DEVELOPING FRACTION SCREENERS TO IDENTIFY CHILDREN AT RISK FOR MATHEMATICS DIFFICULTIES

by

Jessica Rodrigues

A dissertation submitted to the Faculty of the University of Delaware in partial fulfillment of the requirements for the degree of Doctor of Philosophy in Education

Summer 2017

DEVELOPING FRACTION SCREENERS TO IDENTIFY

CHILDREN AT RISK FOR MATHEMATICS DIFFICULTIES

by

Jessica Rodrigues

Approved: _____

Ralph P. Ferretti, Ph.D.
Director of the School of Education

Approved: _____

Carol Vukelich, Ph.D.
Dean of the College of Education and Human Development

Approved: _____

Ann L. Ardis, Ph.D.
Senior Vice Provost for Graduate and Professional Education

I certify that I have read this dissertation and that in my opinion it meets the academic and professional standard required by the University as a dissertation for the degree of Doctor of Philosophy.

Signed: _____

Nancy C. Jordan, Ed.D.
Professor in charge of dissertation

I certify that I have read this dissertation and that in my opinion it meets the academic and professional standard required by the University as a dissertation for the degree of Doctor of Philosophy.

Signed: _____

Roberta M. Golinkoff, Ph.D.
Member of dissertation committee

I certify that I have read this dissertation and that in my opinion it meets the academic and professional standard required by the University as a dissertation for the degree of Doctor of Philosophy.

Signed: _____

Henry May, Ph.D.
Member of dissertation committee

I certify that I have read this dissertation and that in my opinion it meets the academic and professional standard required by the University as a dissertation for the degree of Doctor of Philosophy.

Signed: _____

Joshua Wilson, Ph.D.
Member of dissertation committee

ACKNOWLEDGMENTS

This dissertation is part of a longitudinal study (2010-2015) that investigated children's mathematical development specifically in the area of fractions. The study was supported by a grant from the Institute of Education Sciences (IES) (R324C100004). The work detailed in this dissertation is done so with the permission of the principal investigator Dr. Nancy C. Jordan.

I would like to thank the many individuals who supported me during my doctoral journey and during the process of completing this dissertation. First, I would like to thank my advisor, Dr. Jordan. Thank you for believing in my potential as a new scholar, for challenging me to push through obstacles, and for serving as my mentor these last four years. Your support has been unwavering. The encouragement and guidance you have provided goes above and beyond what is expected of a doctoral advisor, and I will forever be grateful to you.

I would also like to acknowledge Dr. Golinkoff, Dr. May, and Dr. Wilson for serving on my dissertation committee. The time that you dedicated to providing feedback on this research is greatly appreciated. Thank you, Dr. Golinkoff, for encouraging me and for your insightful feedback. Thank you, Dr. May, for sharing your statistical expertise and for encouraging me to learn new analyses to answer my research questions. Thank you, Dr. Wilson, for your statistical expertise and for sharing my appreciation of ROC curve analyses.

Thank you to all Fraction Project team members who have made time spent in the lab a fun and enjoyable experience. You have all inspired me in different ways.

Thank you, Dr. Nancy Dyson, for supporting my dreams as a scholar and for creating such a rewarding work environment. Your passion for mathematics and your dedication to having a positive impact on all children we worked with in the intervention is nothing short of incredible. Thank you, Deborah Micklos, for your words of wisdom and support. Thank you, Dr. Nicole Hansen, for being my peer mentor when I first began the Ph.D. program. You have continued to support me even after your doctoral graduation, and your kindness and work ethic never cease to amaze me.

Finally, I would like to thank my family for their love and support. Thank you to my grandmother who is one of my favorite role models because of her strength and values. Thank you to my grandfather in heaven who loved to brag about his granddaughter pursuing her doctoral degree. Thank you to my husband and best friend, Henrique, for his encouragement. You have been by my side to celebrate my achievements, to lend a shoulder during moments of stress, and to distract me from work when I needed a break. I am thrilled that you can now say "I survived my wife's Ph.D. dissertation!" I love you, Henrique. A special thank you to my parents Dolores and William. You are my twin pillars without whom I could not stand. You told me recently that you have no idea where my drive comes from; hearing that was shocking to me, because everything I am is because of your love, your support, and your example. I love you, Mom and Dad.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

Fractions are a core topic in the mathematics curriculum in the intermediate grades. Competence with fractions is crucial for success in more advanced mathematics courses, such as algebra. Yet many students struggle to learn fractions and are at risk for later mathematics difficulties. To help prevent such difficulties, effective screening tools must identify students who are likely to struggle with fractions before they experience failure. To address this issue, the present study investigated the effectiveness of three fraction measures (i.e., general fraction concepts, fraction number line estimation, and fraction arithmetic) for screening fourth-, fifth-, and sixth-graders, respectively, who will go on to have trouble with mathematics. In particular, the study used data from a larger longitudinal study to assess the accuracy of the measures for identifying students who will not meet state standards on the end-of-the-year high-stakes mathematics test. Receiver operating characteristic (ROC) curve analyses were used to assess the accuracy of fourth-grade fraction measures to predict mathematics achievement in the spring of fourth grade (n = 411), fifth grade (n = 362), and sixth grade (n = 304); the fifth-grade fraction measures to predict mathematics achievement in the spring of fifth grade (n = 384) and sixth grade (n = 318); and the sixth-grade fraction measures to predict mathematics achievement in the spring of sixth grade (n = 327).

