Western Kentucky University

TopSCHOLAR®

Masters Theses & Specialist Projects | Graduate School

Spring 2020

The Predictive Validity of STAR Early Literacy

Karlissa Pollack
Western Kentucky University, k4pollack@gmail.com

Follow this and additional works at: https://digitalcommons.wku.edu/theses

 Part of the Child Psychology Commons, Educational Psychology Commons, Language and Literacy Education Commons, and the School Psychology Commons

Recommended Citation

Pollack, Karlissa, "The Predictive Validity of STAR Early Literacy" (2020). *Masters Theses & Specialist Projects.* Paper 3191.
https://digitalcommons.wku.edu/theses/3191

This Thesis is brought to you for free and open access by TopSCHOLAR®. It has been accepted for inclusion in Masters Theses & Specialist Projects by an authorized administrator of TopSCHOLAR®. For more information, please contact topscholar@wku.edu.

THE PREDICTIVE VALIDITY OF STAR EARLY LITERACY

A Specialist Project
Presented to
The Faculty of the Department of Psychology
Western Kentucky University
Bowling Green, Kentucky

In Partial Fulfillment
Of the Requirements for the Degree
Specialist in Education

By
Karlissa Pollack

May 2020

THE PREDICTIVE VALIDITY OF STAR EARLY LITERACY

Date Recommended 4/29/2020

Sarah Ochs, Ph.D., Director of Specialist Project

Carl Myers, Ph.D.

Kristy Cartwright, M.A.

Cheryl D Davis Digitally signed by Cheryl D Davis
Date: 2020.06.01 12:03:01 -05'00'

Dean, The Graduate School Date

ACKNOWLEDGMENTS

I would like to express my gratitude to my thesis director, Dr. Sarah Ochs, who has always guided and supported me when I needed it most. She is unfailingly caring and reliable, which was the main foundation that allowed me to successfully complete this project and to believe in my professional self.

I would also like to thank Dr. Carl Myers, not only for his wisdom and careful attention to detail, but for being a brilliant mentor and role model. He continued to teach me when I was at my worst, and will continue to inspire me when I am at my best.

I would also like to express my appreciation to Mrs. Kristy Cartwright, whose unique perspective and insight was incredibly interesting and useful for this project. I am grateful for her valuable comments and guidance with this topic.

Finally, I would like to thank my parents, Matthew, and my Eklipse family for being a constant wall of support and care. Throughout my years of studying and through this process of writing this thesis, they have all been a source of strength and encouragement that allowed all of my accomplishments to be possible. Thank you to all of you.

CONTENTS

LIST OF TABLES

THE PREDICTIVE VALIDITY OF STAR EARLY LITERACY

Karlissa Pollack May 2020 37 Pages

Directed by: Sarah Ochs, Carl Myers, and Kristy Cartwright

Department of Psychology Western Kentucky University

In recent years, paper-and-pencil screenings have, to a degree, given way to computer adaptive tests as a more effective way to screen students, though some measures lack research in regards to their utility. The utility of Star Early Literacy (SEL) for predicting future state test performance was examined by analyzing student performance on SEL and student performance on state testing two years later. The following research questions were examined: What is the relationship among first grade SEL scores and performance on the third grade reading state test? What is the diagnostic accuracy of SEL for predicting reading state test performance two years later? Results indicated that while SEL scores are moderately correlated with state reading test scores, accuracy is only fair. Implications of these findings and future directions on this topic of research are further discussed.

Introduction

Early Literacy

Early literacy skills are the precursor skills that develop toward the beginning of achieving literacy. Specifically, early literacy is an ongoing process towards development of a child's skill to talk, listen, write, and interact with written language (Gillett, 2016). The development of an understanding of language such as verbal and non-verbal communication also supports the development of early literacy skills (Gillett, 2016). Children begin to understand the patterns and systems related to the sounds and structure of spoken languages, and eventually this understanding begins to translate into understanding the relationship between these sounds and visual stimuli (National Center for Family Literacy, 2008).

Some of the essential skills that contribute to the development of literacy are phonemic awareness, phonics, fluency, vocabulary, and text comprehension. (National Reading Panel, 2000) Phonemic awareness is the ability to understand and utilize phonemes, or distinct units of sound that are found in spoken word. An example of a phoneme is the sound of the letter "t", usually expressed within text as /t/. Phonics involves the understanding of the relationship between letters and their respective sounds, such as knowing that the /t/ sound is produced by the letter "t". Fluency is the ability to read with speed and accuracy as well as prosody, such as when a person pauses at commas during reading. Vocabulary is the understanding of the relationships between words and concepts/objects, such as knowing that the word "car" corresponds to the physical object "car" and means a vehicle. Text comprehension is the ability to understand the words and concepts presented within text. This means that if a student

reads the statement “The car was red,” they would understand the meaning of these words combined together and could then comprehend that the car mentioned within the text is red (National Reading Panel, 2000).

These skills can be categorized into two domains known as outside-in and inside-out components (Coyne & Harn, 2006). Outside-in includes mainly oral language skills, including semantic, narrative, and conceptual skills. Inside-out involves coding-based skills such as phonemic awareness and phonics. Both of these components are valuable to the development of reading, but inside-out skills are the most critical for early reading learning in kindergarten and first grade. However, the origins of inside-out skills are also relatively unknown in research, as outside-in skills are the skills typically measured and the beginnings are more easily traced to roots in other skills.

The essential skills of literacy also fall under three general abilities: oral language, phonological processing, and print knowledge (Lonigan, Burgess, & Anthony, 2000). Oral language involves being able to speak and comprehend individual sounds and their connections to spoken words, along with what those words represent, which encompasses vocabulary skills. It also relates to phonological sensitivity, which is the ability to perceive and understand the structure of oral language through syllables, blended phonemes, and other pieces of words and sounds. This encompasses phonemic awareness, as it begins the understanding of how sounds relate and blend together. A majority of poor readers are found to have a phonological deficit, regardless of their cognitive ability (Lonigan et al., 2000). Children who have good phonological sensitivity are capable of learning how to read quicker, even when cognitive ability, receptive vocabulary, memory skills, and social class are controlled as variables. Print knowledge

involves understanding the reading process, distinguishing between pictures and text, and understanding the purpose of the print. This closely associates with some aspects of phonics, but also text comprehension and fluency (Lonigan et al., 2000).

