2018

A Comparison of an Oral Reading Fluency Measure and a Reading Comprehension Measure to Identify Students at Risk for Failure on a State Assessment of Reading

Brian Engler

Recommended Citation

Engler, Brian, "A Comparison of an Oral Reading Fluency Measure and a Reading Comprehension Measure to Identify Students at Risk for Failure on a State Assessment of Reading" (2018). *PCOM Psychology Dissertations*. 489.
https://digitalcommons.pcom.edu/psychology_dissertations/489

Philadelphia College of Osteopathic Medicine

Department of Psychology

A COMPARISON OF AN ORAL READING FLUENCY MEASURE AND A READING COMPREHENSION MEASURE TO IDENTIFY STUDENTS AT RISK FOR FAILURE ON A STATE ASSESSMENT OF READING

By Brian Engler

Submitted in Partial Fulfillment of the Requirements for the Degree of

Doctor of Psychology

July, 2018

PHILADELPHIA COLLEGE OF OSTEOPATHIC MEDICINE
DEPARTMENT OF PSYCHOLOGY

Dissertation Approval

This is to certify that the thesis presented to us by

______________________________ on the ______ day of __________________,

20___, in partial fulfillment of the requirements for the degree of Doctor of Psychology,

has been examined and is acceptable in both scholarship and literary quality.

Committee Members' Signatures:

______________________________, Chairperson

____

____

______________________________, Chair, Department of Psychology

Acknowledgements

To my wife Anne, your love and patience over the last year has been a much needed source of support as I went through the process of writing this dissertation. I will always be grateful for your willingness to let me pursue a doctoral degree and to listen to me talk about reading tests as I paced about the house for the past year.

To my mother, Dr. Janet Cahill, thank you for the time you spent learning about reading and supporting me through this process. Your questions and critiques helped me bolster my own understanding of this topic and improved my writing dramatically. You were also instrumental in getting me though my moments of panic and doubt by remaining positive and insistent that I could successfully do this.

To my father, Rick Engler, thank you for your encouraging words and interest you showed throughout the writing of this dissertation. Your tireless work and dedication have always inspired me and taught me that serving others is a rewarding way to spend one's career.

Thank you to Dr. George McCloskey for your encouragement, commitment, and guidance over the last year. I am grateful for the many hours you spent working with me and making sure that nothing got in our way. To Dr. Katy Tresco and Dr. Fernando Cavallo, thank you both for your support and insightful feedback during the writing of this dissertation. When I look back at the various stages of this project, I cannot see how I could have gotten to this final document without your help.

To the staff at Resica Elementary, thank you for your dedication and willingness to support important research by compiling years' worth of data that has led to the completion of numerous dissertations.

Abstract

Significant incentives exist for educators to efficiently identify, students at risk of failing statewide assessments. This study strives to add to the body of research on curriculum-based assessments (CBAs) used, in part, for this purpose. This study compares the ability of two commonly used CBAs to identify students at risk for failure on the Pennsylvania System of School Assessment (PSSA). To this end, results from the Dynamic Indicators of Basic Early Literacy Skills (DIBELS) Oral Reading Fluency (DORF) were compared with the results from the Measures of Academic Progress Reading Comprehension Test (MAP-R). The study analyzed the scores of 93 fourth grade students from a suburban/rural elementary school, all of whom were administered DORF and MAP-R in the fall and spring of the 2016–2017 school year and took the PSSA in 2017. Results from each assessment were examined in terms of several indices (Improvement, Stability, Specificity, and Sensitivity) and were compared, using chi-squared analyses. Overall, the DORF and MAP-R performed comparably, with few statistically significant differences between them.

Table of Contents

List of Tables

CHAPTER 1

INTRODUCTION

The use of standardized assessments of academic proficiency has become the norm across the country. School districts have a vested interest in diverting significant resources in programs designed to identify and track the progress of students at risk for academic problems. This is due, in part, to a desire to improve statewide test results, which influence funding, staffing, and other outcomes (No Child Left Behind, 2001). Although there appears to be a mounting public backlash against high-stakes standardized testing (New York Times Editorial Board, 2013; Layton, 2015), schools still face incentives to make progress on these tests and require reliable methods for identifying students who are likely to struggle. Many schools utilize curriculum-based assessments (CBAs) of reading and math in order to track student progress and target interventions.

The ability to determine whether or not a student is at risk for failure is important for ensuring that students are provided with appropriate interventions as early as possible. Furthermore, although CBAs are helpful for informing instruction, conducting the assessments can take time away from that instruction. Decisions regarding which CBA to use to identify students at risk for poor standardized test outcomes must be informed by understanding how each option is most likely to align with the desired outcome measure, in this case, the PSSA. In 2015, the Pennsylvania Department of Education implemented changes in the Pennsylvania System of School Assessment (PSSA) that resulted in a sharp drop in the number of students scoring in the proficient or advanced range (McCorry, 2015). The drop in scores across both reading and math in all grades (except fifth grade reading) indicates that students who may not have been identified as being at risk for failure on the PSSA in previous years may now be at risk. Further the change in scores makes it necessary to revisit the relationship between the methods used to identify students at risk for failure on the PSSA and outcomes on the test itself.

Statement of the Problem

Given the importance and the scope of the use of CBAs to influence results on standardized tests, it is important that the nature of this relationship is based upon a strong empirical basis. Attempting to replicate and refine an understanding of the relationship between commonly used CBAs and the PSSA constitutes an important endeavor. In choosing tools for screening students for risk of failure on statewide assessments, a number of factors need to be considered, including cost, ease of use, and reliability (Goo, Watt, Park, & Hosp, 2012). Although CBAs tend to be brief and relatively inexpensive, they take instructional time and require school resources in order to be purchased. Identifying which CBAs are most efficient and effective can help inform decisions regarding how much time and money to expend on these assessments.

Purpose of the Study

This study re-examines the relationship between curriculum-based measures in reading and Pennsylvania’s statewide literacy assessment. It attempts to identify which of the two measures currently in use by a particular elementary school is more effective at various points in time during the school year. The study specifically examines two measures of reading: the Dynamic Indicators of Early Literacy Skills (DIBELS) Oral Reading Fluency (DORF) assessment (Good & Kaminski, 2002) and the Measures of Academic Progress reading comprehension assessment (MAP-R) (Northwest Evaluation Association, 2004).

CHAPTER 2

REVIEW OF THE LITERATURE

Introduction

Reading comprehension is an essential skill for learning and life. In 2015, only 36% of fourth grade students in the U.S. were reading at or above the proficient level (National Center for Education Statistics, 2015). In Pennsylvania, 39% of fourth graders scored at the basic or the below basic level for literacy on the 2017 statewide test (Pennsylvania Department of Education, 2017). Federal laws such as The No Child Left Behind Act (NCLB) (US Department of Education, 2001) and its successor, The Every Student Succeeds Act (US Department of Education, 2015), obligate states and local education agencies to provide data regarding reading proficiency and yearly progress toward improving scores. As a result, school districts are incentivized to identify, efficiently, those students at risk for failure on statewide assessments and to implement interventions.

Curriculum-based assessments constitute useful tools for measuring student progress within a curriculum and have been demonstrated to be reliable indicators of academic functioning (Van Der Heyden & Burns, 2010). Many school districts utilize them to screen for academic problems efficiently and to monitor student progress over time (Deno, 1985). Significant research has been conducted examining the relationship between CBAs and statewide testing (Shapiro, Solari, & Petscher, 2008; Scherr, 2011; Dorshimer, 2009; Northwest Evaluation Association, 2016). In Pennsylvania specifically, a number of studies have examined the efficacy of CBAs' ability to predict PSSA results (Lucas, 2013; Weinstein, 2011; Dorshimer, 2009).

This study strives to refine the knowledge base with regard to the use of reading CBAs to predict outcomes on the annual statewide assessment in Pennsylvania. The study specifically examines the predictive ability of two reading CBAs used in an elementary school to screen students for reading problems.

Standardized Testing

Federal and State Laws

The NCLB, passed in 2001, requires that all states measure reading and math progress at least once each year in grades three through eight (US. Department of Education, 2004). Results are submitted to the US Department of Education and are used as part of a complex set of factors determining whether or not schools have made, “adequate yearly progress” (AYP). Schools that do not make AYP and receive funding under Title 1 potentially face consequences, such as staff replacement, conversion into a charter school, or other types of restructuring. As a result, there is significant pressure to achieve AYP, including identifying students who are at risk for poor performance on statewide assessments.

PSSA

The PSSA is a standardized test of literacy, math, and science taken annually by students in grades three through eight (Pennsylvania Department of Education, 2009). Results on the PSSA fall into four categories: advanced, proficient, basic, and below basic. The test uses “cut scores” to assign these categories. It should be noted that in 2015, the state revised the test and adjusted the cut scores, effectively raising the bar for being considered proficient or above. As illustrated in Table 2.1, across all grade levels (except for fifth grade reading), these changes led to an average drop in scores of 9.4% in literacy and 35.4% in math, compared with the previous year.

Table 2.1

PSSA Score Differences Between 2013-2014 and 2014-2015

Grade	Reading 2013–14	ELA 2014–15	Difference	Math 2013–14	Math 2014–15	Difference
3	29.7	37.9	-8.2	24.9	51.5	-26.6
4	31.3	41.4	-10.1	23.7	55.5	-31.8
5	39.4	38	1.4	22.8	57.2	-34.4
6	35.5	40.2	-4.7	28	60.2	-32.2
7	27.9	41.4	-13.5	23.3	66.9	-43.6
8	20.4	41.7	-21.3	26.4	70.1	-43.7
Average:			-9.4			-35.4

This is a notable difference in scores for most grades and warrants further investigation into methods for identifying students who may now be at risk for failure, but may not have been in previous iterations of the test. This study employs PSSA results from the current (i.e., post 2014–2015) version of the test.