Analyses revealed that the general fraction concepts and fraction number line estimation measures consistently emerged as accurate screeners of risk status across the grades while the fractions arithmetic measure did not consistently meet accuracy

standards. That is, only the general fraction concepts measure and the fraction number line measure consistently demonstrated high diagnostic accuracy for predicting the outcome measure as determined by the area under the curve (AUC). In fourth and sixth grades, the general fraction concepts measure and fraction number line estimation measure did not significantly differ for the prediction of the outcome; the two measures were thus combined at each of these grades. In fifth grade, the general fraction concepts measure alone emerged as the best fifth-grade screener. To improve practicality for classroom use, the length of each screener was reduced using best subsets automatic linear modeling. Items with low predictive values for predicting the outcome were eliminated from the final best subset screeners.

Screener statistics are reported for all six best subset screeners in ways that allow researchers and practitioners to administer the screeners with their own sample of students and make predictive interpretations of students' scores. Specifically, the present study demonstrates a method for reporting ROC curve statistics for a single best subset screener that has not yet been introduced in the educational literature. First, a table is provided that reports ROC curve statistics for all cut scores along the screener, allowing researchers and practitioners to select any cut score based on their available resources and research goals. Second, a helpful figure called a probability nomogram is provided that allows readers to easily determine predictive interpretations of students' screener scores. A separate method is demonstrated for reporting a screener that combines two best subset measures. Using logistic regression, a predicted probabilities matrix is provided that allows readers to make interpretations of a student's later mathematics achievement based on the student's scores on both measures.

Addressing an important gap in the literature, the study provides recommendations for mathematics screening in the intermediate grades. The study also demonstrates how to report educational screeners in ways that allow researchers and practitioners to translate screener statistics into readily usable information. The following documents are appended: (a) the full general fraction concepts measure, (b) the full fraction arithmetic measure, and (c) a glossary of statistical terms.

Chapter 1

INTRODUCTION

Proficiency in mathematics is important for success in science, technology, engineering, and mathematics (STEM) vocations (National Mathematics Advisory Panel [NMAP], 2008). Fractions, in turn, are foundational for learning algebra (Booth, Newton, & Twiss-Garrity, 2014) and thus are an important part of the mathematics curriculum in the elementary and intermediate grades (National Governors Association Center for Best Practices & Council of Chief State School Officers [NGACBP & CCSSO], 2010; NMAP, 2008). Fraction procedures, such as finding common denominators, are vital for manipulating algebraic equations, and fraction magnitude understanding can help students reason about answers to algebraic equations (Siegler et al., 2012). A student who incorrectly writes that $x = 40$ given that $(5/4)x = 10$ can use his/her fraction understanding to realize that this answer is not plausible: 5/4 is greater than one, so x needs to be less than ten. Not surprisingly elementary students' understanding of fractions predicts future general mathematics achievement in high school, even after controlling for other types of mathematical knowledge, overall IQ, and family income (Siegler et al., 2012).

Fractions have consistently proven to be a challenging topic for students. Many students struggle to develop even a basic understanding of fractions (e.g., Bailey, Hoard, Nugent, & Geary, 2012; Hansen, Jordan, & Rodrigues, in press) and show minimal growth in fraction knowledge in fourth through sixth grades (Resnick et al., 2016), the period when fractions are taught in school. Students who do not master fractions during the intermediate grades are ill-prepared for subsequent mathematics courses. A weak understanding of fractions can have serious and enduring consequences. Students who

struggle with middle school mathematics are less likely to pursue advanced courses necessary for entry into STEM careers (NMAP, 2008; Sadler & Tai, 2007). Furthermore, college students' success in STEM courses such as science is predicted by their high school mathematics grades (Tai, Sadler, & Mintzes, 2006).