Phonological sensitivity and letter knowledge (a part of print knowledge) account for 54% of the variance in the decoding abilities of kindergarteners and first graders. Oral language abilities also had direct and indirect effects, but it was measured to be a key part of the development of phonological processing skills. There has been a debate in research whether the impact of oral language is only felt within its ability to develop phonological sensitivity, or if it continues to play a key role in the development of reading. The perspective of the phonological sensitivity approach expresses that oral language is the basis for phonological sensitivity which then becomes the basis for language skills (Dickinson, McCabe, Anastasopoulos, Peisner-Feinberg, & Poe, 2003). The other viewpoint is the comprehensive language approach which notes that the effects of oral language are more widespread because it affects emerging literacy directly and continues to play a role in reading development even after phonological sensitivity develops.

The International Reading Association and the National Association for the Education of Young Children (NAEYC) created a joint position statement emphasizing the importance of shifting the then-current trajectory of reading instruction. These organizations detail the importance of several milestones that build the aforementioned skills. Among these milestones are identifying letters, identifying sounds of their language, understanding the relationship between letters and sounds, and connecting words to meaning (NAEYC, 1998). Research has analyzed the link between early learning and later achievement, and the findings indicate that children that develop more

skills in their younger years (e.g., preschool) are more likely to perform better in the primary grades (National Center for Family Literacy, 2008). This is because early literacy skills have a strong relationship with later literacy skills, such as writing and spelling (NCFL, 2008).

The age of a student (younger classmates in preschool and kindergarten vs. older) has been controversially considered to be a risk factor for their school readiness, and therefore, their readiness to read in an academic setting (Crone & Whitehurst, 1999). Older children within each grade were found to have better literacy skills than younger peers in the same grade before they were enrolled in formal schooling and throughout kindergarten. However, by the end of first grade, after being exposed to formal reading instruction for the year, these differences disappeared, and the reading performance of both younger and older children were on the same level. The most important attribute was reading instruction in first grade, which allowed a large improvement in reading skills between first and second grade, with 81% of the student's improvement being attributed to formal schooling. A student's environment can also affect their ability to learn basic literacy skills. Low-income elementary school students that lacked basic literacy skills in the first grade were likely to continue to experience problems in the fourth grade (Greenfield-Spira, Bracken, & Fischel, 2005). The probability of this likelihood occurring was as high as .88 (Juel, 1988). The more severe the student's problem, the more likely their problem grew worse as time progressed.

Early literacy is a crucial foundational skill that supports the development of subsequent academic skills in reading, writing, and math. As such, it is critical that we

find efficient and accurate ways to measure early literacy in order to identify students who may need targeted intervention.

Screening

Universal screening is a method of evaluation used to identify individuals within a population that are potentially at risk for having a certain condition or falling under or over an expected threshold, depending on what is being measured (National Center on Response to Intervention, 2010). For example, a universal screener for depression screens individuals for the likelihood that they are at risk of having the condition. Academically, a universal math screener may screen individuals to determine if their mathematical abilities are falling below expected levels, which may indicate a student needs additional intervention, or are measured above expected levels, which may indicate giftedness. These processes are common in many fields, though perhaps the most well-known is the use of screening in the medical field, such as cancer screenings. Regardless, all types of screening serve the same purpose—they are administered to everyone within a population in order to identify the potential need for further evaluation or intervention within a smaller group or subset of the population. The key is to address any potential problems early in order to prevent these problems from growing in the future (National Center on Response to Intervention, 2010). In the field of school psychology, universal screening is an essential component of multi-tiered systems of supports, as it provides a way to identify students that are potentially at risk for a myriad of problems, such as reading or math difficulties. The process of screening in the schools is very brief and conducted with all students at least two or three times per year in order to identify students with the potential for poor learning outcomes (National Center on Response to Intervention

(NCRTI), 2010). The measures used must be valid, age-appropriate, and reliable to get the most accurate assessment of a student's potential for risk.

The need for screening within an academic context partially arose from the development of the Response to Intervention (RTI) method of identifying children that fall below an expected level of performance to provide early intervention to these children that are at risk for experiencing academic difficulties in school (NCRTI, 2010). With the RTI model, the approach to screening focused on prevention and intervention rather than diagnosis or classification, and its multi-tiered design allowed for students to receive intervention at varying levels of intensity. From the first tier providing school-wide, simple interventions, to the top tier that individualizes and personalizes treatments for a specific student's needs, attempts are made to improve even minor academic issues before they progress into becoming larger problems (NCRTI, 2010).

As a part of education, the practice of screening started to become prevalent after the re-authorization of the Individuals with Disabilities Education Improvement Act (Fuchs & Fuchs, 2005; IDEA, 2004). Previous to this law, practitioners referred to the discrepancy between IQ scores and a student's achievement level to identify students with learning disabilities, sometimes referred to as a "wait to fail" model due to students having to meet the discrepancy criteria before receiving services. Screening can allow educators to identify children that may need additional instruction to reach the level of achievement they should be reaching at a particular grade or time of year. In this way, the student can be identified as needing more instruction, so that they can get help to reach their full potential, early. Without screening, it would be substantially more difficult to

identify whether a child is struggling or not, what they are having difficulty with, and what area of ability to target to best help the child to succeed.

Curriculum-based measures. Curriculum-based measures (CBMs) are among one of the more popular choices for screening tools (Shinn, 1998). They are capable of reliably assessing student achievement within each basic academic area, such as reading, math, and writing. For reading CBM (R-CBM), students are asked to read aloud a passage within a specific time limit, while the administrator notes errors to assess reading accuracy and words read per minute. The validity of the measure was assessed via correlations with criterion, norm-referenced measures, specifically the Stanford Diagnostic Reading Test, the Woodcock Reading Mastery Test, and the Reading Comprehension subtest from the Peabody Individual Achievement Test (Shinn, 1998). The correlation coefficient of this criterion-related validity ranged from .73 to .91, though most correlations were above .80. Another correlation coefficient between the CBM and published measures of global reading skills ranged from .63 to .90, with most coefficients above .80. Test-retest reliability coefficients ranged from .82 to .97, though most correlations were above .90 (Shinn, 1998). The correlations can be interpreted utilizing standards from Schober, Boer, & Schwarte (2018), where .40 to .69 is a moderate correlation, .70 to .89 is a strong correlation, and .90 to 1.00 is a very strong correlation. With most CBM correlations above .80, it correlates well with measures of global reading skills. Its test-retest reliability also has a strong to very strong correlation, meaning CBM is consistent in producing similar scores with repeated administrations over time.