Curriculum-Based Assessment of Reading

Many school districts utilize CBAs to measure student progress and identify those students who are struggling to improve. Furthermore, CBAs can be used to assess a variety of basic skills and measure growth in those skills over time. To this end, specific measures are chosen to align with instructional goals and establish a baseline of proficiency in a particular skill. As instruction is conducted, the measures are repeated to track progress over time. Typically, if sufficient progress is not being made, instructional changes are implemented. This concept is broadly referred to as response to intervention (RTI). Appropriately conducted, CBAs provide a number of advantages over other methods of measuring student progress. First, many CBAs are very brief, and many can be conducted in fewer than five minutes. In addition, CBAs also measure classroom skills and can be tailored to measure a student’s specific ability in an area of instruction. Beyond this, CBAs are also simple to track and compare current skills with previous measurements. In effect, CBAs allow educators to quickly identify students who may be struggling, make instructional choices based on their skills, measure their progress over time, and use that data to further inform instruction. In some states and localities, CBAs are also used as part of broader RTI programs employed to determine eligibility for special education.

Elements of Reading

Reading Comprehension

Reading comprehension comprises the ability to understand written material. It is critically important both for learning and for real life applications. Reading comprehension is considered “the essence of reading” (Durkin, 1993) and is affected by a number of different skills, including vocabulary development, reading fluency, prior knowledge, and cognitive skills, such as reasoning and working memory (National Reading Panel, 2000; Nouwens, Groen, & Verhoeven, 2017). The literacy components of the PSSA and the MAP-R are designed to assess reading comprehension directly (Pennsylvania Department of Education, 2009; Northwest Evaluation Association, 2009), but DORF assesses only oral reading fluency.

Oral Reading Fluency

Oral reading fluency (ORF) refers to the ability to read aloud quickly and accurately (Adams, 1994). This is often used by schools as a primary method of measuring overall reading skills (Shapiro & Clements, 2009), as well as predicting performance on state assessments. Additionally, ORF is also able to measure, directly, phonological skills and word recognition, and it can further constitute an indicator of reading comprehension.

Research has supported the notion that ORF can be predictive of overall reading ability, as well as performance on state assessments. Early research into the connection between ORF and comprehension was conducted by L.S. Fuchs, Fuchs, and Maxwell (1988). Their study utilized three types of reading comprehension assessments and an ORF assessment, all created for the purpose of the study, comparing results from each to the Reading Comprehension subtest of the Stanford Achievement Test (Gardner, Rudman, Karlsen, & Merwin, 1982). Reading comprehension was measured in several ways. The first involved answering questions based on the written text. The next measured reading comprehension through passage recall, in which students were asked to retell as much detail as possible from a 400-word passage. The third method was a cloze task, in which every seventh word in a 400-word passage was deleted and replaced with a blank. Students were asked to fill in the blank with a word that fit the context of the passage. The oral reading task required reading two 400-word passages, with scores based on correct words per minute. A total of 70 middle school students, all of whom had a reading disability, participated in the study. The average correlations were .82 for answering questions, .70 for passage recall, and .72 for cloze. The oral reading task was the most strongly correlated among the study’s measures at .91. The authors cited this as evidence supporting the idea that ORF constitutes a valid measure of broad reading abilities.

There is also evidence that ORF is a superior measure of reading capacity, compared with reading individual words in isolation. Jenkins, Fuchs, Espin, van den Broek, and Deno (2000) conducted a study of 113 fourth grade students' performances on the Iowa Test of Basic Skills, a comprehension test, and two short measures of reading fluency. The first measure was a 400-word folktale. The second was a list of randomly ordered words from the folktale. Students were provided one minute to read each measure. The correlation between the word list and the comprehension test was .53. The correlation between the folktale and the comprehension test was .83, suggesting that ORF is a stronger measure of overall reading than the ability to read words in isolation.

Curriculum-Based Measures of Reading

Measures of Academic Progress (MAP-R)

The MAP-R is a computer-based measure of reading comprehension. It dynamically assesses the reader's skill and adjusts the difficulty accordingly (Northwest Evaluation Association, 2013). The tests are not timed and typically take between 15 and 30 minutes to complete. School districts that conduct MAP-R testing typically do so two to three times per year.

A study of MAP-R conducted by its publisher (Northwest Evaluation Association, 2016) examined its relationship to the PSSA. The study produced cut scores that correspond to the four levels of performance on the PSSA. The study concluded that MAP-R can be used reliably to predict PSSA results. The study was able to calculate the odds of a student receiving a score of proficient or higher on the PSSA, based on scores on the MAP-R. For example, a third grade student who obtains a score of 197 on the reading portion of the MAP-R in the spring has a 62% likelihood of a proficient score on the PSSA. A third grade student who obtains a score of 204 or higher has a >99% chance of scoring in the proficient or higher range in the PSSA. It is relevant to note that the study utilized the 2015 PSSA, which is the equivalent dataset used in this dissertation. The study further developed a series of charts that allow for comparing MAP-R scores with projected proficiency on the PSSA.

There appears to be limited research regarding the use of MAP-R as a general outcome measure outside of articles published by the MAP-R's developer. Results from one large study (Cordray, Pion, Brandt, Molefe, & Toby 2012) attempted to assess the impact of the MAP-R reading intervention and used the MAP-R assessment. Fall MAP-R reading scores were highly correlated with the Illinois state pre-test, but MAP-R scores from the spring appeared to overestimate performance on the statewide assessment. This raises the possibility that the MAP-R resulted in false positives in this study. Further, the MAP-R intervention program did not appear to influence reading skills significantly despite increases in MAP-R reading scores.

Klingbeil, McComas, Burns, and Helman (2015) compared three universal screening measures using spring MAP-R scores as an outcome variable: ORF probes from AIMSweb (Pearson, 2012), the Fountas & Pinnell Benchmark Assessment System (BAS) (Heinemann Publishing, n.d.), and fall MAP-R scores. The BAS was used to assess both reading comprehension and ORF, with results combined into a single score. Unsurprisingly, fall MAP-R scores were most predictive of spring MAP-R scores. When used in isolation, both AIMSweb and the BAS demonstrated high rates of false negatives; students who were at risk were not identified. When combined, the three measures were highly accurate in identifying students who were at risk.

One study found that ORF probes from AIMSweb were directly predictive of MAP-R scores from elementary students in Nebraska (Merino & Beckman, 2010). Interestingly, AIMSweb Maze scores (a cloze reading task) did not significantly predict MAP-R scores, suggesting that ORF alone is sufficient to measure overall reading ability.

January and Ardoin (2015) reported that little data existed concerning the technical adequacy of MAP. Their study of 802 students between first and fifth grades found high correlations between ORF probes and MAP-R scores. The correlation did not differ significantly between grades. As part of the study, 86 elementary school teachers completed a survey regarding their beliefs and understandings about CBAs. The surveys revealed that teachers reported being more likely to use MAP-R data than data from other CBAs available to them. The results also indicated that teachers doubted the validity of ORF probes for assessing reading comprehension.

DIBELS Oral Reading Fluency (DORF)

Dynamic Indicators of Basic Early Literacy Skills (DIBELS (Good & Kaminski, 2002) constitutes another commonly used tool for assessing reading. The DIBELS suite of assessments includes measures of ORF, phonemic awareness, phonics, vocabulary, and reading comprehension. According to its authors, DIBELS was designed to be a set of short and efficient screening and progress-monitoring tools that could serve as a general outcome measure for reading.

DORF (Good & Kaminski, 2002) is conducted in a one-on-one setting with a teacher. Students are provided with written passages and asked to read aloud for one minute. Scores are recorded as correct words per minute. Words omitted or substituted, as well as hesitations of more than three seconds, are scored as errors. Words self-corrected within three seconds are scored as accurate. Passages are leveled to reflect vocabulary at a difficulty level commensurate with the student's grade level expectations, with vocabulary becoming increasingly advanced at higher grade levels.

DIBELS has been extensively studied in terms of its utility as a progress monitoring and pre-referral tool for special education evaluations. One dissertation (Pollard, 2015), examined the predictive strength of DORF for reading scores from the Wechsler Individual Achievement Test, Third Edition (WIAT-III) (Breaux, 2009) ORF and Reading Comprehension subtests. The WIAT-III is a commonly used norm-referenced test of academic achievement used in psychoeducational assessments. The results indicated that DORF scores were predictive of WIAT-III scores in grades two through five as a group, but that the correlations were not significant at every grade level.

The literature also contains research on the ability of other DIBELS subtests to predict broader outcomes. Research conducted by the authors of DIBELS (Good, et al., 2004) examined correlations between four DIBELS measures and the Woodcock-Johnson Psycho-Educational Battery Total Reading Cluster (Mather & Woodcock, 2001). The researchers examined four related reading measures provided to kindergarten and first grade students: Initial Sound Fluency (ISF), Phoneme Segmentation Fluency (PSF), Letter Naming Fluency (LNF), and Nonsense Word Fluency (NWF). All of the DIBELS assessments measured were found to be moderately correlated with the Woodcock-Johnson. When data from repeated measures were examined (three to four exposures for each measure), the predictive validity was strong, ranging from r = .91 to .96.

DIBELS subtests may possess some utility for predicting reading ability in young children; however, there is evidence to suggest that they do not perform as well as ORF alone. Goffreda, Diperna, and Pedersen (2009) conducted a study of all DIBELS subtests' relationships both to the PSSA and to the TerraNova California Achievement Test (CAT) (CTB/McGraw-Hill, 2005). Scores from four DIBELS subtests (LNF, PSF, NWF, & ORF) conducted in first grade were compared with second grade CAT scores and third grade PSSA scores. This study suggested that of the various DIBELS assessments, only the scores from ORF were significant predictors of future performance on either the PSSA or CAT.