To help prevent mathematics failure, teachers need screening tools to help them identify students who are likely to struggle. A fourth-grader, for example, who shows difficulties with fraction knowledge on a screening measure can be targeted for additional supports, such as intervention or mentoring (Gersten et al., 2009). Previous research, however, has focused primarily on mathematics screeners administered in kindergarten through second grade for identifying students at risk for mathematics difficulties (e.g., Duncan et al., 2007; Jordan, Kaplan, Ramineni, & Locuniak, 2009; Morgan, Farkas, & Wu, 2009). In a review of screening measures administered in these early grades, Gersten and colleagues (2012) recommend that screening measures be extended to the intermediate grades. Even students who meet mathematics benchmarks in the early grades can encounter difficulties in later grades when fractions are introduced, typically in third or fourth grade (Gersten, Clarke, Haymond, & Jordan, 2011; Gersten et al., 2012; NGACBP & CCSSO, 2010).

Specific Aims

To address the need for mathematics screeners in the intermediate grades the present study uses longitudinal data to examine the predictability of fraction screening tools. In particular, a collection of fraction measures was administered in fourth, fifth, and sixth grades to determine their potential usefulness for identifying students who are struggling in mathematics. Fraction skills were purposefully selected for examination because they comprise a large and potentially unifying portion of mathematics content during these grades (NGACBP & CCSSO, 2010). Moreover, level of fraction knowledge is associated with later mathematics (e.g., Booth & Newton, 2012). Fractions stand out as

a particularly challenging topic for many students (e.g., Bailey et al., 2012; Ni & Zhou, 2005), which further suggests the need to screen for difficulties in this area of mathematics.

The specific aims of the present study are: (a) to assess the diagnostic accuracy of fraction concepts and fraction arithmetic measures in fourth through sixth grades, (b) to identify the measure(s) in each grade with highest diagnostic accuracy for predicting later mathematics achievement while also considering practicality for classroom use, and (c) to demonstrate an improved way for educational researchers to report results from receiver operating characteristic (ROC) curve analyses, which assess the predictive strength of potential screener measures. Each aim is discussed next.

Assess the Diagnostic Accuracy of Different Measures of Fraction Knowledge Administered in Fourth through Sixth Grades

As noted earlier, fraction knowledge is a unique and reliable predictor of later mathematics achievement (e.g., Resnick et al., 2016), even when controlling for the contributions of cognitive abilities and family income (Siegler et al., 2012). This finding, along with fractions being a critical component of mathematics education starting in mid-elementary school (NGACBP & CCSSO, 2010; NMAP, 2008), suggests that fraction measures administered during these grades are strong candidates for screeners of later mathematics difficulties.

The present study examines two different types of fraction knowledge: fraction concepts and fraction arithmetic. Acquisition of mathematics knowledge in any area requires a student to be accurate and fluent with both concepts and arithmetic procedures (Geary, 2004). Some studies have suggested that fraction concepts and procedures are relatively independent constructs (e.g., Hallett, Nunes, Bryant, & Thorpe, 2012; Hansen et al., in press; Ye et al., 2016).

To consider the fraction measures as potential screeners, analyses were conducted to determine each measure's ability to predict students who fall into one of two groups: students who do not meet the mathematics standards on a validated end-of-the-year statewide achievement test versus students who do meet the mathematics standards; this diagnostic ability is called *diagnostic accuracy* (Jordan, Glutting, Ramineni, & Watkins, 2010; Youngstrom, 2014). Diagnostic accuracy and screening potential of all measures administered in the fall/winter of fourth, fifth, and sixth grades for predicting later, general mathematics achievement on a statewide mathematics achievement test are reported.

Identify the Fraction Measure(s) with Highest Diagnostic Accuracy per Grade While also Considering Practical Value for Teachers

A second goal of the present study was to identify a screening measure or a combination of measures that not only accurately predicts mathematics difficulties by also has practical value for teachers. The first step was to identify the fraction predictor measure(s) per each grade with high predictive power, as determined by high diagnostic accuracy. In the scenario of multiple measures emerging as valid screeners with high diagnostic accuracy, the measures would be combined and tested.

The second step was to assess if the selected screener or combination of screeners could be modified to improve time efficiency and usability for the classroom, without sacrificing the high diagnostic accuracy associated with the full screener(s) for the prediction of later mathematics performance. For example, a screener with 15 items is quicker to administer and easier to score than a measure that includes those same 15 items plus 20 additional items. The best combinations of items for each grade (i.e., fourth, fifth, and sixth grade) that do not sacrifice the high diagnostic accuracy of the screener are tested and reported.

Demonstrate an Improved Method of Reporting Screener Statistics in Educational Studies

A final goal of the study was to propose a more complete set of recommendations for reporting screener results in ways that are understandable and usable for teachers than presently exists in the educational literature. The type of screener (i.e., single screener versus combined measure) determined the type of statistics reported in the present study.