Kilgus, Methe, Maggin, and Tomasula (2014) conducted a meta-analysis to determine overall diagnostic accuracy of R-CBM. Criterion-related validity revealed a strong relationship between R-CBM and several criterion measures, namely statewide achievement tests like the North Carolina End-of-Grade Tests and published norm-referenced tests like the Stanford Achievement Test—10th edition. The weighted mean correlation was .67. However, correlations were found to be higher between R-CBM and published norm-referenced tests, specifically when they were individual measures of word identification administered less than two years after R-CBM was administered. A follow-up study that compared R-CBM to statewide achievement tests only found an overall mean correlation coefficient of .69 (Yeo, 2010). The meta-analysis found that the correlation was positively related to the study's sample size and negatively related to the proportion of English Language Learners and special education students (Yeo, 2010).

Early literacy screening. Early literacy can also be assessed using CBMs. These CBMs are general outcome measures (GOMs) that can also progress monitor the development of early reading skills as the school year progresses (Utchell, Schmitt, McCallum, McGoey, & Piselli, 2016). Measures like the Dynamic Indicators of Basic Early Literacy Skills (DIBELS) or Tests of Early Literacy (TEL) determine a student's development in sub-skills and compare it to normative scores for other students in the same grade. These measures are typically valued due to their relationship with state test scores, to which they highly correlate for a variety of grades. However, these measures are also useful because they can determine whether a student's skill development is progressing normally for their grade level (Utchell et al., 2016). They are typically administered as a screener at the beginning of the year in order to identify students that

are potentially at risk as early as possible, so that intervention can begin (Coyne & Harn, 2006). It is important to accurately identify children with delays in early literacy development, as identifying these students and providing targeted instruction will likely increase a student's performance, as prevention interventions tend to have more chances of success when started sooner in the child's difficulties instead of later, when these challenges have significantly impeded their learning (Pool, n.d.).

However, the appropriate time to start screening children has been a debated topic. Some literature claims that early literacy skills should be screened before literacy instruction formally begins (Pool, n.d.). The supporting argument for this is that reading problems can be identified and properly prevented before children are sent into the structure of literacy instruction that a child might soon fall behind in. However, other research claims that screening at any point between preschool and the beginning of kindergarten is more advantageous. This may be because the confounding problems of behavior, attention, and task motivation that might influence performance are minimized due to the children already having experience within a classroom setting. Another benefit to this timeframe for screening is that the children should be at a point where they have experienced an ample amount language and literacy, and therefore they won't be tested too early before they have gotten the opportunity to encounter these experiences (Pool, n.d.).

Early literacy screening tools need to accomplish four important goals. They must examine the core early literacy skills such as phonemic awareness and phonics, be sensitive enough to accurately differentiate between children at-risk and not, be efficiently and easily administered, and meet the minimum standards for validity and

reliability (Pool, n.d.). One of the key components of a screening measure is that it must accurately identify students that are at risk, while avoiding misclassifying any students. Classification accuracy, or predictive power, means a measure is able to accurately and reliably identify students that may experience reading difficulty in the future based on their current performance (Coyne & Harn, 2006). It includes sensitivity and specificity—sensitivity is defined as accurately identifying at-risk students that are actually at risk, and specificity is defined as accurately identifying not at-risk students that are actually not at-risk based on the results of the screening measure compared to a particular criterion measure, such as a state test (Pool, n.d.). Another important component of reading screeners is that they are useful in answering important questions about a child's progress and can guide teachers in their decision making on how to proceed with further analysis in order to best meet the needs of their student (Coyne & Harn, 2006). Data can give a brief look into a child's current ability, which can tell teachers which children are at risk and which students will likely need additional interventions. Some measures used for screening, or more frequent progress monitoring, also have additional scores such as grade equivalents, literacy classifications (emergent, transitional probable reader), or domain scores identifying performance in specific subskills like phonemic awareness. This allows them to reliably select students for interventions in a timely manner and utilize interventions that match the intensity of the problem identified (Coyne & Harn, 2006).

Early literacy measures administered in kindergarten and first grade can predict future state reading assessment performances up to seven years later (Utchell et al., 2016). The DIBELS-oral reading fluency (ORF) test predicted state reading performance

in the third grade with reasonable accuracy, with much of the variance being accounted for by the DIBELS measures of phoneme segmentation fluency and initial sound fluency. The results from this study provide evidence towards DIBELS-ORF scores being used to predict standardized state assessment performance up to five years after the screener was given (Utchell et al., 2016). However, beyond five years, the previous state reading scores from earlier grades become a more significant predictor of future state test performance, with little of the variance being accounted for by the early literacy measures (Utchell et al., 2016).

Traditional early literacy measures are given in a pencil-and-paper format. They are norm-referenced and criterion-referenced assessments that either compared a student to a sample of similar students across the country or compared the student to a set of objectives or standards they are assumed to have learned within the course of their school year, respectively (Van Horn, 2003). However, these measures lack a strong "improvement-referenced" approach to evaluation, which tracks a student's progress with repeated measurements over the course of the school year. Grade-level testing with pencil-and-paper measures can be less accurate in determining what grade level a student is reading at. For example, if a fourth grader is given a fourth grade paper-and-pencil reading probe and scores below his peers, it is hard to determine from this one probe whether he is reading one grade level below average or two. Lower grade level probes can be administered as well, but having to administer multiple probes would be very time-consuming and less effective than having a probe that adapts to the level of the student.