Reidel (2007) conducted a survey of over 1,500 urban first grade students and compared their scores on DORF with two tests of reading comprehension: the Group Reading Assessment and Diagnostic Evaluation (conducted in first grade) and the TerraNova (conducted in second grade). The DORF was able to predict correctly 80% of the students who would score poorly in first grade and 71% who would do so in second grade. Other DIBELS subtests (PSF, NWF, LNF, and Retell Fluency (RF)) were unable to predict reading comprehension scores accurately more than 40% of the time. The RF subtest, a proposed measure of comprehension (with no currently published norms), constituted a weaker predictor of overall comprehension. Neither did combining scores from ORF and RF improve the predictive power of ORF alone. Despite these findings, Riedel warned that the extremely short duration of DORF was insufficient to measure comprehension, and the amount of information that could be obtained in that time was too limited, particularly for older students.

The use of DIBELS as a general outcome measure and a tool for choosing curriculum has also been found to have limitations. At least one study (Cavallo, 2012) found that DIBELS does not predict performance on specific curricula. This study examined the ability of various DIBELS subtests to predict which students would benefit from instruction using the Fast For Word language program. There were no significant findings for any DIBELS subtest, including ORF, indicating that DIBELS is not an effective tool for tailoring instruction.

Specifically, DORF has not been universally accepted as a tool for measuring reading. Samuels (2007) argued that DIBELS does not measure fluency in a meaningful sense, noting that beginning readers and the more advanced readers engage in different processes. Beginning readers focus on decoding, whereas later readers are able to focus on comprehension and decoding simultaneously. Samuels also noted that older students who read too quickly tend to miss content and comprehend less, despite showing strong fluency. DORF, with its focus solely on speed, could underestimate problems with comprehension. Although DIBELS is a well-supported progress monitoring tool overall, one study found that its design may make administrators more easily prone to scoring errors, compared with other options (Ardoin & Christ, 2009).

Comparison of MAP-R and DORF

Although both MAP-R and DORF measure reading, they do so in very different ways, which may affect usability and outcomes. DORF is conducted by a teacher using brief reading passages aligned to grade level vocabulary. Student abilities are measured through correct words per minute with a one minute time limit. Teachers are able to adjust the difficulty of the passages to help determine the instructional level at which a student is currently functioning.

MAP-R testing is conducted on a computer, with no direct scoring observation needed from a teacher. The test is untimed, but takes 15–30 minutes to complete according to its publisher (NWEA, 2013). MAP-R testing can be conducted in groups. It utilizes a dynamic approach because it adjusts item presentation based upon the responses of the child. If a student responds correctly, the next question is harder. An incorrect response prompts the program to make the next question easier. MAP-R also directly assesses reading comprehension, whereas

DORF assesses only words per minute. The authors of DORF maintain that it is rare to find children who read fluently, but do not comprehend.

One major difference between the two systems is cost. According to the publisher of MAP-R, the annual cost per student is $13.50, not including a minimum $1,500 annual license fee. By contrast, the DIBELS suite, which includes ORF, comprehension, phonological awareness, and other assessments, costs $1 per student per year. This difference in cost may be enough for some school districts to choose DIBELS over MAP-R, regardless of any other factors.

Both assessments allow data to be graphed over time, and scores can be compared with national and local norms. This allows educators to better track the progress of individual students and compare them with their peers at the classroom, school, state, and national level.

Use of CBAs to Predict Statewide Outcomes

Measures of reading fluency have been found to be highly predictive of strong reading results in several states' assessments. In Oregon, 96% of children in third grade who met the benchmark for ORF scored at or above expectations on the Oregon Statewide Assessment (Good, Simmons, Kame'enui, 2001). In that study, students who read grade-level material at or above 110 words per minute were highly likely to meet expectations, but only 28% of students who read under 70 words per minute were rated as "meets expectations." Roehrig, Petscher, Nettles, Hudson, and Torgesen (2008) published similar findings between measures of DORF and Florida's state assessment. Notably, this study was able, consistently, to identify students at low risk for poor state scores. However, it was only moderately consistent in its ability to identify students at risk for reading problems on other measures. A meta-analysis of 41 studies of one-minute ORF probes found that ORF scores accounted for an average of 67% of variance in the reading scores in norm-referenced achievement tests for students in grades one through six (Reschly, Busch, Betts, Deno, & Long, 2009).

Previous research has examined the predictive validity of DORF, specifically with regard to the PSSA (Dorshimer, 2009; Sherr, 2011; Good, Simmons, & Kame'enui, 2001). Shapiro, Solari, and Petscher (2008) found DORF to be a strong predictor of students in grades three through five who were likely to score at or above grade level expectations on the PSSA, but it could not predict students who would be at moderate risk for reading problems. This body of research supports the use of ORF as a proxy for broad reading skills. Two previous studies (Dorshimer, 2009; Scherr, 2011) examined the use of DORF and MAP-R to screen for risk of failure on the PSSA in the same elementary school that has been used for this study (albeit with different cohorts of students).

Dorshimer (2009) used the PSSA as an outcome measure for curriculum-based measures of reading. He used both the DIBELS and MAP-R as possible predictors of the PSSA. He concluded that both measures were sensitive enough to measure student responses both to regular classroom instruction and to reading interventions. These findings were consistent across each cohort in the study, as well as within cohorts. Results from the DORF and MAP-R indicated that as students progressed, the percentage identified as "at risk" dropped. This supports the idea that these measures are closely aligned with curriculum and constitute reasonable predictors of academic performance.

Dorshimer's (2009) overall findings appear to provide support for the use both of DORF and of MAP-R to monitor progress and identify students at risk for failure on the PSSA. The study was less successful in predicting specific factors, such as which interventions were likely to be efficacious. Dorshimer also noted that the MAP-R scores did not provide evidence that one of the two domains assessed by MAP-R (vocabulary and comprehension) provided better predictors of intervention effectiveness. This finding questions the notion put forth by the publishers of MAP-R that it is effective at recommending specific interventions. The study also could not identify clear differences between the DIBELS' and MAP-R's false positives (students identified as at risk, but who passed the PSSA) and false negatives (students not identified as at risk, but later failed the PSSA). Both measures were found to be less effective at predicting which students would remain not at risk in later grades. For example, a student who was not considered to be at risk in third grade may be at risk for failure in fifth grade, but not be identified via DIBELS or MAP-R testing until those measures will have been conducted in the fall of fifth grade. Another relevant finding was that regardless of which measure was used, it became more difficult for the remaining at-risk students to improve as time went on. There was no difference in the ability of the DIBELS or MAP-R to identify these students who continued to be at risk. These students who were found to be at risk from year to year consistently failed to pass the PSSA.

One limitation of Dorshimer's (2009) study was that when drawing conclusions about which of the two measures was a more successful predictor of PSSA outcomes, the study relied exclusively on a discussion of raw percentages of students predicted to be at risk. The study did not utilize any multivariate analyses, which would have allowed for a more thorough examination of the relative strength of the measures, noting that the study did not include analyses to determine if the differences were statistically significant. Neither was the author able to identify the relative merits of using each measure, including whether or not measures of reading fluency could reliably and broadly assess reading comprehension.

Scherr (2011) analyzed the efficacy of MAP-R as a predictor of performance in an RTI program with the PSSA as an outcome measure. The study found that in a cohort of 82 fourth grade students, 62 scored as proficient on the PSSA. Of the 20 students who did not score as proficient, the MAP-R was able to predict accurately the outcomes of students who were not enrolled in intensive intervention programs. However, one key finding of the study was that the MAP-R was less than 50% accurate at predicting which students in higher intervention tiers would score as proficient on the PSSA. More than half of students in those groups who were not proficient on the spring MAP-R were able to pass the PSSA. This may mean that the MAP-R constituted an effective tool for targeting interventions.

Scherr also asserted that DIBELS is an adequate predictor of PSSA outcomes, but the MAP-R is a superior tool, because it is able to provide a more in-depth assessment of reading problems and is better suited to informing interventions. However, this conclusion does not address which tool is more useful and efficient for predicting PSSA outcomes. If DIBELS and MAP-R are both consistently able to predict reading comprehension outcomes, it may follow that the more efficient tool should be used as a universal screener, and the more in-depth assessment for those identified as at risk. A major limitation of Scherr's study was the small sample size drawn from a single cohort. The author himself reported that the small sample size made it difficult to identify trends in the data.

One major limitation of both Scherr's and Dorshimer's studies is that they relied on descriptive statistics to predict PSSA outcomes and did not directly compare DIBELS and MAP-R. As a result, it is impossible to determine whether or not a statistically significant difference exists between the two measures. Furthermore, data from the 2015 changes to the PSSA were not available, creating a need for further analysis.

Concluding Summary of the Literature

Schools are under significant pressure to utilize a variety of CBAs to help identify struggling readers and predict academic outcomes. However, these tests require time and resources, and so should be utilized efficiently. There is significant evidence that ORF is predictive of overall reading skills. Previous research has demonstrated that MAP-R Reading is highly predictive of outcomes on the PSSA. Research is needed to determine whether or not shorter, less expensive CBAs are able to produce similar outcomes.

Current Study

The current study attempts to build on prior literature by directly comparing DORF and MAP-R's relationship with PSSA outcomes through the use of statistical analysis. It expands on this body of questions by conducting a statistical analysis of differences between the two measures beyond the use of descriptive statistics. Variables of outcomes, such as true positives and true negatives, are compared in order to examine the efficacy of each measure in predicting results. It also provides a statistical analysis of the relationship between the measures and the current version of the PSSA following the changes implemented in 2015.

Research Questions

The following research questions are addressed in this study:

1. Is the proportion of fourth grade students in the current research study sample that were identified as not proficient on the PSSA reading assessment significantly different from the proportion of the population of 4th grade students in the state of Pennsylvania that were identified as not proficient on the PSSA reading assessment?