The primary analysis used in studies of single educational screeners is the receiver operating characteristic (ROC) curve analysis (e.g., Jordan, Glutting, & Ramineni, 2008; Wilson, Olinghouse, McCoach, Santangelo, & Andrada, 2016). ROC curve analysis is a statistical method cited in the literature for determining the diagnostic accuracy of a screener (e.g., Gersten et al., 2012). For the present study, ROC analyses evaluate the ability of a screener measure to identify students who are at risk for later mathematics difficulties (i.e., true positives) and to rule out the concern for students who are not at risk (i.e., true negatives; e.g., Cummings & Smolkowski, 2015). In the present study, ROC curve analyses are conducted to assess the diagnostic accuracy of fraction measures for the prediction of students at risk and for computing statistics associated with screener cut scores.

Currently, there are two sets of guidelines for conducting and reporting ROC curve analyses; one set is reported in the educational literature (Cummings & Smolkowski, 2015) and the other is reported in the medical literature for clinical decision-making (Bossuyt et al., 2003; Youngstrom, 2014). The medical guidelines for *conducting* ROC curve analyses do not differ from the guidelines in the educational literature. However, the two sets of guidelines differ in recommendations for *reporting* the ROC data for use by clinicians or teachers in real-world settings. The medical guidelines appear more helpful for reporting ROC statistics in ways that are easy to use for clinicians. Even though the context of clinicians working with patients in a medical setting differs from working with students in a classroom setting, the ROC analysis itself

and the corresponding statistics remain the same across disciplines, just with different interpretations.

In response, the present study *conducts* ROC analyses by following the general ROC framework that is supported by both the educational literature (Cummings & Smolkowski, 2015) and the medical literature (Bossuyt et al., 2003; Youngstrom, 2014). For the final step of *reporting* the ROC data in a way that translates the statistics into usable information for teachers, the present study follows the medical guidelines. By comparing the two sets of guidelines, an enhanced list of ROC recommendations is proposed for educational purposes. The goal is to help educational researchers report results in a way that is most interpretable for real-world classroom application.

If combining one screener with discrete data and a second screener with continuous data (e.g., measuring students' estimates on the fraction number line measure with percent absolute error), the ROC analysis cut score statistics become difficult to interpret. The researcher must now consider a *combination* of scores to make predictions of risk status. For this scenario, the present study presents a matrix that allows a researcher or teacher to easily make empirically-driven decisions of a student's later mathematics achievement (Clemens, Keller-Margulis, Scholten, & Yoon, 2016).

Educational Significance

Screeners in the intermediate grades help teachers identify children who are at risk for later difficulties. Fractions are foundational to mathematics curricula in upper elementary school (NGACBP & CCSSO, 2010), and competence with fractions prepares students for succeeding in algebra and higher-level mathematics courses. Yet research on effective screeners of fractions, in particular, and mathematics difficulties, more generally, in the intermediate grades is scarce (Gersten et al., 2011; 2012). Identification of students who need help in fractions can lead to intervention, additional support, and mentoring. As such, identifying students who are at risk for mathematics is the crucial

step for helping struggling students (Clemens et al., 2016). Additionally, improving the educational guidelines for reporting results in ways that are understandable and usable for teachers can inform data-driven decision-making in the classroom, such as teachers deciding which students should receive additional supports. Without access to valid screening measures in the intermediate grades, students who are at risk for later difficulties may go unnoticed, never receiving the supports they need to prepare them for later success.

Chapter 2

REVIEW OF THE LITERATURE

The present literature review has several goals. First, it is argued that fractions play a key role in mathematical development, with understanding of fraction magnitudes being especially important. Second, to show that fraction knowledge is multifaceted, the review addresses the differences and connections between fraction concepts and fraction procedures. Third, the review examines studies showing fraction knowledge to be a unique predictor of later mathematics achievement. Fourth, the review discusses the importance of mathematics screeners for detecting at-risk students, identifies existing mathematics screeners for the intermediate grades, and summarizes a method for improving the efficiency of screeners for classroom use. Finally, the review addresses the importance of receiver operating characteristic (ROC) curve analysis for assessing screener measures, outlines statistical terms that are foundational to the analysis, and shares an improved way for educational researchers to report ROC statistics that is more useful for teachers.

Numerical Development, Magnitude Knowledge, and Fractions

Earlier theories of numerical development have focused on the development of whole number knowledge and depict this whole number learning as relatively discontinuous with learning fractions (Geary, 2006; Gelman & Williams, 1998). Geary (2006) posits that knowledge of fraction concepts and fraction procedures are biologically secondary competencies. These theories view whole number and fraction learning as segmented processes in numerical development. While whole number

knowledge is depicted as developing in a somewhat natural way, fraction knowledge is viewed as a challenging skill that is acquired later in development.