Defining Computer Adaptive

Computer-adaptive tests (CATs) provide students with a test that adjusts itself to become the most optimal test for each individual examinee (Meijer & Nering, 1999). The content and difficulty of the assessment adjusts each question to the student's unique performance using an item response theory (IRT) model. With this model, the examinee's "trait level" is estimated based on his or her performance during the test administration, and items that have been determined to be approximately close to this level are selected from a large item pool and administered. For example, if a student answers correctly, a slightly more challenging question is selected to be administered next—if the student answers incorrectly, the next question will be less challenging. In this way, the test attempts to match closely to each individual student's achievement level, unlike traditional paper-and-pencil (P&P) test questions that remain on the same level of difficulty regardless of a student's performance on the previous questions (Meijer & Nering, 1999). Another advantage of using a CAT versus a P&P test is that the number of questions outside of a student's current reading level are reduced, as high achievers are not exposed to questions that are too easy and lower achievers are not subjected to questions that are too difficult (Renaissance Learning, 2015). In addition to this, tests are able to be administered and scored at a faster pace (Meijer & Nering, 1999). They are unique in that they require minimal supervision for administration, as they use computer graphics, audio, and automatic instruction and test questions to give the process of testing a sense of automaticity and ease of use (Renaissance Learning, 2015). A study conducted to investigate the implications of computer-adaptive testing found that the computer-adaptive test produced greater measurement precision and less error, and an overall more positive test experience compared to paper-and-pencil tests (Martin & Lazendic, 2018).

Computer-adaptive tests are utilized in a variety of ways within academic and clinical settings. For example, a CAT was developed to more precisely measure depressive symptoms, which increased the precision with which patients were assessed for depression, along with decreasing the burden of responding to a long strand of questions by adapting to the respondent's answers and allowing for less questions to be asked (Fliege et al., 2005). In an academic setting, a CAT was used to measure predictive validity in math and was compared to Curriculum Based Measures of math to determine its relative ability to predict state assessment scores (Shapiro & Gebhardt, 2012). The relationship to the math state assessment was stronger for the CAT than it was for both CBM comparison tests. In regard to progress monitoring for math, computer adaptive tests showed a linear growth pattern as students were continually assessed in the third, fourth, and fifth grade, while the CBM measure showed linear growth in only third and fifth grades (Shapiro, Dennis, & Fu, 2015). The strongest predictive validity was demonstrated by the computer adaptive test and its relationship to the third and fourth grade outcomes. The frequency and duration of the progress monitoring may also have an effect on the quality of the predictive validity, as longer and more frequent assessment produced more accurate results (Van Norman, Nelson, & Parker, 2017).

R-CBM was compared to the computer-adaptive test, the Measures of Academic Progress (MAP) in order to determine the unique contribution each test had on students' overall scores on the Iowa Test of Basic Skills-Total Reading (ITBS-TR) composite, an achievement test for students in grades K-8 (January & Ardoin, 2015). Entered into a regression, both tests together explained 76% of the variance in the ITBS-TR composite. If the CBM-R was entered into the regression first, it accounted for 55% of the variance,

while the MAP accounted for 21% of the variance. However, if the MAP was entered into the regression first, it accounted for 75% of the variance, and the CBM-R failed to explain any significant unique variance (January & Ardoin, 2015).

Defining Star Early Literacy

STAR Early Literacy (SEL) is a computer-adaptive assessment designed to measure the early literacy skills of young children from pre-k to third grade, assessing their grasp on literary concepts and skills that directly influence their future ability to read. One of its advantages is the short amount of time it takes—only about ten minutes for administration, in addition to the fact that it can be administered to a group of students at a time, due to its computer-adaptive nature (Renaissance Learning, 2015). SEL is also a cost-effective option for assessing early literacy in comparison to three other popular early literacy tests—namely, the Dynamic Indicators of Basic Early Literacy Skills (DIBELS), the Group Reading Assessment and Diagnostic Evaluation (GRADE), and the Texas Primary Reading Inventory (TPRI) (McBride, Ysseldyke, Milone, & Stickney, 2010). However, there is limited peer-reviewed research on the use of this particular test (Clemens et al., 2015).

One study was conducted with 98 kindergarteners in order to determine the predictive validity of SEL for reading skills (Clemens et al., 2015). SEL was administered along with the Letter-Naming Fluency (LNF) CBM, Letter-Sound Fluency (LSF) CBM, Word-Reading Fluency (WRF) CBM, and the Word Identification (WID) and Word Attack (WAT) subtests from the Woodcock Reading Mastery Test—Revised/Normative Update. SEL was administered as a screening assessment in the fall, winter, and spring of the students' kindergarten year. The other paper-based measures

were administered one month after the administration of the spring SEL, and in the spring of their first grade year (Clemens et al., 2015).

SEL correlated moderately with all of the paper-administered tests, with little variance across fall, winter, and spring administrations (LNF: .39-.42; LSF: .47-.51; WID: .59-.61; WAT: .54-.57; WRF: .53-.58) (Clemens et al., 2015). Additionally, SEL correlated with WID, WAT, and WRF with a similar magnitude (.58-.60) to the LNF and LSF measures (.57-.66). For the first grade, kindergarten SEL moderately predicted the obtained scores (.51-.60), and was a stronger predictor of WID and WAT than kindergarten LNF and LSF (Clemens et al., 2015). However, LNF and LSF were stronger predictors of first grade reading fluency than kindergarten SEL. SEL scores for fall, winter, and spring also predicted word reading skills at the end of the kindergarten year and explained a statistically significant proportion of variance. For the first grade, the spring SEL significantly predicted reading accuracy and fluency. However, the amount of variance it accounted for (37% for reading accuracy, 33% for reading fluency) was less than the amount of variance accounted for by the spring WRF (43% for reading accuracy, 54% for reading fluency) (Clemens et al., 2015).

Performance on the SEL test in kindergarten were moderately predictive of end-of-year reading skills (Clemens et al., 2015). The paper-based assessment of letter-naming and letter-sound fluency explained more variance than SEL—48% variance if measured independent of SEL, and 20% more when used along with SEL. However, including SEL instead of utilizing only CBM measures allowed for 13% more accounted variance, in comparison to adding letter-naming or letter-sound fluency, which resulted in minimal variance accounted for. While SEL appears to be capable of predicting reading

skills with reasonable accuracy, it is currently unclear if its predictive power is greater than traditional paper testing.

