *Psychological Bulletin, 110*(3), 557-570.

DeMaris, A. (1993). Odds versus probabilities in logit equations: A reply to Roncek. *Social Forces, 71*(4), 1057-1065.

DeMaris, A. (1995). A tutorial in logistic regression. *Journal of Marriage and Family, 57*(4), 956-968. Retrieved 11 20, 2014, from http://www.jstor.org/stable/353415

DeMaris, A., Teachman, J., & Morgan, S. P. (1990). Interpreting logistic regression results: A critical commentary. *Journal of Marriage and Family, 52*(1), 271-277. Retrieved 12 2, 2014, from http://www.jstor.org/stable/352857

Department, N. M. (n.d.). *NM PED.* Retrieved 2017, from Instructional Material Bureau: http://ped.state.nm.us/InstructionalMaterial/dl10/Adoption%20Framework1.pdf

Dewey, E. N., Powell-Smith, K. A., Good, R. A., & Kaminski, R. A. (2015). *DIBELS Next Technical Adequacy Brief.* Eugene, OR: Dynamic Measurement Group, Inc.

Entwisle, D. R., Alexander, K. L., & Olson, L. S. (2007). Early Schooling: The handicap of being poor and male. *Sociology of Education, 80*, 114-138.

Farnia, F., & Geva, E. (2013). Growth and predictors of change in English language learners' reading comprehension. *Journal of Research in Reading, 36*(4), 389-421.

Fergusson, D. H. (2008). The transmission of social inequality: Examination of the linkages between family socioeconomic status in childhood and educational achievement in young adulthood. *Research in Social Stratification and mobility, 26*, 277-295.

Fiester, L. (2010). *Early Warning: Why Reading by the end of third grade matters.* Baltimore, MD: The Annie E. Cassey Foundation. Retrieved from www.aecf.org

Fuchs, D., Compton, D. L., Fuchs, L. S., Bryant, V. J., Hamlett, C. L., & Lambert, W. (2012). First-grade cognitive abilities as long-term predictors of reading comprehension and disability status. *Journal of Learning Disabilities, 45*(3), 217-231.

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

Gilbertson, D., & Bluck, J. (2006). Improving responsive to intervention for English-language learners: A comparison of instructional pace on letter naming rates. *Journal of Behavioral Education, 15*, 131-147. doi:10.1007/s10864-006-9013-y

Goffreda, C. T., & DiPerna, J. C. (2010). An empirical review of psychometric evidence for the Dynamic Indicators of Basic Early Literacy Skills. *School Psychology Review, 39*(3), 463-483.

Goffreda, C. T., DiPerna, J. C., & Pedersen, J. A. (2009). Preventive screening for early readers: predictive validity of the Dynamic Indicators of Basic Early Literacy skills (DIBELS). *Psychology in the Schools, 46*(6), 539-552. doi:10.1002/pits.20396

Good, R. H., & Kaminski, R. A. (2016, October). *DIBELS Next.* Retrieved from Dynamic Measurement Group: https://dibels.org/dibelsnext.html

Good, R. H., Kaminski, R. A., Dewey, E. N., Wallin, J., Powell-Smith, K. A., & Latimer, R. J. (2011). *DIBELS Next Technical Manual.* Eugene, OR: Dynamic Measurement Group, Inc.

Good, R., Kaminski, R. A., Cummings, K., Dufour-Martel, C., Petersen, K., Powell-Smith, K., . . . Wallin, J. (2011). *DIBELS Next Assessment Manual.* Longmont, CO, USA: Cambium Learning Group/Sopris.

Gottardo, A., Collins, P., Baciu, I., & Gebotys, R. (2008). Predictors of grade 2 word reading and vocabulary learning from grade 1 variables in Spanish-speaking children: Similarities and differences. *Learning Disabilities Research & Practice, 23*(1), 11-24.

Gunn, B., Smolkowski, K., Biglan, A., & Black, C. (2002). Supplemental instruction in decoding skills for Hispanic and non-Hispanic students in early elementary school: A follow-up. *The Journal of Special Education, 36*(2), 69-79.

Hecht, S. A., & Greenfield, D. B. (2001). Comparing the predictive validity of first grade techer ratings and reading-related tests on third grade levels of reading skills in young children exposed to poverty. *School Psychology Review, 30*(1), 50-69.

Hernandez, D. (2010). *How third-grade reading skills and poverty influence high school graduation.* Baltimore, MD: Annie E. Casey Foundation. Retrieved from www.aecf.org

Hickey, D. W. (2012). *The Developmental Reading Assessment and the third grade Reading Ohio Achievement Assessment: an investigation of possible relationships (Doctoral dissertation).* Ann Arbor, MI: Retrieved from PorQuest Information and Learning Company.

Hosmer, D. W., & Lemeshow, S. (1989). *Applied Logistic Regression.* New York: John Wiley & Sons.