Segmented theories of numerical development focus on the discontinuities between whole numbers and fractions, meaning the properties of whole numbers that do not always hold true for fraction reasoning. For example, children are familiar with each whole number having a unique successor, such as "four" always follows "three." Children are predisposed to make the assumption that all numbers have one specific successor, but this principle does not apply when working with fractions, which are infinitely divisible (Siegler, Fazio, Bailey, & Zhou, 2013; Siegler, Thompson, & Schneider, 2011). Furthermore, each whole number is represented by solely one symbol, whereas fraction magnitudes can be represented by many different symbols (Siegler & Lortie-Forgues, 2014) due to fraction equivalence (e.g., 1/2 = 4/8 = 50/100). The struggles that many children experience with fractions have been attributed to these discontinuities between whole numbers and fractions. Students often incorrectly apply whole number knowledge to fraction tasks, such as reasoning that 1/4 is smaller than 1/10 because four is smaller than ten, and this tendency can be seen in both children and adults (DeWolf & Vosniadou, 2011; Ni & Zhou, 2005).

The integrated theory of numerical development (Siegler & Lortie-Forgues, 2014; Siegler et al., 2011) argues that, while the many discontinuities between whole numbers and fractions distort fraction understanding, the discontinuities do not offer a complete picture of numerical development. Siegler and colleagues propose a *continuous* process of development that is unified by one key understanding: all real numbers have magnitudes that can be represented on a number line. The emphasis on the continuity between all real numbers while still acknowledging the discontinuities that impact development creates a more unified depiction of the process of numerical development than offered by prior theories. According to the integrated theory, the one understanding that serves as a unifying theme for numerical development is the ongoing advancement

and growth of understanding numerical magnitudes (Siegler & Lortie-Forgues, 2014). The development of numerical knowledge is considered as a process of broadening the set of numbers whose magnitudes can be accurately represented.

The importance of magnitude understanding in numerical development has been highlighted in many studies, even early in development. Children's success in preschool on non-symbolic approximate number system (ANS) tasks to identify whether an array of dots on a screen shows more blue dots or more yellow dots) correlates with mathematics ability after controlling for age and verbal skills (Libertus, Feigenson, & Halberda, 2011). Mazzocco, Feigenson, and Halberda (2011) showed that precision on these non-symbolic ANS tasks again measured in preschool predicts mathematics achievement two years later.

The importance of magnitude understanding has also been reported in studies using symbolic numerical tasks involving fractions (e.g., Resnick et al., 2016; Siegler et al., 2011; Torbeyns, Schneider, Xin, & Siegler, 2015). Fraction magnitude understanding is the ability to comprehend, estimate, and compare the sizes of fractions (Fazio, Bailey, Thompson, & Siegler, 2014). A common measure of fraction magnitude understanding is a number line task in which students estimate the locations of individual fractions on a number line (e.g., Booth et al., 2014; Siegler et al., 2011). Siegler et al. (2011) administered a 0-1 fraction number line task (i.e., a number line that begins with zero on the left endpoint and extends to one on the right endpoint) and a 0-5 fraction number line task to students in sixth and eighth grades. The researchers report strong correlations in both grades between accuracy on each fraction number line task with fraction arithmetic performance and overall mathematics achievement. Resnick et al. (2016) looked longitudinally from fourth through sixth grade at students' accuracy when estimating fraction magnitudes on 0-1 and 0-2 number lines. Although students showed positive linear growth overall, latent class growth analyses revealed three empirically distinct growth trajectory classes: Students who were highly accurate on the fraction number line

task in fourth grade and become even more accurate by sixth grade; students who were inaccurate in fourth grade but improved greatly over the course of the study; and students who were inaccurate in fourth grade and showed minimal growth. Student membership in the different growth classes was highly predictive of achievement on a statewide standardized mathematics test in the spring of sixth grade, demonstrating the importance of growth in fraction magnitude understanding to mathematics achievement more broadly.

Overall, magnitude understanding in particular is an important concept that is relevant to both whole numbers and fractions. Fraction magnitude understanding emerges as a key aspect of numerical development. Students who struggle on fraction number line tasks lack a deep understanding of fraction magnitudes and often fail to see the relation between the numerator and the denominator (Resnick et al., 2016). For example, struggling students often inaccurately place the fraction 1/19 to the far right of a 0-2 number line. The students view the 19 in the denominator as a "big number" and estimate that the fraction has a large magnitude (Rodrigues, Dyson, Hansen, & Jordan, 2017). Students who cannot accurately place fractions on a number line are likely to continue to struggle in mathematics classes, at least without receiving additional supports (Resnick et al., 2016; Siegler et al., 2011). The next section differentiates between fraction concepts and fraction procedures.