A second study examined the technical adequacy of SEL in comparison to other early reading measures (McBride et al., 2010). Assessments were given in grades K-2 to 200 students per grade from 7 states, with SEL and DIBELS being administered twice to establish retest reliability for each grade. A cost-benefit analysis was also conducted in order to determine which test would produce the most benefit in comparison to the cost of implementation. Compared to the Dynamic Indicators of Basic Early Literacy Skills (DIBELS), the Group Reading Assessment and Diagnostic Evaluation (GRADE), and the Texas Primary Reading Inventory (TPRI), when SEL and these tests were given to kindergarten students, the three non-SEL tests correlated with each other in skill-tests of phonemic awareness and vocabulary at slightly higher rates than SEL correlated to these tests. (McBride et al., 2010). However, for grade 1, SEL correlated more highly with skill-tests of phonemic awareness, phonics, and comprehension, than the other non-SEL tests. In grade 2, SEL correlated higher with skill-tests of phonics and comprehension compared to the non-SEL tests. For the retests of SEL and DIBELS, an alternate form of DIBELS was given and an approximated alternate form of SEL was given by administering items not given in the first administration. The retest reliability for SEL was good overall for K-2. DIBELS's test-retest reliability was good for kindergarten Letter Naming, first grade Nonsense Words, and second grade Oral Reading, and acceptable for kindergarten Word Usage and second grade Retell (McBride et al., 2010). For the cost-benefit study, SEL was found to be the most cost-effective, with DIBELS being three times the cost and TPRI being six to seven times the cost of SEL (McBride et

al., 2010). It was reported that SEL was fast and easy to administer, and the reports given by the program were detailed and suitable for parents. The results were also immediate, compared to the paper versions of DIBELs and TPRI that require hand-scoring or hand-entering and even a waiting period for results, in the case of TPRI. It should be noted that DIBELS was only administered in its paper version, and TPRI was administered in both paper and computer formats, though the computer version was reported to be more efficient than the paper version. Opportunity costs were also lower for SEL than for DIBELS and TPRI, as the latter two tests require training and practice to be able to correctly administer. It was reported that though the assessments were consistent with each other, SEL was the most efficient of the three tests (McBride et al., 2010).

Purpose of the Current Study

Some previous studies (e.g., McBride et al., 2010; Clemens et al., 2015) have suggested that SEL is a cost-efficient, time-efficient, and relatively valid and reliable instrument in comparison to other early literacy tests, but results are mixed and studies are minimal. Clemens et al. (2015) concluded that SEL appears to be capable of predicting reading performance a year after its administration with a reasonable amount of accuracy. However, further analysis is required to determine its predictive accuracy beyond one year. This project examined the utility of SEL for predicting future state test performance. Specifically, the following research questions were examined:

1.) What is the relationship among first grade SEL scores and performance on the state's third grade reading test?

2.) What is the diagnostic accuracy of SEL scores for predicting reading state test performance two years later?

Method

Participants

Participants were selected from a suburban school district in the southeastern part of the United States, across six elementary schools. Out of the 234 students, 115 were female (49.1%) while 119 were male (50.9). Over half of the students were White, Hispanic (53.1%), almost half were White, Non-Hispanic (41.0%), and a few were Black (2.8%), Asian (1.5%), two or more races (1.0%), or American Indian or Alaskan Native (0.6%). These students were first graders during the 2012-2013 school year and completed third grade during the 2014-2015 school year.

Measures

STAR Early Literacy (SEL; Renaissance Learning, 2015). Since very limited peer-reviewed research has been conducted with SEL, the information from this section comes exclusively from the test publisher. SEL is a computer adaptive measure of early literacy skills for students in kindergarten through second grade, although it has been researched and administered satisfactorily to children in pre-kindergarten through third grade (Renaissance Learning, 2015). The measure examines early literacy skills in three broad domains (Word Knowledge and Skills, Comprehension Strategies, and Constructing Meaning and Numbers and Operations), which contain ten sub-domains and assess 41 different sets of skills or concepts. The software is unique in that it is computer administered and designed to require little teacher supervision. It is also computer-adaptive, which means it adjusts the difficulty of its administration depending on the student's performance. It is a very brief test, administering only 27 test items, which takes an average time of about 11 minutes, but there is no time limit (Renaissance

Learning, 2015). It correlates highly with more time-intensive measures of early literacy, providing a briefer assessment option that may be just as accurate as longer assessment. The program contains an automated database of previous and current administrations, allowing for immediate access to progress monitoring data in order to track growth over time and provide an estimate of early reading skills relative to criterion-referenced norms. The program is designed for this assessment to be administered multiple times within a year, providing criterion-referenced scores and creating reports that detail the growth of both individuals and classes of students. This also means it allows for screening and progress-monitoring (Renaissance Learning, 2015).

The ten sub-domains assessed are Alphabetic Principle, Concept of Word, Visual Discrimination, Phonemic Awareness, Phonics, Structural Analysis, Vocabulary, Sentence-Level Comprehension, Paragraph-Level Comprehension, and Early Numeracy (Renaissance Learning, 2015). In relation to the five broad categories related to reading development developed by the U.S. National Research Panel (i.e., phonemic awareness, phonics, vocabulary, comprehension, fluency), SEL measures all except for fluency (McBride et al., 2010). Alphabetic Principle assesses knowledge of letter names and sounds. Concept of Word assesses an understanding of print concepts, such as the difference between words and letters and how print is typically structured. Visual Discrimination assesses the ability to discern whether words are the same or different, and to differentiate between capital and lowercase letters. Phonemic Awareness assesses the understanding of word sounds/phonemes and identifying sounds and parts of sounds. Phonics assesses the understanding of consonants and vowels, and their variations in sounds and forms. Structural Analysis assesses the understanding of decoding syllables

and affixes and identifying the parts that make up compound words. Vocabulary assesses the knowledge of words and their meanings. Sentence-Level Comprehension assesses the ability to identify the contextual meaning of words. Paragraph-Level Comprehension assesses the ability to understand the main topic of a text and to be able to answer questions about the text. Early Numeracy assesses the ability to identify numbers, understand their correspondence with real-world objects, and understand how numbers relate to one another and to real-world concepts, such as weight and volume (Renaissance Learning, 2015).

The item bank contains more than 2300 items, with the program designed to keep track of the questions presented and to not repeat questions within a 30-day period (Renaissance Learning, 2015). To be computer-adaptive, every item had to be placed on a continuum of difficulty. This was accomplished by a Calibration Study, which tested over 250 forms of the test and almost 3000 items in order to determine the approximate difficulty of every item. Over 300 schools were included in the sample, and tested students from pre-kindergarten to third grade. The Rasch model was utilized in order to determine a difficulty value for each item and to assign a score to every student (Renaissance Learning, 2015). A psychometric review of all of all results was conducted in order to ensure the accurate difficulty rating of the items. Currently, dynamic calibration allows new test items to be tested by placing them randomly within a normal SEL test. The new items do not count towards the student's score, but data are collected on them in order to properly calibrate their level of difficulty in order to incorporate them into future tests (Renaissance Learning, 2015).