2. What proportion of fourth grade students are identified as at risk of earning a not-proficient category rating on the PSSA reading assessment based on fall and spring DIBELS DORF score categories?
3. What proportion of fourth grade students are identified as at risk of earning a not-proficient category rating on the PSSA reading assessment based on the fall and spring MAP-R score categories?
4. What is the relationship between DORF score ratings (at risk/not at risk) or MAP-R category ratings (not proficient/proficient) and PSSA score categories (proficient/not proficient) for fourth grade students?

    a. What proportion of fourth grade students identified as at risk with fall and spring DORF and MAP-R earned PSSA scores in the proficient range (operationally defined as the Improvement Index), and how do the fall and spring DORF and MAP-R compare as indicators of improvement?

    b. What proportion of students identified as not at risk with fall and spring DORF and MAP-R earned PSSA scores in the proficient range (operationally defined as the Instability Index), and how do the fall and spring DORF and MAP-R category scores compare as indicators of instability?

    c. What is the proportion of fourth grade students identified as at risk with fall and spring DORF and MAP-R who earned PSSA scores in the not-proficient range (operationally defined as the Sensitivity Index), and how do fall and spring DORF and MAP-R category scores compare as indicators of sensitivity?

d. What is the proportion of fourth grade students that earned PSSA scores in the proficient rang who were also identified as not at risk with the fall and spring DORF and MAP-R score categories (operationally defined as the Specificity Index), and how do the fall and spring DORF and MAP-R category scores compare as indicators of specificity?

e. What is the percentage of improvement over chance represented by the relationship between fall and spring DORF categories and PSSA reading score categories and the relationship between fall and spring MAP-R categories and PSSA reading score categories (operationally defined as the Kappa Index) for fourth grade students, and how do the fall and spring DORF and MAP-R category scores compare as predictors of PSSA outcomes in terms of improvement over chance?

5. How does success or failure in CBAs relate to future PSSA results?
    a. What percentage of students classified as at risk were able to pass the PSSA? Was there a significant difference between the PSSA passage rates for students identified as at risk on either DORF or MAP-R?
    b. What types of DORF-PSSA and MAP-R-PSSA score change patterns were exhibited by fourth grade students?
    c. What effect does considering combined DORF and MAP-R data have on predicting passage rates?

Summary of Research Questions

These research questions are intended to assess the relationship between DORF and MAP-R in a number of ways. They seek to examine the extent to which each assessment accurately identifies students at risk for failure on the PSSA and the relative level of an overestimation or an underestimation of those students. This study also seeks to examine whether or not combinations of the two CBAs affect the reliability of identification.

CHAPTER 3

METHODS

Overview

This study examined the relationship between PSSA results and results of DORF and MAP-R assessments. It attempted to determine if a measure of DORF is capable of identifying students at risk for failure on the PSSA with the same fidelity as a direct measure of reading comprehension (MAP-R).

Source of Data

Shelf data were accessed from one suburban/rural elementary school in Northeast Pennsylvania. The archived data that were accessed includes MAP-R, DORF, and PSSA scores for students who attended fourth grade during the 2016–2017 school year. The district used in the study is diverse, with a population that is 44.7% White, 22.5% Black, and 27.1% Hispanic. In this district, 19% receive special education services (Pennsylvania State Data Center, 2018)

Inclusion/Exclusion Criteria

Data from students who took the DORF and MAP-R in fall and/or spring, and the PSSA in the spring, were included. Data from students who do not have DORF and/or MAP-R scores or did not have a PSSA score on file were excluded. Student names were removed from the dataset and replaced with numbered identifiers to ensure that confidentiality was maintained. DORF and MAP-R scores were available for all of the 93 students who took the PSA during the 2016-2017 school year.

Measures and Materials

PSSA

The PSSA utilizes four performance level descriptors to report results (PA State Board of Education, 2007). Each level possesses different expectations that vary, depending on the student's grade level. Students must obtain a score of at least *proficient* to be considered as having met the state performance expectations. Data from the most recent (2017) PSSA results were used for this study (Pennsylvania Department of Education, 2017). The four performance levels for the English Language component of the PSSA are described as follows in Table 3.1:

Table 3.1
Description of PSSA Performance Levels

Advanced	The Advanced Level reflects superior academic performance. Advanced work indicates an in-depth understanding and exemplary display of the skills included in the Pennsylvania Academic Content Standards.
Proficient	The Proficient Level reflects satisfactory academic performance. Proficient work indicates a solid understanding and adequate display of the skills included in the Pennsylvania Academic Content Standards.
Basic	The Basic Level reflects marginal academic performance. Basic work indicates a partial understanding and limited display of the skills included in the Pennsylvania Academic Content Standards. This work is approaching satisfactory performance, but has not yet reached such. There is a need for additional instructional opportunities and/or increased student academic commitment to achieve the Proficient Level.
Below	The Below Basic Level reflects inadequate academic performance. Below Basic work indicates little understanding and minimal display of the skills included in the Pennsylvania Academic Content Standards. There is a major need for additional instructional opportunities and/or increased student academic commitment to achieve the Proficient Level.

MAP-R

The MAP-R is a computerized assessment of reading that adjusts its level of difficulty based on the user's performance (NWEA, 2003). Results are reported as Rasch Unit (RIT) scores. These scores are the level at which a student is able to answer 50% of the questions correctly, providing educators information regarding where to target instruction. MAP-R assessments are conducted twice per year, in the spring and fall.

The publishers of MAP-R conducted a study of the concordance between the performance levels of the PSSA using MAP-R's 2015 norming data (NWEA, 2015). The study established cut scores that correspond to the performance levels on the PSSA for grades three through eight. These performance levels correspond to the performance levels of the PSSA: advanced, proficient, basic, and below basic.

Table 3.2
Fall MAP-R Cut Scores for Grades 3–6.

Grade	Advanced	Proficient	Basic	Below Basic
3	209–350	185–208	164–184	100–163
4	213–350	198–212	175–197	100–174
5	223–350	206–222	183–205	100–182
6	225–350	211–224	188–210	100–187

Table 3.3
Spring MAP-R Cut Scores for Grades 3–6.

Grade	Advanced	Proficient	Basic	Below Basic
3	215–350	204–214	195–203	≤ 194
4	230–350	217–229	204–216	≤ 203
5	243–350	229–242	215–228	≤ 214
6	246–350	232–245	216–231	≤ 215

DIBELS Oral Reading Fluency

The DIBELS scoring utilizes three "cut points for risk" (Dynamic Measurement Group, 2010) described as "At or Above Benchmark," "Below Benchmark," and "Well Below Benchmark." Students who have scores in the range of at benchmark or above benchmark are described as having an 80–90% chance of achieving literacy goals. Students who score in the below benchmark range are reported as having a 40–60% chance of achieving literacy goals. Students who score in the Well Below Benchmark range are only 10–20% likely to reach their literacy goals and will need significant support. For DORF, proficiency is based on the number of words correctly read within one minute (WPM). Reading probes are leveled based on grade, resulting in different WPM expectations for different grades. For example, there is a lower WPM expectation for sixth graders than for fifth graders due to the more difficult words. The DORF cut scores are listed in the tables below.

Table 3.4
Fall DORF Cut Scores for Grades 3–6.

Grade	At or Above Benchmark	Below Benchmark	Well Below Benchmark
3	70 ≥ WPM	54–69 WPM	53 ≤ WMP
4	90 ≥ WPM	70–89 WPM	69 ≤ WPM
5	111 ≥ WPM	96–110 WPM	95 ≤ WPM
6	107 ≥ WPM	90–106 WPM	89 ≤ WPM

Table 3.5
Spring DORF cut scores for Grades 3–6.

Grade	At or Above Benchmark	Below Benchmark	Well Below Benchmark
3	100 ≥ WPM	80–99 WPM	79 ≤ WMP
4	115 ≥ WPM	95–114 WPM	94 ≤ WPM
5	130 ≥ WPM	105–129 WPM	104 ≤ WPM
6	120 ≥ WPM	95–119 WPM	94 ≤ WPM

Research design

Procedure

This quantitative research design was modeled after prior studies that examined the relationship between oral reading fluency and reading comprehension, using cross-tabulation data and calculation of indices of agreement including sensitivity, specificity and kappa (Baker et al., 2008; Barger, 2003; Buck & Torgesen, 2003; Crawford et al., 2001; Fuchs et al., 2001; Good, Simmons & Kame'enui, 2001; Good et al., 2002; Keller & Shapiro, 2005; Kim et al., 2005; McGlinchey & Hixson, 2004; Roehrig et al., 2007; Schilling et al., 2007; Shapiro et al., 2006; Shapiro et al., 2007; Shaw & Shaw, 2002; Stage & Jacobsen, 2001; Vander Meer et al., 2005; Wiley & Deno, 2005; Wilson, 2005; Wood, 2006). The current study expanded on these previously used procedures by examining additional indices of agreement, identified as the Improvement and the Stability Indexes and by applying Chi Square analyses, a nonparametric test of statistical significance. These procedures were used to examine the relationship between the MAP-R fall and spring administrations and the PSSA spring reading assessment, the relationship between the fall and spring administrations of the DORF and the PSSA reading assessment, and the relationship between the fall and spring MAP-R reading assessment and the fall and spring DORF assessment.

Prior to conducting the analyses, the MAP-R, DORF, and PSSA scores were transformed into dichotomous categorical scores. The PSSA performance level data are converted into the dichotomous score categories as follows: PSSA performance levels of below basic and basic were converted into the category of not proficient, and the performance levels of proficient and advanced were converted into the category of proficient. The MAP-R scores that have been converted into PSSA performance level equivalents were converted in a manner similar to that used with the PSSA performance levels. The DORF scores that were transformed into DIBELS benchmark categories were converted into dichotomous score categories by converting DIBELS at or above benchmark scores into the category of proficient and by converting DIBELS below benchmark and well below benchmark scores into the category of not proficient.

Frequency counts for the proficient and not-proficient categories are tabled for description. The frequency counts for the categories are used to calculate proficiency percentages and Sensitivity Indexes. The Sensitivity Index values are used to conduct chi-squared analyses to determine if differences in Sensitivity Index values are statistically significant.