Differentiating Fraction Concepts from Fraction Procedures

According to Geary (2004), mastery of mathematical knowledge depends on both concepts and procedures. As applied to fraction learning, conceptual knowledge involves not only an understanding of fraction magnitudes (e.g., Siegler et al., 2013), but also an understanding of fraction notations and the recognition that an infinite number of fractions exist between any two fractions (Van Hoof, Janssen, Verschaffel, & Van Dooren, 2015). The tasks used to measure fraction concepts, however, differ from study

to study. Some researchers use a broad set of items that touch on several different fraction concepts, such finding parts of whole and parts of a set, understanding the larger of two fractions, and the ability to order fractions according to their magnitudes (e.g., Fuchs et al., 2013; Jordan et al., 2013; Seethaler, Fuchs, Star, & Bryant, 2011; Vukovic et al., 2014), while others use more fine-grained assessments of fraction magnitudes, such as the aforementioned fraction number line task (e.g., Resnick et al., 2016; Siegler et al., 2011). On the other hand, procedural knowledge involves knowing procedures for adding, subtracting, multiplying, and dividing fractions as well for solving other fraction problems (e.g., cross multiplying; Hecht & Vagi, 2012; Siegler et al., 2013). Proficiency in one domain of fraction knowledge does not always imply proficiency in the other; for example, a student who successfully uses an algorithm for a fraction arithmetic problem is not always aware of *why* the algorithm works (Hecht & Vagi, 2012).

Both concepts and procedures are important aspects of fraction learning. Hecht, Close, and Santisi (2003) conducted a study with fifth-grade students to investigate the relation between whole number arithmetic and fraction concepts (which tapped into many conceptual understandings, including fraction comparisons and area model representations of fractional quantities) on three different outcomes: fraction computation, fraction arithmetic word problems, and fraction estimation for which students estimated the sums of fraction computation problems. Using structural equation modeling, the data showed that fraction conceptual knowledge uniquely contributed to performance in all three fraction outcomes. In contrast, whole number arithmetic knowledge independently contributed only to the fraction computation outcome (Hecht et al., 2003). These findings suggest that while both conceptual and procedural knowledge are important for success with fractions, conceptual understanding may be of higher importance and may even influence performance in fraction procedures. For example, consider a fifth-grader who has a strong conceptual understanding of fractions. When solving the computation problem 3/4 + 1/4, the student impulsively writes down the

answer 4/8, incorrectly adding together the denominators. However, the student's conceptual understanding of fractions helps her to realize that 4/8 is equivalent to 1/2 and thus not a reasonable answer to the problem. A different student who applies memorized procedures to a computation problem without reasoning conceptually about the problem is less likely to catch this error.

In an investigation of individual differences in how children combine conceptual and procedural fraction knowledge, researchers Hallet, Nunes, and Bryant (2010) found that some children rely more on procedural knowledge, whereas other children depend more on conceptual knowledge; the latter appear to have an advantage on both conceptual and procedural problems compared to the students who rely primarily on procedural knowledge, again pointing to the importance of conceptual understanding. Hecht and Vagi (2012) showed that students who exhibited better conceptual knowledge than procedural knowledge demonstrated higher accuracy on a fraction computation task than students with low conceptual knowledge. However, the researchers also found that some students with relatively high procedural knowledge use this knowledge to compensate for their weaker conceptual knowledge (Hecht & Vagi, 2012).

The timing of conceptual and procedural development in mathematics is also important to consider. There is disagreement in the literature regarding whether concepts or procedures are acquired first, with some studies showing that children develop conceptual knowledge before procedural knowledge (e.g., Peck & Jencks, 1981) and others documenting cases in which procedural knowledge develops within a domain before conceptual knowledge (e.g., Byrnes & Wasik, 1991). A more recent perspective is the iterative model proposed by Rittle-Johnson, Siegler, and Alibali (2001). This model portrays conceptual and procedural knowledge as influencing one another continuously throughout development in a bi-directional fashion, rather than one preceding the other. Growth in conceptual knowledge leads to increases in procedural knowledge and vice versa.

Bailey, Hansen, and Jordan (2017) examined whether this bi-directional relationship exists between fraction magnitude understanding (as measured by accuracy on a fraction number line task) and fraction arithmetic. Using a state-trait modeling approach, the researchers found evidence of transfer from fraction arithmetic skill to fraction magnitude understanding between two waves of measurement between fifth and sixth grade (i.e., the fall of fifth grade to the spring of fifth grade and the spring of fifth grade to the winter of sixth grade); transfer from fraction magnitude understanding to fraction arithmetic skill was found between the spring of fifth grade and the winter of sixth grade. These findings suggest a bi-directional relationship between fraction magnitude understanding and fraction arithmetic. In other words, children's knowledge of fraction magnitudes supports their learning of fraction arithmetic and vice versa.