A validation study was conducted in 2001 in order to determine the reliability and validity of SEL and to calculate score distributions (Renaissance Learning, 2015). The test sampled 84 schools from the US and Canada with approximately 11,000 students, categorized based on four regions, four groups dependent on their student enrollment, and three groups based on the overall socioeconomic status of their students. Approximately 9,000 students took SEL two or more times in order to determine test-retest reliability data. For the validation study, the scaled score split-half reliability for all US grades was .91, the scaled score retest reliability for all US grades was .86, and the scaled score internal consistency reliability for all US grades was .92 (Renaissance Learning, 2015). The most recent study was completed in 2012 after the latest version of SEL was created, and was intended to provide data on the new version of SEL in comparison to the older version. This study utilized 7,420 students from 50 schools in the US and Canada, with both the older and newer version of SEL being administered to all students. Overall, the Scaled Scores for both versions of SEL correlated at .78. The newer version of SEL had an internal consistency reliability of .86 for all US grades (Renaissance Learning, 2015).

State of Texas Assessments of Academic Readiness (STAAR). STAAR is the statewide norm-referenced achievement test that is meant to be a criterion measure for reading for third through eighth grades, intended to measure whether students meet statewide academic standards. Student performance is categorized into three levels. Level 1 students performed below average and serves as a cautionary sign that the student may need intervention. Level 2 students are performing at an average level, while Level 3 students are performing at an above average level. The reading test consists of multiple choice questions that are answered after reading grade-level appropriate

passages, and assesses domains of reading skills such as passage comprehension and vocabulary. The total number of passages and questions increase with grade level, but the time limit remains the same. Stratified alpha for the overall Reading test is .89.

Procedures

Data for this project came from existing screening data. These data were collected from 2012-2015 by trained school personnel as part of their regular screening data collection. Screening (i.e., SEL) data were collected three times per year (i.e., fall, winter, spring) in first grade and the state test was administered two years later, in the spring of third grade. Institutional Review Board approval was granted to examine these data (IRB #1408955-1).

Data Analysis

Preliminary analysis included examining descriptive statistics (i.e., mean, standard deviation, skewness, and kurtosis). To answer the first research question regarding the relationship among first grade SEL scores and performance on the state's third grade reading test, Pearson's correlations were used. To answer the second research question regarding the diagnostic accuracy of SEL scores for predicting reading state test performance two years later, receiver operator characteristic (ROC) curves were used to identify a cut score that (a) maximizes both sensitivity and specificity, and (b) yields a sensitivity of .90. Area under the curve, sensitivity, and specificity were determined. Area under the curve (AUC) is an overall measure of accuracy. Sensitivity measures the proportion of true positives, or students that are identified as at risk who are actually at risk. Specificity measures the proportion of true negatives, or students that are identified as not at risk who are actually not at risk. To calculate diagnostic accuracy statistics, SEL

was the continuous, state variable and the STAAR scores were dichotomized into pass/fail and used as the state variable in SPSS.

Results

Data were collected from the 2012-2013 school year using fall, winter, and spring SEL screening for 234 students in first grade and from the third grade spring state test administration in 2015. Overall, 47 (20.1%) students had unsatisfactory performance on the state test, while 187 (79.9) had satisfactory or above. This is similar to the overall passing rates across the state (23% unsatisfactory performance, 77% satisfactory or above; Texas Education Agency, 2015). Table 1 displays descriptive statistics from the sample including mean, standard deviation, minimum and maximum values, and skewness and kurtosis. Skewness and kurtosis values are in the acceptable range (±1.96, Madansky, 1988), indicating normally distributed data.

Table 1

Descriptive Statistics

	Mean	SD	Min	Max	Skewness	Kurtosis
STAAR	1477.04	128.59	1093	1911	.47	.54
Fall SEL	675.02	86.06	327	872	-.19	.27
Winter SEL	750.52	70.92	476	955	-.52	.76
Spring SEL	777.79	53.65	576	896	-.40	.50

Note. STAAR = State of Texas Assessment of Academic Readiness, SEL = Star Early Literacy

The first research question examined the relationship among SEL scores and state test (STAAR) scores. To examine the relation between SEL and state reading testing, Pearson's correlations were employed. Table 2 displays results of the correlations. Utilizing the correlation standards from Schober, Boer, & Schwarte (2018), the

relationships between third grade reading state test scores and fall, winter, and spring SEL scores and fall SEL scores were all moderate and significant at the $p < .01$ level for a 2-tailed test. Of note, each SEL screening administration was also only moderately correlated with each other SEL screening administration.

Table 2

Correlations Among STAAR and Fall, Winter, and Spring SEL scores

Variable	STAAR	Fall SEL	Winter SEL	Spring SEL
STAAR	1			
Fall SEL	.45*	1		
Winter SEL	.45*	.46*	1	
Spring SEL	.48*	.49*	.45*	1

Note. STAAR = State of Texas Assessment of Academic Readiness, SEL = Star Early Literacy, * $p < .01$

To answer the second research question, ROC curves were used to analyze diagnostic accuracy. The state test score cutoff was determined by utilizing the score that corresponded to 2015's pass/fail criteria. The area under the curve for fall, winter, and spring SEL scores were .75, .74, and .72 respectively. As a comparison criteria, Cicchetti, Volkmar, Klin, and Showalter (1995) discuss interpretations of scores and determine that scores of .70 and below are poor, .70 to .79 are fair, .80 to .89 are good, and .90 and above are excellent. Applying these standards, the area under the curve for fall, winter, and spring SEL scores is fair. Utilizing the curves, it was determined that the cut score for the state test that maximized sensitivity and specificity for Fall SEL scores was 647.50, achieving an optimal maximized sensitivity of .66 and an optimal maximized specificity