Hung, H., Jean Chiu, Y., & Yeh, H. (2013). Multimodal assessment of and for learning: A theory-driven design rubric. *British Journal of Educational Technology, 44*(3), 400-409. doi:10.1111/j.1467-8535.2012.01337.x

Kingdon, D., Serbin, L. A., & Stack, D. M. (2017). Understanding the gender gap in school performance among low-income children: A developmental trajectory analysis. *International Journal of Behavioral Development, 41*(2), 265-274.

Kvalseth, T. O. (1985). Cautionary note about R Squared. *The Merican Statistician, 39*, 279-285.

Lakin, J. M., & Young, J. W. (Fall 2013). Evaluating growth for ELL students: Implications for accountability policies. *Educational Measurement, 32*(3), 11-26.

Lesnick, J., Goerge, R. M., Smithgall, C., & Gwynne, J. (2010). *Reading on grade level in third grade: How is it related to high school performance and college enrollment?* Chicago: Chapin Hall at the University of Chicago. Retrieved from http://www.chapinhall.org/sites/default/files/Reading_on_Grade_Level_111710.pdf

Lewin, N. L. (2011). *Is the Developmental Reading assessment (DRA2) a predictor of success on the third grade Texas Assessment of Knowledge and Skills (TAKS) Reading test? (Doctoral dissertaion).* retrieved from ProQuest.

Lubienski, S. &. (2010). Beyond free lunch: Which family background measures matter? *Education Policy Analysis Archives, 18*, 1-39.

Martz, E. (2013, April). *Enough is Enough! Handling Multicollinearity in Regression Analysis.* Retrieved 2017, from The Minitab Blog: http://blog.minitab.com/blog/understanding-statistics/handling-multicollinearity-in-regression-analysis

Measured Progress. (2014). *New Mexico Standards Based Assessment 2013-2014 Technical Report.* Retrieved from http://ped.state.nm.us/AssessmentAccountability/AssessmentEvaluation/sba/2012/2013-14%20NM%20Technical%20Report.pdf

Menard, S. W. (2001). *Applied Logistic Regression Analysis* (Second ed.). Thousand Oaks, CA: Sage University Papers Series on Quantitative Applications in the Social Sciences, 07-106.

Morgan, S. P., & Teachman, J. D. (1988). Logistic Regression: Description, examples, and comparisons. *Journal of Marriage and the Family, 50*(4), 929-936. Retrieved 11 20, 2014, from http://www.jstor.org/stable/352104

National Alliance for Public Charter Schools. (2017). Retrieved from www.publiccharters.org/dashboard/students/page/lep/state/NM/year/2012

National Institute of Child Health and Human Development. (1998). *Preventing reading difficulties in young children.* Retrieved from http://www.nap.edu/catalog/6023.html

National Reading Council. (1998). *Preventing reading difficulties in young children.* Washington, DC: National Academy Press.

National Reading Panel. (2000). *Teaching children to read: An evidence-based assesment of the scientific research literature on reading and its implications for reading instruction: Reports of the subgroups (NIH Publication No. 00-4754).* Washington, DC: National Institute of Child Health and Human Development. Retrieved from http://www.nap.edu/catalog/6023.html

Nations Report Card.gov. (n.d.). Retrieved 2017, from Institute of Education Sciences, National Center for Education Statistics: https://nces.ed.gov/help

Nelson, J. R., Benner, G. J., & Gonzalez, J. (2005). An investigation of the effects of a prereading intervention on the early literacy skills of children at risk of emotional disturbance and reading problems. *Journal of Emotional and Behavioral Disorders, 13*(1), 3-12.

New Mexico K-3 Plus. (2011). *mCLASS:DIBELS Guide to Managing Data.* Retrieved from http://www.ped.state.nm.us/EarlyChildhood/dl11/New%20Mexico%20K3%20Plus%20Guide%202011%202012.pdf

New Mexico Public Education Department. (2012). *K-3 Plus Education Annual Report for School Year 2011- 2012.* Santa Fe.

Olson, L. (2007). *DIBELS Involved in 'Reading First' Controversies*. Retrieved from Education Week: http://www.edweek.org/ew/articles/2007/05/02/35form-nmside.h26.html

Pampel, F. (2000). *Logistic Regression: A Primer.* Thousand Oaks, CA: Sage University Papers Series on Quantitative Applications in the Social Sciences, 07-132.

Pearson Education, Inc. (2009). *K-8 Technical Manual Developmental Reading Assessment.* Upper Saddle River, New Jersey: Pearson Education, Inc.

Peduzzi, P. C. (1996, December). A simulation study of the number of events per variable in logistic regression analysis. *Journal of Clinical Epidemiology*(12), 1373-1379.

Peng, C. Y., & So, T. S. (2002). Logistic regression analysis and reporting: A primer. *Understanding Statistics, 1*(1), 31-70.