A recent study examining pathways to fraction knowledge lends empirical support for considering fraction concepts and fraction procedures as distinct types of fraction knowledge (Ye et al., 2016). Sixth-grade fraction conceptual understanding was assessed with a measure that included various fraction concepts items, including set model items (e.g., "Shade 2/5 of ten circles") and estimation (e.g., "Estimate the sum: 7/8 + 12/13"). Sixth-grade fraction procedural understanding was assessed with fraction addition, subtraction, multiplication, and division items. Researchers used separate mediation analyses to explore pathways to fraction concepts knowledge and fraction procedural knowledge via third-grade cognitive skills (e.g., attentive behavior, verbal ability, nonverbal ability, and working memory) and fifth-grade numerical skills (e.g., magnitude reasoning and calculation). Distinct pathways for fraction conceptual knowledge and fraction procedural knowledge emerged. Whole number magnitude reasoning ability in fifth grade fully mediated the relationship between third-grade cognitive processes and sixth-grade fraction conceptual knowledge. In contrast, whole number multiplication and division abilities emerged as the key intermediaries between cognitive processes and sixth-grade fraction procedural knowledge.

Overall, the literature on fraction concepts and procedures suggests that they are relatively separate but mutually supportive competencies, with fraction concepts seeming to be most important to mathematics achievement (e.g., Hallet et al., 2010; Hecht et al., 2003).

Fraction Knowledge and Later Mathematics Achievement

In 2008, the National Mathematics Advisory Panel (NMAP) concluded that proficiency with fractions is foundational for later mathematics; in particular, NMAP highlighted the importance of fraction understanding for learning algebra. A few years later, Siegler and colleagues (2011) emphasized the importance of fraction magnitude understanding for overall numerical development. As described earlier, their integrated theory suggests that a vital part of numerical development is learning that many whole number properties (e.g., having one and only one successor) do not apply to all numbers. Importantly, the introduction of fractions is a child's first opportunity to learn the inconsistencies between whole numbers and other numbers. As such, the integrated theory implies that the acquisition of fraction knowledge is crucial to overall mathematics achievement. Yet, at the time of the NMAP (2008) publication and the proposal of the integrated theory of numerical development (Siegler et al., 2011), empirical support showing a link between fraction knowledge and later mathematics achievement was lacking (Booth & Newton, 2012). Since then, several studies have provided strong empirical support for the assertion that fraction knowledge predicts both algebra skill and later mathematics achievement more broadly.

Booth and Newton (2012) explored the relations between middle school students' fraction and whole number magnitude knowledge on algebra readiness. Students completed a fraction number line task (i.e., 0-1 number line) and two whole number line tasks (i.e., 0-100 and 0-6257). Algebra readiness was assessed by three measures: feature knowledge (e.g., knowledge of the equals sign), equation solving, and word problem

solving. The researchers found that fraction magnitude knowledge, more so than whole number magnitude knowledge, predicted early algebra skill. This finding supports the idea that fraction understanding, in particular, is important for later algebra achievement. Siegler and colleagues (2012) not only explored the relation between fraction knowledge and algebra performance but also the relation between fraction knowledge and overall mathematics achievement. Using two nationally representative longitudinal data sets (one from the United States and the one from the United Kingdom), knowledge of fraction arithmetic at age ten uniquely predicted algebra and overall mathematics achievement in high school even after controlling for family education, family income, intellectual abilities, and whole number arithmetic. Furthermore, the aforementioned study conducted by Resnick et al. (2016) shows that growth in fraction magnitude understanding, in particular, predicts student performance on a statewide, standardized test of general mathematics achievement. Thus, there exists empirical support for the idea that fraction knowledge benefits students' general mathematics achievement.

The relation between fraction understanding and algebra achievement is not unique to young children. In a recent investigation, Hurst and Cordes (2017) replicated this finding with adults. The researchers found that fraction skills remain an important predictor of algebra ability years after targeted fraction instruction in the classroom. This finding suggests that the predictive relationship between fraction understanding and algebra ability is not dependent on recent classroom instruction. Rather, the relationship holds even years after students' schooling in both fractions and basic algebra concepts. Researchers must ask *why* fraction knowledge emerges as a predictor of mathematics achievement more generally. Siegler and colleagues (2012) discuss four possible reasons why fraction knowledge uniquely predicts later mathematics performance. The first explanation is that measures of algebra knowledge and general mathematics achievement are essentially measuring fraction knowledge, as fractions are prevalent in later mathematics problems (e.g., $[6/8]x = 12$). A second and more speculative hypothesis is

that students who struggle with fractions in the intermediate grades may feel frustrated and hopeless in the area of mathematics; these students may give up trying in subsequent mathematics classes and rely on rote memorization. A third possibility is that the unique predictive value of fraction knowledge stems from fractions being an abstract and difficult topic for students, thus measuring more advanced thinking or general intelligence. A fourth and intriguing explanation is that fraction understanding is representative of an underlying structure of number that is essential for more advanced mathematics.