of .70. Using the same standard of Cicchetti et al. (1995), sensitivity and specificity scores for Fall SEL fell into the poor range, meaning that the sensitivity and specificity achieved are unreliable in accurately determining true positives and true negatives for passing the state test. The cut score that produced sensitivity closest to .90 was 698.16, with a sensitivity of .89 and a specificity of .48. Applying the previously state standards of Cicchetti et al. (1995), while SEL would be able to predict true positives (passing the state test) with reasonable accuracy, the specificity suggests that its ability to predict true negatives (failing the state test) is equivalent to 50/50 chance. The cut score for the state test that maximized sensitivity and specificity for Winter SEL scores was 735.50, achieving an optimal maximized sensitivity of .70 and an optimal maximized specificity of .69. These scores still fall into the poor range and indicate that Winter SEL is incapable of being an accurate measure of determining state test pass/fail. The cut score that produced sensitivity closest to .90 was 786.50, with a sensitivity of .89 and a specificity of .36. It is worth noting the specificity is significantly under a chance level of .50. The cut score for the state test that maximized sensitivity and specificity for Spring SEL scores was 777.39, achieving an optimal maximized sensitivity of .62 and an optimal maximized specificity of .69. The cut score that produced sensitivity closest to .90 was 798.50, with a sensitivity of .89 and a specificity of .36. Similar to Fall and Winter SEL scores, the Spring SEL scores produced poor optimal sensitivity and specificity and the specificity when sensitivity was set closest to .90 was significantly below a chance level of .50, which will be discussed.

Table 3

Diagnostic Accuracy

	Cut Score	Sensitivity	Specificity	AUC
	Maximized Sensitivity and Specificity			
Fall SEL	647.50	.72	.70	.75
Winter SEL	735.50	.70	.69	.74
Spring SEL	777.39	.62	.69	.72
	Sensitivity closest to .90			
Fall SEL	698.16	.89	.48	.75
Winter SEL	786.50	.89	.36	.74
Spring SEL	798.50	.89	.36	.72

Note. STAAR = State of Texas Assessment of Academic Readiness, SEL = Star Early Literacy

Discussion

This project examined the utility of SEL for predicting future state test performance. The first research question examined the relationship among first grade SEL scores and performance on the third-grade reading state test. STAAR test scores were moderately correlated with fall, winter, and spring administrations of SEL, indicating that students who score well on one measure typically also score well on the other measure. This suggests that SEL is similar to other reading screeners, such as MAP, as both measures have a positive correlation with the scores obtained in the state test they were analyzed with (January & Ardoin, 2015). However, in comparison to other screeners, the correlations between SEL and the STAAR test were not particularly strong, indicating that the two tests may measure different aspects of literacy and utilize different standards that produce different evaluations of a student's competency in literacy (January & Ardoin, 2015).

Interestingly, SEL scores did not correlate well between its own fall, winter, and spring scores, suggesting that the SEL tests may measure different aspects of literacy or may use different standards that produce different evaluations of a student's competency in literacy as the student progresses through the school year. This is particularly problematic when progress monitoring typically measures the same skill across time while utilizing the same standards. SEL may not accurately progress monitor if it is measuring different skills with each test or utilizes different standards with each administration. This could cause a student's scores to fluctuate wildly, which could make it difficult to find a trend in their data to accurately determine a need for intervention.

The second research question examined the diagnostic accuracy of SEL scores for predicting reading state test performance two years later. The area under the ROC curve is somewhat acceptable for fall, winter, and spring SEL scores, as all scores are within the .70 to .79 range of fair of Cicchetti et al. (1995), suggesting it has some ability to discriminate between people that may pass or fail the state test. However, an ideal level would be at least .80 or higher, which would improve its diagnostic accuracy and thus its reliability as an accurate test of literacy skills (Cicchetti et al., 1995). As time went on through the year, optimal maximum SEL sensitivity dropped from fall to spring. Given the relatively weak correlations between fall, winter, and spring SEL, it can be speculated that the variations between tests may have had an impact on their ability to accurately determine students that would pass and fail the state test. The evaluation standards may have been developed using the fall SEL, which could explain why sensitivity was highest at this point. If the developers assumed that fall, spring, and winter SEL would be approximately the same and would measure the same aspects of literacy, they may have been compelled to not include spring and winter SEL in creating their standards. However, if there were unaccounted for variations in the winter and spring SEL tests that may not have matched well with the original evaluation standards, the winter and spring SEL would be less likely to accurately determine students' true early literacy abilities, and thus also less likely to accurately discriminate between students that would pass and fail the state test.

The optimal maximized sensitivity of SEL indicates that it would be able to identify approximately 72%-62% of students that would fail the state test. In addition, the optimally maximized specificity of SEL indicates that it would be able to identify 69%-

70% of students that will pass the state test. Both of these levels are somewhat above chance levels of 50%, but do not reach a more reliably acceptable level of .80 (Cicchetti et al., 1995). If sensitivity is set close to a desirable level of .90, to where it would be able to identify 90% of students that would fail the state test, then SEL would only be able to identify 48%-36% of students that will pass the state test. Interestingly, the specificity levels are significantly below a chance level of 50%, which could indicate that SEL is actively not identifying true positives. Coupled with the higher standard of sensitivity, it appears that requiring the test to accurately identify students that would fail biases it into overly identifying students as potentially failing even if they would not, producing a significant amount of false negatives. Given the large amount of false negatives, it completely diminishes the benefit of having a higher ability to identify true negatives. In comparison to other reading screeners, the predictive validity achieved by SEL is much less than what can be produced by other measures (Clemens et al., 2015).

Implications

Given these results, while SEL positively correlates at a moderate level with STARR (though at the lower end of the .40 to .69 level of moderate from Schober, Boer, & Schwarte (2018)), it is not that reliable in accurately discriminating between students that will pass or fail the state test. Since the scores produced by SEL do not reliably indicate whether a student will pass or fail the state test, it should not be used as a measure to predict a student's future performance. SEL may be utilized as a screener to indicate current student reading performance, and decisions to investigate potentially at-risk students can be made based on current data. However, SEL is not reliable as a screener to identify future reading performance, and therefore should not be used in

decisions to potentially intervene with a child based on their potential to not perform well in the future.

Limitations

The current study utilized students from one school district, which limits the ability to generalize findings to other school districts. The current study also uses one state's reading test, which may not generalize the results to reading tests from other states. Results may also vary as SEL scores were only compared to 3rd grade state testing, whereas comparison to subsequent years of state testing may produce different results.