Powell-Smith, K. A., Good, R. H., Habedank Stewart, L., & Dewey, E. N. (2011, March). The utility of DIBELS Next for predicting GRADE outcomes. San Francisco, CA: Poster presented at the National Association of School Psychologists' Annual Convention.

Public Education Department. (2007). *NMPED New Mexico Accountability Data (2009-2015). Proficiencies State, District and School, by Grade.* Retrieved October 11, 2016, from ped.state.nm.us/AssessmentAccountability/AcademicGrowth/NMSBA.html: http://www.ped.state.nm.us/AssessmentAccountability/AcademicGrowth/NMSBA.html

Public Education Department. (2015). *K-3 Plus*. Retrieved June 2, 2016, from http://www.ped.state.nm.us/Literacy

Public Education Department. (n.d.). *New Mexico Public Education Department, Instructional Material Bureau.* Retrieved from Instructional Material Adoption Framework: http://ped.state.nm.us/InstructionalMaterial/dl10/Adoption%20Framework1.pdf

Riedel, B. (2007). The relation between DIBELS, reading comprehension, and vocabulary in urban first-grade students. *Reading Research Quarterly, 42*(4), 546-567. doi:10.1598/RRQ.42.4.5

Rouse, H. L., & Fantuzzo, J. W. (2006). Validity of the Dynamic Indicators for Basic Early Literacy Skills as an indicator of early literacy for urban kindergarten children. *School Psychology Review, 35*(3), 341-355.

Samuels, S. J. (2007). The DIBELS tests: Is speed of barking at print what we mean by reading fluency? *Reading Research Quarterly, 42*(4), 563-567.

Sarkar, S. K., Midi, H., & Rana, S. (2011). Detection of outliers and influential observations in binary logistic regression: An empirical study. *Journal of Applied Sciences, 11*(1), 26-35. doi:10.3923/jas.2011.26.35

Seals, M. P. (2013). *Impact of leveled reading books on the fluency and comprehesion levels of first grade students (Doctoral dissertation.* Ann Arbor, MI: Retrieved from ProQuest.

Shaw, R., & Shaw, D. (2002). *DIBELS Oral Reading Fluency-based indicators of third grade reading skills for Colorado State Assessment Program (CASP).* Eugene, OR: University of Oregon.

Skandera, H. (2011, July 1). *New SBA Performance Standards, Scale and Data [Memorandum].* Retrieved from NMPED: http://ped.state.nm.us/AssessmentAccountability/AssessmentEvaluation/dl11/New%20SBA%20Cut%20Score%20Memo.pdf

Snow, C. E., Burns, M. S., & Griffin, P. (n.d.). Predictors of Success and Failure in Reading. In N. R. Counsel, *Preventing Reading Difficulties in Young Children.* Retrieved from http://www.nap.edu/catalog/6023.html

Sparks, S. D. (2011, April 8). *Inside School Research.* Retrieved from Education Week: http://blogs.edweek.org/edweek/inside-school-research/2011/04/the_disquieting_side_effect_of.html

Stoltzfus, J. C. (2011). Logistic regression: A brief primer. *Academic Emergency Medicine, 18*, 1099-1104. doi:10.1111/j.1553-2712.2011.01185.x

The National Early Literacy Panel. (2009). *Early beginnings: Early literacy knowledge and instruction.* National Institute for Literacy, National Institute for Child Health and Human Development.

Thomson, S. &. (2010). *Challenges for austrailian education: Results from PISA 2009.* Melbourne: ACER.

UCLA: Statistical Consulting Group. (n.d.). Retrieved December 2016, from http://www.ats.ucla.edu/stat/spss/webbooks/reg/chapter2/spssreg2.htm

US Department of Education. (n.d.). *Reading First State Profile: New Mexico.* Retrieved December 30, 2015, from Reading First: http://www2.ed.gov/programs/readingfirst/index.html

Walsh, A. (1987). Teaching understanding and interpretation of logit regression. *TeachingSociology, 15*(2), 178-183. Retrieved from http://www.jstor.org/stable/1318033

Wang, C., & Algozzine, B. (2011). Oral reading rates of second-grade students. *Journal of Educational Psychology, 103*(2), 442-454.

Wanzek, J., & Vaughn, S. (2009). Students demonstrating persistent low response to reading intervention: Three case studies. *Learning Disabilities Research & Practice, 24*, 151-163.

Wilson, J. (2005). *The relationship of Dynaimic Indicators of Basic Early Literacy Skills (DIBELS) Oral Reading Fluency to performance on Arizona Instrument to Measure Standards (AIMS).* Tempe, AZ: Tempe School District No.3.

Xie, X., Pendergast, J., & Clarke, W. (2008). Increasing the power: A practical approach to goodness-of-fit test for logistic regression models with continuous predictors. *Computational Statistics & Data Analysis, 52*, 2703-2713.