Statistical Analyses

The cross-tabulation table values for each of these three comparisons were used to calculate values for the Improvement, Instability, Sensitivity, Specificity, and Kappa Indexes and to test the statistical significance of differences between the agreement percentages derived from the fall and spring administrations of the MAP-R, compared with PSSA results and the fall and spring administrations of the DORF with PSSA results.

Fall and spring DORF and MAP-R Index values were derived from the generation of 2 x 2 cross-tabulation tables using the formulas shown in Table 6.

Table 3.6
Construction of Crosstabulation Tables and Calculation Formulas Used to Derive the Index Values Used in Statistical Analyses of Data

DORF or MAP-R Fall or Spring Score Category	PSSA Score Category	
	Not Proficient	Proficient
At-Risk or Not Proficient	A	B
Not At-Risk or Proficient	C	D

Improvement Index = (B/(A+B)) x 100

Instability Index = (D/(C+D) x 100

Sensitivity Index = (A/(A+C)) x 100

Specificity Index = (D/(B+D)) x 100

Kappa = ((po-pe)/(1-e)) x 100 where:

Po = pA +pD

Pe = ((pA +pC)(pA+pB)) + ((pB +pD)(pC+pD))

pA=A/Total N pB=B/Total N pC=C/Total N pD=D/Total

In addition to the calculation and statistical analysis of the index scores, a descriptive analysis was completed to examine the relationship between fall and spring DORF and PSSA results and fall and spring MAP-R and PSSA results. To accomplish this descriptive analysis, status change categories were constructed as illustrated in Table 7.

A status change pattern was determined for each student by examining the category scores obtained on the fall and spring administrations of the DORF and the spring administration of the PSSA, and then categorizing patterns of status changes from fall DORF to spring DORF to spring PSSA. Students were assigned to categories based on the pattern of relationship among these three scores. Percentages of students exhibiting each status change pattern were calculated and tabled for descriptive analysis. This procedure was repeated for the fall and spring MAP-R category scores and the spring PSSA score. Status change patterns indicate how a student performed on the fall and spring administration either of DORF or of MAP-R and the PSSA. They are written with results in order of administration. For example, N - P - N would indicate that a student was not proficient on the fall administration, was proficient on the spring administration, and not proficient on the PSSA. A pattern of P – P – P would indicate that a student was proficient on both fall and spring CBA administrations and the PSSA, but a pattern of N – N – N would indicate that a student was not proficient on any of the administrations. A student who is not proficient on one or both of the CBA administrations is considered to be at-risk.

Operational definitions for terms as well as the indices and patterns used to analyze the data and interpret findings in this study are as follows:

At-Risk: A student is deemed to be "At-Risk" of failing to obtain a proficient score on the PSSA if the student obtains a not proficient score either on the fall or on the spring administration of a CBA.

Percentage of Students at Risk: The percent of students at risk is operationally defined as the percentage of students at risk of not being proficient based on the results of a DORF or MAP-R administration.

Improvement Index: The Improvement Index is operationally defined as the percentage of students categorized as not proficient on a DORF or MAP-R administration, but who were identified as proficient on the PSSA reading assessment. The Improvement Index represents the success rate of students identified as At-Risk of being Not Proficient on the PSSA.

Instability Index: The Instability Index is operationally defined as students who were identified as proficient on a DORF or MAP-R administration, but who conversely earned scores in the not-proficient range on the PSSA reading assessment during that same school year.

Sensitivity: Sensitivity is operationally defined as the proportion of students who were identified as not proficient on the PSSA but were also identified as not proficient on a DORF or MAP-R administration.

Specificity: Specificity is operationally defined as the proportion of students who were identified as proficient on the PSSA but were also identified as proficient on a DORF or MAP-R administration.

Kappa: The Kappa statistic indicates the percentage of increase over chance level represented by the overall percentage of agreement of DORF or MAP-R categories with PSSA results.

Performance Patterns and Categories: Performance patterns are based on the relationship among the score descriptive categories (Proficient or Not Proficient) assigned to a student's fall and spring administrations of the DORF or the MAP-R and the PSSA results. The eight possible performance patterns are shown in Table 3.2. The eight performance patterns are grouped into four performance categories: Consistently Not Proficient, Negative Change, Positive Change, and Consistently Proficient.

CHAPTER 4

RESULTS

The results of the statistical analyses conducted to address the research questions are presented in this chapter.

Research Question 1: Is the proportion of 4th grade students in the current research study sample that were identified as not proficient on the PSSA reading assessment significantly different from the proportion of the population of 4th grade students in the state of Pennsylvania that were identified as not proficient on the PSSA reading assessment?

It was predicted that the percentage of students who failed the PSSA would not significantly differ from statewide results. In order to determine this, a chi-squared test was conducted to compare the two groups. Results indicated that the difference between the two groups was not statistically significant (p = 0.6926). In this study, 37% of the fourth grade students in the sample earned scores in the not proficient range on the PSSA reading assessment. This is similar to the statewide results, for the same year; 39% of students statewide earned scores in the not proficient range on the PSSA reading assessment (PA Department of Education, 2017).

Research Question 2: What proportion of 4th grade students were identified as At-Risk of earning a Not Proficient category rating on the PSSA reading assessment based on fall and spring DIBELS Oral Reading Fluency (DORF) score categories?

Of the 93 fourth grade students who took DORF in the fall, 42 (45%) were identified as at risk of earning a not-proficient category rating on the PSSA reading assessment. In the spring, the percentage of students identified as at risk increased to 52%.

Research Question 3: What proportion of 4th grade students are identified as At-Risk of earning a Not Proficient category rating on the PSSA reading assessment based on the fall and spring Measures of Academic Progress Reading Assessment (MAP-R) score categories?

Of the 93 students who took MAP-R in the fall, 47 (51%) were identified as at risk of earning a not-proficient category rating on the PSSA reading assessment. In the spring, 41 students (44%) were identified as at risk of earning a not-proficient category rating on the PSSA reading assessment.

Research Question 4: What is the relationship between DORF score ratings (At-Risk/Not At-Risk) or MAP-R category ratings (Not Proficient/Proficient) and PSSA score categories (Proficient/Not Proficient) for 4th grade students?

Research Question 4a. What proportion of 4th grade students identified as At-Risk with fall and spring DORF and MAP-R earned PSSA scores in the Proficient range (operationally defined as the Improvement Index) and how do the fall and spring DORF and MAP-R compare as indicators of improvement?

It was hypothesized that fall and spring administrations of DORF and MAP-R would possess minimal differences in terms of ability to identify accurately those students at risk for scoring not proficient on the PSSA. The proportion of students identified as at risk with fall and spring DORF and MAP-R who earned PSSA scores in the proficient range is reported as the

Improvement Index. Improvement Index values calculated for the fall and spring administrations of the DORF and MAP-R are provided in Table 8.

Table 4.1
Improvement Index values for the fall and spring administrations of the DORF and MAP-R with a sample of 4th grade students (n = 93) Improvement Index

	Fall	Spring	Fall	Spring
Grade 4	DORF	DORF	MAP-R	MAP-R
Improvement	36%	42%	34%	32%

Fall administration of the DORF and MAP-R yielded Improvement Index values of 36% for the DORF and 34% for the MAP-R. A Chi-square analysis showed that there is no statistically significant difference between the Improvement Index values obtained with the fall administration of the DORF and the MAP-R (*p* = 0.7755).

Spring administrations of the DORF and MAP-R yielded Improvement Index values of 42% for DORF and32% for MAP-R. A Chi-square analysis indicated that there is no statistically significant difference between the Improvement Index values obtained with the spring administration of the DORF and MAP-R (*p* = 0.1590).

Research Question 4b: What proportion of students identified as Not At-Risk with fall and spring DORF and MAP-R earned PSSA scores in the Proficient range (operationally defined as the Instability Index) and how do the fall and spring DORF and MAP-R category scores compare as indicators of instability?

It was hypothesized that the Instability of each test would be very low. The Instability Index is defined as the percentage of students not identified as At-Risk on DORF or MAP-R who then went on to obtain a not proficient score on the PSSA. Instability Index values calculated for the fall and spring administrations of the DORF and MAP-R are provided in Table 9.

Table 4.2

Instability Index values for the fall and spring administrations of the DORF and MAP-R with a sample of 4th grade students (n = 93)

Instability Index	Fall	Spring	Fall	Spring
Grade 4	DORF	DORF	MAP-R	MAP-R
Instability	12%	11%	4%	10%

Fall administrations both of the DORF and of the MAP-R yielded very low Instability Index values. In the fall, the DORF results yielded an Instability Index of 12% and the MAP-R results yielded an Instability Index of only 4%. A Chi-square analysis indicated that there was a significant difference between the Instability Index values of the two assessments ($p = 0.0449$), with the DORF yielding a significantly higher Instability Index value than the MAP-R.

In the spring, the DORF results yielded an Instability Index of 11% and the MAP-R results yielded an Instability Index of 10% in the spring. A Chi-square analysis indicated that there was no significant difference between the two assessments in the spring ($p = 0.08244$).

Research Question 4c: What is the proportion of 4th grade students identified as At-Risk with fall and spring DORF and MAP-R who earned PSSA scores in the Not Proficient range (operationally defined as the Sensitivity Index) and how do fall and spring DORF and MAP-R category scores compare as indicators of sensitivity?

It was hypothesized that both DORF and MAP-R would yield comparable Sensitivity Index values. Sensitivity is defined as the percentage of students who were identified as At-Risk who also went on to score as not proficient on the PSSA. Sensitivity Index values calculated for the fall and spring administrations of the DORF and MAP-R are provided in Table 10.

Table 4.3
Sensitivity Index values for the fall and spring administrations of the DORF and MAP-R with a sample of 4th grade students (n = 93)

Sensitivity Index	Fall	Spring	Fall	Spring
Grade 4	DORF	DORF	MAP-R	MAP-R
Sensitivity	82%	85%	94%	85%

In the fall, among the students who went on to obtain not-proficient scores on the PSSA, 82% of students were identified as at risk on DORF, and 94% were identified as at risk, based on MAP-R scores. A Chi-squared analysis indicated a significant difference between the two assessments in the fall (*p* = 0.0120), with MAP-R demonstrating greater sensitivity. In the spring, both assessments had a Sensitivity Index of 85% and did not significantly differ (*p* = 1).