Future Directions

Future studies on this topic should consider taking samples from multiple school districts across the country and compare scores to their respective state tests, in order to identify if SEL's usefulness in predicting future state test scores is more viable in other states. Comparison to 4th and 5th grade state testing may indicate that SEL is more accurate in identifying state test pass/fail in these grades, or perhaps identify a pattern of becoming more or less accurate as more time passes

References

Cicchetti, D., Volkmar, F., Klin, A., & Showalter, D. (1995). Diagnosing autism using ICD-10 criteria: A comparison of neural networks and standard multivariate procedures. *Child Neuropsychology, 1*, 26-37.

Clemens, N. H., Hagan-Burke, S., Luo, W., Cerda, C., Blakely, A., Frosch, J., . . . Jones, M. (2015). The predictive validity of a computer-adaptive assessment of kindergarten and first-grade reading skills. *School Psychology Review, 44*(1), 76-97.

Coyne, M. D., & Harn, B. A. (2006). Promoting beginning reading success through meaningful assessment of early literacy skills. *Psychology in the Schools, 43*(1), 33-43.

Crone, D. A., & Whitehurst, G. J. (1999). Age and schooling effects on emergent literacy and early reading skills. *Journal of Educational Psychology, 91*(4), 604-614.

Dickinson, D. K., McCabe, A., Anastasopoulos, L., Peisner-Feinberg, E. S., & Poe, M. D. (2003). The comprehensive language approach to early literacy: interrelationships among vocabulary, phonological sensitivity, and print knowledge among preschool-aged children. *Journal of Educational Psychology, 95*(3), 465-481.

Dynamic Measurement Group. (2008). *DIBELS 6th Edition Technical Adequacy Information.* Eugene, OR.

Fliege, H., Becker, J., Walter, O. B., Bjorner, J. B., Klapp, B. F., & Rose, M. (2005). Development of a computer-adaptive test for depression (D-CAT). *Quality of Life Research*, 14, 2277-2291.

Fuchs, D., & Fuchs, L. S. (2005). Responsiveness-to-intervention: A blueprint for practitioners, policymakers, and parents. *Teaching Exceptional Children, 38(1),* 57-61.

Gillett, E. (2016). Early literacy development. In D. Couchenour, & K. J. Chrisman, *The SAGE encyclopedia of contemporary early childhood education* (pp. 501-502). Thousand Oaks, CA: SAGE Publications, Inc.

Greenfield-Spira, E., Storch Bracken, S., & Fischel, J. E. (2005). Predicting improvement after first-grade reading difficulties: the effects of oral language, emergent literacy, and behavior skills. *Developmental Psychology, 41(1),* 225-234.

Individuals with Disabilities Education Act, 20 U.S.C. § 1400 (2004).

January, S. A. A., & Ardoin, S. P. (2015). Teachnical adequacy and acceptability of curriculum-based measurement and the measures of academic progress. *Assessment for Effective Intervention, 41(1),* 3-15.

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

Kilgus, S. P., Methe, S. A., Maggin, D. M., & Tomasula, J. L. (2014). Curriculum-based measurement of oral reading (R-CBM): A diagnostic test accuracy meta-analysis

of evidence supporting use in universal screening. *Journal of School Psychology, 52(1),* 377-405.

Lonigan, C. J., Burgess, S. R., & Anthony, J. L. (2000). Development of emergent literacy and early reading skills in preschool children: evidence from a latent-variable longitudinal study. *Developmental Psychology, 36*(5), 596-613.

Madansky, A. (1988). *Prescriptions for working statisticians.* New York, NY: Springer-Verlag.

Martin, A. J., & Lazendic, G. (2018). Computer-adaptive testing: Implications for students' achievement, motivation, engagement, and subjective test experience. *Journal of Educational Psychology*, *110(1),* 27-45.

McBride, J. R., Ysseldyke, J., Milone, M., & Stickney, E. (2010). Technical adequacy and cost benefit of four measures of early literacy. *Canadian Journal of School Psychology*, *25(2)*, 189-204.

Meijer, R. R., & Nering, M. L. (1999). Computerized adaptive testing: overview and introduction. *Applied Psychological Measurement, 23(3),* 187-194.

National Association for the Education of Young Children. (1998). Learning to read and write: developmentally appropriate practices for young children. *Young Children, 53(4),* 30-46.

National Center for Family Literacy. (2008). *Developing early literacy: report of the national early literacy panel.* Jessup, MD: ED Pubs.

National Center on Response to Intervention. (n.d.). *Essential components: screening webinar.* (p. 11). rti4success.org.

National Reading Panel. (2000). *Teaching children to read: An evidence-based assessment of the scientific research literature on reading and its implications for reading instruction.* Washinbgton, DC: National Reading Panel.

Pool, J. L., & Johnson, E. S. (n.d.). *Screening for reading problems in preschool and kindergarten: an overview of select measures.* Retrieved from http://www.rtinetwork.org/.

Renaissance Learning. (2015). *Technical Manual.* Wisconsin Rapids: Renaissance Learning.

Schober, P., Boer, C., & Schwarte, L. (2018). Correlation coefficients: Appropriate use and interpretation. *Anesthesia & Analgesia, 126*(5), 1763-1768.

Shapiro, E. S., & Gebhardt, S. N. (2012). Comparing computer-adaptive and curriculum-based measurement methods of assessment. *School Psychology Review, 41(3),* 295-305.

Shapiro, E. S., Dennis, M. S., & Fu, Q. (2015). Comparing computer adaptive and curriculum-based measures of math in progress monitoring. *School Psychology Quarterly, 30(4),* 470-487.

Shinn, M. (1998). *Advanced applications of curriculum-based measurement.* New York, NY: Guilford Press.

Texas Education Agency. (2015). *Summary report: Grade three reading.* Retrieved from https://tea.texas.gov/sites/default/files/staar-sum2015-04spring-g3-fr.pdf

Utchell, L. A., Schmitt, A. J., McCallum, E., McGoey, K. E., & Piselli, K. (2016). Ability of early literacy measures to predict future state assessment performance. *Journal of Psychoeducational Assessment, 34*(6), 511-523.

Van Horn, R. (2003). Computer adaptive tests and computer based tests. *Technology*, *567*, 630-631.

Van Norman, E. R., Nelson, P. M., & Parker, D. C. (2017). Technical adequacy of growth estimates from a computer adaptive test: Implications for progress monitoring. *School Psychology Quarterly*, *32(3)*, 379-391.

Yeo, S. (2010). Predicting performance on state achievement tests using curriculum-based measurement in reading: A multilevel meta-analysis. *Remedial and Special Education*, *31*, 412-422.