Research Question 4d: What is the proportion of 4th grade students that earned PSSA scores in the Proficient range who also were identified as Not At-Risk with the fall and spring DORF and MAP-R score categories (operationally defined as the Specificity Index) and how do the fall and spring DORF and MAP-R category scores compare as indicators of specificity?

It was hypothesized that both DORF and MAP-R would have comparable Specificity Index values. Specificity is defined as the percentage of students who were identified as not at risk and who also went on to score as proficient on the PSSA. Specificity Index values calculated for the fall and spring administrations of the DORF and MAP-R are provided in Table 11.

Table 4.4
Specificity Index values for the fall and spring administrations of the DORF and MAP-R with a sample of 4th grade students (n = 93)

Specificity Index	Fall	Spring	Fall	Spring
Grade 4	DORF	DORF	MAP-R	MAP-R
Specificity	75%	67%	73%	78%

In the fall, 75% of students who were proficient on the PSSA were not at risk on DORF. Of the students with proficient PSSA scores, 73% were identified as not at risk, based on MAP-R scores in the fall. A Chi-squared analysis indicated no significant difference between the specificity for the fall administration of the two assessments (p = 0.7565). In the spring, DORF had a Specificity Index of 67% and MAP-R had a Specificity Index of 78%, with no significant differences between the two measures (p = 0.0939).

Research Question 4e: What is the percentage of improvement over chance represented by the relationship between fall and spring DORF categories and PSSA reading score categories and the relationship between fall and spring MAP-R categories and PSSA reading score categories (operationally defined as the Kappa Index) for 4th grade students and how do the fall and spring DORF and MAP-R category scores compare as predictors of PSSA outcomes in terms of improvement over chance?

It was hypothesized that both DORF and MAP-R would have comparable Kappa proportions. Kappa is defined as the percentage of increase over chance level represented by the overall percentage of agreement between each measure and the PSSA. Kappa Index values calculated for the fall and spring administrations of the DORF and MAP-R are provided in Table 4.5.

Table 4.5
Kappa Index values for the fall and spring administrations of the DORF and MAP-R with a sample of 4th grade students (n = 93)

Kappa Index	Fall	Spring	Fall	Spring
Grade 4	DORF	DORF	MAP-R	MAP-R
Kappa	.54	.47	.61	.60

The relationship between the fall DORF and PSSA categories indicated a 54% improvement over chance (Kappa =, 54). The relationship between the fall MAP-R and the PSSA categories indicated 51% improvement over chance (Kappa = .51). A Chi-square analysis indicated no significant difference between the Kappa values of the DORF and the MAP-R (*p* = 0.3355). In the spring, the relationship between the DORF and PSSA produced a Kappa Index of .47 and the relationship between the MAP-R and the PSSA produced a Kappa Index of .60. A Chi-square analysis indicated no significant difference between the two spring DORF and MAP-R kappa values (*p* = 0.0763).

Research Question 5: How does success or failure on the DORF and the MAP-R relate to future PSSA reading assessment results?

Research Question 5a: What percentage of students classified as At-Risk were able to pass the PSSA? Was there a significant difference between the PSSA passage rates for students identified as At-Risk on either DORF or MAP-R?

Table 4.6 shows the percent of students at risk on the DORF and the MAP-R and the percent of those at-risk students that earned passing scores on the PSSA.

Table 4.6
Percent of students at risk and percent of those students at-risk that earned passing scores on the PSSA

Students at Risk	DORF	MAP-R
Percentage at Risk	57%	55%
At Risk Passage Rate	42%	49%

Of the 93 students in the sample, 53 (57%) were deemed at risk based on their DORF performance, and 40 (43%) students were to be not at risk. Of the at-risk students, 22 (42%) were able to obtain proficient scores on the PSSA. Of the not-at-risk students, 38 (95%) were able to obtain passing scores on the PSSA.

Of the 93 students in the sample, 52 (55%) were deemed at-risk based on their MAP-R performance, and 41 (44%) students were considered not at risk. Of the at-risk students, 20 (49%) were able to obtain proficient scores on the PSSA. Of the not at-risk students, 38 (95%) were able to obtain passing scores on the PSSA.

A Chi-squared analysis was conducted to compare the percentage of students who were deemed to be at risk based on their DORF and MAP-R performances, revealing no significant difference between the two groups (p = 0.7841). A chi-squared analysis was also conducted to compare the PSSA passage rate of students who were deemed to be at risk. There was no significant difference between the two groups (p = 0.4735).

In order to examine individual outcomes and the overall relationship between the DORF and the PSSA and the MAP-R and the PSSA, student performance data on fall and spring administrations of the DORF and MAP-R and the PSSA were sorted into performance patterns. The resulting patterns of performance are shown in Table 4.7.

Table 4.7
Student Performance Patterns

Performance Pattern Fall – Spring – PSSA	DORF Number of Students (n = 93)	MAP-R Number of Students (n = 93)
Consistently Non-Proficient	N	n
N* – N – N	24	27
Negative Change		
P*– N – N	4	1
P – P – N	2	1
N – P – N	3	4
Positive Change		
P – N – P	7	4
N – N – P	13	9
N – P – P	2	7
Consistently Proficient		
P – P – P	38	40

* *Student Performance Pattern N indicates a score that was not proficient; P indicates a score that was proficient.*

Among the 37 students who did not obtain proficient scores on either the fall or spring DORF administrations, 24 failed to obtain a proficient score on the PSSA. However, 13 students were able to obtain a proficient score. Among the 11 students who obtained proficient DORF scores in the fall, but not in the spring, seven were able to obtain proficient scores on the PSSA, and four were not proficient. A total of five students who obtained not-proficient scores in the fall DORF received proficient DORF scores in the spring. Among those five, two passed the PSSA. A total of 40 students received proficient DORF scores both in fall and in spring. Among those students, 38 were able to pass the PSSA.

On the MAP-R, among the 36 students who did not obtain proficient scores on either the fall or spring MAP-R administrations, 27 failed to obtain a proficient score on the PSSA; nine were able to obtain a proficient PSSA score. Among the five students who obtained proficient MAP-R scores in the fall, but not in the spring, four were able to obtain proficient scores on the PSSA, and four were not proficient. A total of 11 students who obtained not-proficient scores in the fall MAP-R received proficient MAP-R scores in the spring. Among those 11, seven passed the PSSA. A total of 41 students received proficient MAP-R scores in both fall and spring. Among those students, 40 were able to pass the PSSA.

Data were combined across the DORF and MAP-R as shown in Table 4.8 in order to examine student performance patterns more closely.

Table 4.8
DORF and MAP-R outcomes compared by PSSA outcome

Total Proficient Scores DORF and/or MAP-R	Proficient Score on PSSA	Not Proficient Score on PSSA
0	6	22

1	8	4
2	3	7
3	10	0
4	33	0
Total Passed and Failed	60	33

In the combined analysis, a total of 28 students obtained no proficient scores on the fall and spring administrations of either the DORF or MAP-R. Among these students, six (28%) were able to obtain proficient scores on the PSSA. A total of 12 students received a proficient score on only one administration of the DORF or MAP-R during the school year. Among those students, eight (67%) were able to obtain proficient PSSA scores. A total of 10 students obtained a proficient score on two administrations of the DORF or MAP-R during the school year. Among those students, three (30%) were able to obtain proficient PSSA scores. A total of 10 students obtained a proficient score on three of the four administrations of the DORF and MAP-R during the school year. Among those students, 10 (100%) were able to obtain proficient PSSA scores. Finally, 33 students obtained a proficient score on all four administrations of the DORF and MAP-R during the school year. Among those students, 33 (100%) were able to obtain proficient PSSA scores.

CHAPTER 5

DISCUSSION

Summary of Findings

In this study, the utilized sample demonstrated a high goodness of fit regarding the percentage of students who were not proficient on the PSSA, compared with statewide results. This suggests that the sample in the study is representative of the statewide fourth grade PSSA results.

In the fall, 45% of students were identified as at risk of earning a not-proficient score on the PSSA, based on DORF scores. In the spring, the percentage was similar, with 52% of students identified as at risk, based on DORF scores. This indicates that although students made progress, the proportion of students at risk increased slightly, meaning they did not keep pace with DORF expectations.

Of the students who took MAP-R in fall and spring, the number identified as at risk remained stable across the two administrations. This suggests that students at risk in the fall, although making some progress, are not able to make sufficient progress to bring them out of the at-risk category.

Both DORF and MAP-R appear to measure similar rates of improvement. In other words, both assessments demonstrated similar percentages of students who were identified as at risk for failure on the PSSA, but were able to pass the PSSA. For both DORF and MAP-R, it appears that it is unusual for a student to pass either assessment and then fail to obtain a proficient score on the PSSA. Although there is a statistically significant difference between DORF and MAP-R in the fall, the relatively small sample size of the study makes it difficult to determine any meaningful difference between the measures.

There was a significant difference between the sensitivity of the two measure in the fall. Fall administrations of DORF over-predicted students being at risk of failing the PSSA. This may be a strength of DORF because it may be less likely to miss students who may benefit from reading interventions. In the spring, both DORF and MAP-R had the same Sensitivity Index values.

There were no statistically significant differences in the measures in their specificity in accurately predicting PSSA proficiency. However, there was a substantial percentage of students who were misclassified and were thought to be at risk but passed the PSSA. Conversely, there were few students thought not to be at risk that went on to be not proficient on the PSSA. This indicates that there remains a notable percentage of students whose PSSA outcomes were not accurately identified either by DORF or by MAP-R, but that the majority of students thought to be not at-risk were indeed able to obtain a proficient score on the PSSA. DORF and MAP-R appear to have similar rates of agreement with PSSA scores, above chance.

Both DORF and MAP-R demonstrated similar percentages of students who fell into the at-risk category. Students in the at-risk category were able to pass the PSSA at similar rates. This indicates that neither of the measures over or under identifies students at risk.

For both DORF and MAP-R, passing both fall and spring administrations had a notable impact on a student’s likelihood of passing the PSSA. Very few students who passed both fall and spring administrations either of DORF or of MAP-R went on to obtain a not proficient score on the PSSA. Conversely, the majority of students who did not obtain at least one proficient score on either the fall or spring administration were much less likely to pass the PSSA.

When student data from both DORF and MAP-R are combined, a clearer pattern emerges. Of the 43 students who obtained three or more proficient scores on either DORF of

MAP-R, 100% were able to obtain proficient scores on the PSSA. Of the 28 students who were unable to obtain proficient scores on any of the administrations of DORF or MAP-R, only six were able to pass the PSSA. These results are consistent with other research in which CBA scores were combined (Klingbeil, McComas, Burns, & Helman, 2015).

Significance of the Findings

The major focus of this study was on comparing the ability of two reading assessments, DIBELS DORF and the MAP-R, in order to identify students who were likely to fail to obtain proficient scores on the PSSA. This is important due to the high-stakes nature of the PSSA and the need to reassess the relationship between CBAs and the PSSA following the changes to the test implemented in 2015.

Overall, there appear to be few differences in the ability of these two measures to identify students at risk of failure on the PSSA. Both measures have demonstrated approximately equivalent efficacy based on several indices. For the Improvement Index (percent of students categorized as not proficient on either DORF or MAP-R, but who were identified as proficient on the PSSA), there was no difference in either the fall or spring administrations. For the Instability Index (students who were identified as proficient either on DORF or on MAP-R and earned scores in the not-proficient range on the PSSA), there was a small difference in the fall, and no difference in the spring. Fewer students who were identified as not at risk on the MAP-R in the fall went on to fail the PSSA. In other words, the MAP-R may be slightly more effective at accurately identifying students who were at risk. For the Sensitivity Index, the fall administration of MAP-R was slightly more likely to identify accurately those students who would go on to obtain not proficient scores on the PSSA. This is an important finding because the Sensitivity Index is very highly relevant for identifying students in need of remediation. The two measures were not significantly different in their Specificity Indices (proportion of students who were identified as proficient on the PSSA who also were identified as proficient on DORF or MAP). Interestingly, both measures appear to over-identify students who may be at risk, particularly in the fall. This is not a surprising finding, given the fact that they are administered far in advance of the PSSA, and students are likely to receive interventions that reduce the likelihood of failure. The Kappa Index values (a comparison between the two measures that accounts for chance) were also equivalent. These results are consistent with previous research on the relationship between oral reading fluency and reading comprehension (Gardner, Rudman, Karlsen, & Merwin, 1982; L.S. Fuchs, Fuchs, and Maxwell, 1988; Goffreda, Diperna, and Pedersen, 2009; Reidel, 2007).

When examining the two CBAs, there appear to be no significant differences in the PSSA passage rate among students who were identified as being at risk. These students were much less likely to pass the PSSA than students not identified as being at risk, regardless of which CBA was used to identify them. It appears that the relationship between the PSSA and CBA results grows significantly stronger when scores from DORF and MAP-R are combined. In this study, when scores are combined, students who obtained proficient scores on three or more CBA administrations were able to obtain proficient scores on the PSSA 100% of the time.

Impact of the Findings

The results of this study suggest that there are minimal differences between DORF and MAP-R with regard to predicting PSSA outcomes. However, there are significant differences in the cost and methods of administration between the two measures. The cost of DORF is substantially lower per student and takes far less time to administer. This study does not find significant evidence that the added cost of MAP-R leads to greater efficacy. School districts with limited resources may benefit from this finding in making decisions regarding which assessments to purchase. The time required to administer each assessment is also very different. DORF probes are time limited at one minute each. This is in contrast to the MAP, which can take up to 30 minutes per student according to its publisher (though, anecdotally, this author has heard from teachers who reported that the test can take up to 90 minutes). Although MAP can be distributed to groups of students, thereby reducing the overall time spent by the administrator, it remains a substantial undertaking, particularly for young students with short attention spans. Testing can be a stressful, expensive, and time-consuming process. Educators must strike a balance between efficiency and efficacy when making decisions regarding which assessments to utilize.

Another impact of the study is to lend support to the empirical base of the MAP. The MAP has very little published research outside of studies conducted by its developer; however, DIBELS has an extensive and diverse research base. This study adds to the relatively small body of literature examining the efficacy of MAP compared with other CBAs.

Although DORF and MAP-R appear to be essentially equivalent when used in isolation, they appear to have increased efficacy when combined. Combined results were more likely to predict accurately which students were not at risk, compared with using only one CBA. This is consistent with previous research examining the use of multiple measures to assess student achievement (Klingbeil, McComas, Burns, & Helman 2015).

School psychologists and other educators may find this and other similar studies useful in analyzing data from RTI programs. Understanding that poor ORF can have wide-reaching implications for student achievement is important for decision making and recommending interventions.

Limitations

There are several limitations to this study. This study utilized data from a single district in one state, which may have unique characteristics that can limit the generalizability of the study. This study also studied only two of the many CBAs used for monitoring reading progress and predicting reading problems. Furthermore, the study examined the relationship between two CBAs and the current version of the PSSA. Pennsylvania may continue to make changes to the PSSA that could result in further differences in student results. The study also did not examine the nature of interventions employed for students identified with reading problems. Certain interventions may result in different performance outcomes for different measures.

This study also was not able to determine the reason why a small percentage of students who were thought not to be at risk went on to obtain not-proficient scores on the PSSA. This could be due to a number of idiosyncratic factors, including motivation, health, family issues, or other unknown factors. It also did not consider demographic factors such as special education eligibility. This dataset did not differentiate between students in special education and those who do not have a disability.

This study was also limited by its relatively small sample size and narrow population. The study utilized 93 fourth grade students; although their PSSA passage rates appeared to be comparable with the state as a whole, the small number of students leaves these results more open to changes due to chance. Furthermore, the study may not generalize to students in other grades. This may be particularly true of higher grades, where reading content is more complex and abstract. With higher levels of complexity, reading fluency’s ability to predict comprehension seems likely to be reduced.

Another limitation of this study was its use of categorical data. Although categorical data is useful in terms of its simplicity, it precludes a level of nuance that could be helpful in analyzing scores. All three assessments examined for this study employ categorical cut-off scores, but also have continuous scores within each score category. In other words, a student with a score one point below proficient cannot be differentiated from a student with the lowest possible score, and so on.

Future Directions

Future research is needed to create a robust body of literature concerning the efficacy and overall strengths and weaknesses of various CBAs. DORF and MAP-R appear to have largely overlapping ability to predict PSSA results; however, there may be other assessments that are better suited for that role. Furthermore, there may be combinations of CBAs that, when used in conjunction, yield increased accuracy. Additional research is also need to determine if different CBAs are better able to aid in identifying specific interventions that possess a greater chance of moving from at-risk status.

Future studies would benefit from larger sample sizes that include more grade levels and more descriptive information on the characteristics of the students, such as socioeconomic status or special education classification. A study conducted with continuous data may also improve the ability to identify students that are on the cusp of proficiency or are likely to move from a not-at-risk status to an at-risk status.

References

Adams, M. J. (1994). *Beginning to read: Thinking and learning about print*. MIT press.

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

Baker, S. K., Smolkowski, K., Katz, R., Fien, H., Seeley, J. R., Kame'enui, E. J., Beck, C. T.(2008). Reading fluency as a predictor of reading proficiency in low-performing, high poverty schools. *School Psychology Review, 17*, 18-37. Retrieved from http://www.naspweb.org/publications/index.html.

Barger, J. (2003). *Comparing the DIBLES oral reading fluency indicator and the North Carolina end of grade reading assessment* (Technical Report). Asheville, NC: NC Teacher Academy.

Buck, J., & Torgesen, J. (2003). *The relationship between performance on a measure of oral reading fluency and performance on the Florida Comprehensive Assessment Test* (Technical Report No. 1). Tallahassee, FL: Florida State University, Florida Center for Reading Research.

Breaux, K. C. (2009). WIAT-III technical manual. San Antonio, TX: NCS Pearson.

Cavallo, F. (2012). *Predicting student responsiveness to Fast ForWord using DIBELS subtests*. Temple University.

Cordray, David S., Georgine M. Pion, Chris Brandt, and Ayrin Molefe. (2013). *The Impact of the Measures of Academic Progress (MAP) Program on Student Reading Achievement*. Society for Research on Educational Effectiveness.

Crawford, L., Tindal, G., & Stieber, S. (2001). Using oral reading rate to predict student performance on statewide achievement tests. *Educational Assessment, 7,* 303-323doi:10.1207/S15326977EA0704_04

CTB/McGraw-Hill. (2005). Product detail: TerraNova, the second edition (CAT/6). Retrieved from http://www.ctb.com.

Deno, S. L. (1985). Curriculum-based measurement: The emerging alternative. *Exceptional Children*, 52, 219-232.

Dorshimer, B. L. (2009). Curriculum-based measures of reading and student performance on statewide achievement tests. Philadelphia College of Osteopathic Medicine (Doctoral Dissertation). Retrieved from http://www/pcom.edu/library

Durkin, D. (1993). *Teaching them to read (6th ed.).* Boston, MA: Allyn & Bacon.

Flindt, N. (2004). *Technical Adequacy of DIBELS: Results of the early childhood Research institute on measuring growth and development* (Technical Report, No. 7). Eugene, OR: University of Oregon.

Fuchs, L. S., Fuchs, D., Hosp, M. K., & Jenkins, J. R. (2001). Oral reading fluency as indicator of reading competence: A theoretical, empirical, and historical analysis. *Scientific Studies of Reading, 5*(3), 239-256. doi:10.1207/S1532799XSSR0503_3

Fuchs, L. S., Fuchs, D., & Maxwell, L.(1988). The validity of informal measures of reading comprehension. *Remedial and Special Education*, 9(2), 20–28.

Gardner, E. F., Rudman, H. C., Karlsen, B., & Merwin, J. C. (1982). *Stanford Achievement Test.* Iowa City, IA: Harcourt Brace Jovanovich.

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.

Goo, M., Watt, S., Park, Y., & Hosp, J. (2012). A guide to choosing web-based curriculum-based measurements for the classroom. *Teaching Exceptional Children*, 45, 34-40.

Good, R.H., Kaminski, R.A., Shinn, M., Bratten, J., Shinn, M., Laimon, D., Smith, S., & Wallin, J. (2002). *Summary of decision rules for intensive, strategic, and benchmark instructional recommendations in kindergarten through third grade* (Technical Report No. 11). Eugene: University of Oregon.

Good, R. H., & Kaminski, R. A. (Eds.). (2002). *Dynamic indicators of basic early literacy skills (6th ed.).* Eugene, OR: Institute for the Development of Educational Achievement.

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

Heinemann Publishing. (n.d.) *Fountas and Pinnell reading assessment resources*. Retrieved from http://www.heinemann.com/fountasandpinnell/researchBAS.aspx

January, S. A. A., & Ardoin, S. P. (2015). Technical adequacy and acceptability of curriculum-based measurement and the Measures of Academic Progress. *Assessment for Effective Intervention*, *41*(1), 3-15.

Jenkins, J. R., Fuchs, L. S., Espin, C., van den Broek, P., & Deno, S. L. (2000). *Effects of task format and performance dimension on word reading measures: Criterion validity, sensitivity to impairment, and context facilitation*. In Pacific Coast Research Conference, San Diego, CA.

Keller, M. A., & Shapiro, E. S. (2005). *General outcome measures and performance on standardized tests: An examination of long-term predictive validity*. Paper presented at the meeting of the National Association of School Psychologists Convention, Atlanta, GA.

Kim, Y., Petscher, Y., Schatschneider, C., & Foorman, B. (2010). Does growth rate in oral reading fluency matter in predicting reading comprehension achievement? *Journal of Educational Psychology, 102*, 652-667. doi:10.1037/a0019643

Klingbeil, D. A., McComas, J. J., Burns, M. K., & Helman, L. (2015). Comparison of predictive validity and diagnostic accuracy of screening measures of reading skills. *Psychology in the Schools*, 52(5), 500-514.

Layton, L. (2015, October 24). Study says standardized testing is overwhelming nation's public schools. The Washington Post. Retrieved from https://www.washingtonpost.com/local/education/study-says-standardized-testing-is-overwhelming-nations-public-schools/2015/10/24/8a22092c-79ae-11e5-a958-d889faf561dc_story.html

Lucas, Joseph H. (2013). *Using math curriculum-based assessments to predict student performance on the Pennsylvania System of School Assessment math test* . PCOM Psychology Dissertations. 244. https://digitalcommons.pcom.edu/psychology_dissertations/244

Mather, N., & Woodcock, R. W. (2001b). *Woodcock-Johnson® III tests of cognitive abilities: Examiner's manual standard and extended batteries*. Itasca, IL: Riverside.

McCorry, K. (2015, July 14). Pa. says 2015 standardized test scores dropped precipitously because of added rigor. WHYY. Retrieved from https://whyy.org/articles/pa-says-2015-standardized-test-scores-dropped-precipitously-because-of-added-rigor/

McGlinchey, M. T., & Hixson, M. D. (2004). Using curriculum-based measurement to predict performance on state assessments in reading. *School Psychology Review, 33*(2), 193-203 Retrieved from http://www.nasponline.org/publications/spr/sprissues.aspx

Merino, K., & Beckman, T. O. (2010). Using reading curriculum-based measurements as predictors for the Measures of Academic Progress (MAP) standardized test in Nebraska. *International Journal of Psychology: A Biopsychosocial Approach*, 6, 85–98.

National Center for Education Statistics (2015). The Nation's Report Card: Reading. Institute of Education Sciences, U.S. Department of Education, Washington, D.C. https://www.nationsreportcard.gov/reading_math_2015/#reading?grade=4

National Reading Panel (U.S.), & National Institute of Child Health and Human Development (U.S.). (2000). *Report of the National Reading Panel: Teaching children to read: an evidence-based assessment of the scientific research literature on reading and its implications for reading instruction: reports of the subgroups*. Washington, D.C.: National Institute of Child Health and Human Development, National Institutes of Health.

Northwest Evaluation, Association. (2016). *Linking the Pennsylvania PSSA Assessments to NWEA MAP Tests*. Northwest Evaluation Association.

Northwest Evaluation, Association. (2013). *Measures of Academic Progress: A comprehensive guide to the MAP K–12 computer adaptive interim assessment*. Northwest Evaluation Association.

Northwest Evaluation Association. (2004). *Reliability and validity estimates: NWEA achievement level tests and measures of academic progress*.

Nouwens, S., Groen, M. A., & Verhoeven, L. (2017). How working memory relates to children's reading comprehension: The importance of domain-specificity in storage and processing. *Reading and Writing: An Interdisciplinary Journal*, 30(1), 105-120.

Pearson Education. (2008). *AIMSweb.* San Antonio, TX.

Pennsylvania Department of Education. (2017). 2017 PSSA state level data. Retrieved from http://www.education.pa.gov/data-and-statistics/PSSA/Pages/default.aspx

Pennsylvania Department of Education. (2009). Pennsylvania System of School Assessment: 2009-2010 assessment handbook. Harrisburg, PA: Retrieved from www.education.state.pa.us

Pennsylvania State Data Center. (2018). Special education data report: LEA performance on state performance plan (SPP) targets, school year 2016-2017. Retrieved From https://penndata.hbg.psu.edu/penndata/documents/BSEReports/Public%20Reporting/2016_2017/PDF_Documents/Speced_Data_Report_SD374_Final.pdf

Pollard, J. R. (2015). *The predictive strength of the DIBELS Next ORF assessment to the WIAT-III ORF and reading comprehension subtests for students referred for special education eligibility evaluations*. Indiana University of Pennsylvania.

Reschly, A. L., Busch, T. W., Betts, J., Deno, S. L., & Long, J. D. (2009). Curriculum-based measurement oral reading as an indicator of reading achievement: A meta-analysis of the correlational evidence. *Journal of School Psychology*, 47427-469. doi:10.1016/j.jsp.2009.07.001

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

Roehrig, A. D., Petscher, Y., Nettles, S. M., Hudson, R. F., & Torgesen, J. K. (2007). Accuracy of the DIBELS oral reading fluency measure for predicting third grade reading comprehension outcomes. *Journal of School Psychology, 46*, 343-366.doi:10.1016/j.jsp.2007.06.006

Samuels, S.J. (2007). The DIBELS tests: Is speed of barking at print what we mean by reading fluency? *Reading Research Quarterly*, 42(4), 563-566.

Scherr, R. (2011). *The efficacy of the MAP reading probe as a predictor of performance in a three-tiered system of reading interventions*. Philadelphia College of Osteopathic Medicine (Doctoral Dissertation).

Shapiro, E.S., & Clements, N.H. (2009). A conceptual model for evaluating system effects of response to intervention. *Assessment for Effective Intervention*, 35, 3-16.

Shapiro, E. S., Keller, M. A., Lutz, J. G., Santoro, L.E., & Hintze, J. M. (2006). Curriculum based measures and performance on state assessment and standardized tests: Reading and math performance in Pennsylvania. *Journal of Psychoeducational Assessment, 24*, 19-35.doi:10.1177/0734282905285237

Shapiro, E.S., Solari, E., & Petscher, Y. (2008). Use of a measure of reading comprehension to enhance prediction on the state high stakes assessment. *Learning and Individual Differences,* 18, 316-328.

Shaw, R., & Shaw, D. (2002). *DIBELS oral reading fluency-based indicators of third grade reading skills for Colorado State Assessment Program (CSAP)*. (Technical Report). Eugene, OR: University of Oregon.

Stage, S., & Jacobsen, M. (2001). Predicting student success on a state-mandated performance based assessment using oral reading fluency. *School Psychology Review, 30*, 407-419. Retrieved from http://www.naspweb.org/publications/index.html

The Editorial Board (2013, July 13). The trouble with testing mania. The New York Times Retrieved from https://www.nytimes.com/2013/07/14/opinion/sunday/the-trouble-with-testing-mania.html

U.S. Department of Education (2015). Every Student Succeeds Act (ESSA) of 2015, Pub.L. 114-95 (2015). Retrieved from https://edworkforce.house.gov/uploadedfiles/every_student_succeeds_act_-_conference_report.pdf

U.S. Department of Education (2001). No Child Left Behind Act of 2001, Pub. L. No. 107-110. Retrieved from http://www.ed.gov/nclb

VanDerHeyden, A. M., & Burns, M. K. (2010). *Essentials of response to intervention (Vol. 79).* John Wiley & Sons.

Vander Meer, C. D., Lentz, F. E., & Stollar, S. (2005). *The relationship between oral reading fluency and Ohio proficiency testing in reading*. (Technical Report). Eugene, OR: University of Oregon.

Weinstein, Elana E., (2011) A study of the relationship between elementary school students' performance on progress monitoring measures of oral reading fluency and reading comprehension to the Pennsylvania System of School Assessment. Philadelphia College of Osteopathic Medicine (Doctoral Dissertation).Wilson, J. (2005). *The relationship of Dynamic Indicators of Basic Early Literacy Skills(DIBELS) oral reading fluency to* *performance on Arizona Instrument to Measure Standards (AIMS).* (Research Brief). Tempe, AZ: Assessment and Evaluation Department of the Tempe School District No. 3.

Wood, D. E. (2006). Modeling the relationship between oral reading fluency and performance on a statewide reading test. *Educational Assessment, 11*, 85-104. doi:10.1207/s15326977ea1102_1