Siegler and colleagues (2012) found support for the fourth explanation and argue that students' who master fractions have a deep understanding of number that is essential for later mathematics more broadly. The researchers report that the predictive strength of students' fraction knowledge at age ten did not differ between students with greater and lesser mathematics achievement in high school. In other words, early fraction knowledge emerged as a unique predictor of later mathematics achievement regardless of ability level. Thus, the researchers claim that the relation between fraction knowledge and later mathematics is not a result of fractions being difficult to master. Rather, the predictive value of fractions seems to be due to fractions being essential to more advanced mathematics.

To extend the findings of Siegler et al. (2012), Bailey and colleagues (2012) tested whether measures of fraction knowledge are proxies for more general cognitive abilities, such as working memory. This explanation aligns with the hypothesis that fraction knowledge is essentially measuring general intelligence since fractions are a difficult topic. To test this possibility, Bailey and colleagues (2012) examined the extent to which students' performance on a fraction comparison measure predicted both mathematics achievement more broadly and word reading skills. If fraction competence represents general intelligence, then students' performance on the fraction comparison measure should be a unique predictor of later literacy skills. The researchers report that

sixth-graders' fraction comparison skill predicted one year gains in mathematics achievement, controlling for working memory and intelligence. Fraction comparison skill did not, however, predict students' word reading skills. They conclude that the findings are inconsistent with the hypothesis that fraction knowledge is a proxy of general intelligence. As such, the findings suggest that fraction competence holds unique importance for mathematics learning and achievement.

Screening for Mathematics Difficulties

Prior studies on mathematics screeners for identifying students at risk for mathematics difficulties have concentrated on kindergarten, first grade, and second grade (e.g., Bryant, Bryant, Gersten, Scammacca, & Chavez, 2008; Clarke, Baker, Smolkowski, & Chard, 2008; Jordan et al., 2008; Lembke & Foegen, 2009; Methe, Hintze, & Floyd, 2008; Seethaler & Fuchs, 2010; VanDerHayden et al., 2011). This focus on early grades is warranted, as school personnel want to identify at-risk students as early as possible in hopes of circumventing later difficulties. However, there is also a need for screening students in the intermediate grades. The following sections discuss the following: (a) the importance of mathematics screeners for students beyond second grade, (b) prior studies on mathematics screeners for the intermediate grades and (c) different categories of screening measures and an approach for improving the efficiency of these measures for classroom use.

Importance of Mathematics Screeners for the Intermediate Grades

In a review of prior studies exploring mathematics screeners, Gersten and colleagues (2011; 2012) urge future research to explore screeners for the intermediate grades. The authors caution readers that their review only addresses screeners for kindergarten through second grade and does not elucidate which students will succeed in mathematics in the early grades but struggle with more "intricate and abstract topics such

as those involving rational number" (Gersten et al., 2011, p.14). They conclude that future longitudinal studies must address these questions and student learning of more advanced mathematics content.

For example, imagine a student who is screened for mathematics difficulties along with her peers in the fall of first grade. Her performance on the screener demonstrates proficiency for basic number sense and as such, does not raise any concerns for her teacher. Now fast-forward to the fall of fifth grade. The student is experiencing new difficulties and is struggling to master fraction arithmetic and fraction concepts. Although the student's teacher considers her a low-performer in his classroom, the teacher decides that her performance does not warrant intervention. The student thus progresses into middle school mathematics. Unfortunately, without a strong understanding of fractions, her mathematics performance continues to decline when algebra is introduced in the curriculum.

The aforementioned example shows a student who meets mathematics benchmarks in the early grades but encounters difficulties in later grades when more abstract topics are introduced, such as fractions (Gersten et al., 2012). Without a valid screener measure in the intermediate grades, the teacher relied on his opinion to determine whether or not she needed additional supports. In educational settings, this occurrence is not uncommon; teachers and other educators often use opinion and intuition to decide which students in their classrooms should receive support services (Smolkowski & Cummings, 2015). However, studies of diagnostic decision-making suggest that judgments grounded on data and statistical models outperform judgments made on intuition alone (e.g., Grove, Zald, Lebow, Snitz, & Nelson, 2000). Decisions can be improved with the help of efficient diagnostic screeners (Smolkowski & Cummings, 2015).

Without valid screener measures for the intermediate grades, schools may "miss" students who are at risk for later difficulties and ill-prepared for more advanced

mathematics such as algebra. Identification of students at risk is the foundation of prevention, and the best method of identifying these students is with valid diagnostic screeners (Smolkowski & Cummings, 2015). Yet, the scarcity of research for intermediate grades in the area of mathematics screeners is striking. The next section explores the small subset of studies that have begun to address this gap in the literature.

Prior Studies Assessing Mathematics Screeners for the Intermediate Grades

The What Works Clearinghouse practice guide for assisting students struggling with mathematics cites only two studies that assess screeners beyond the second grade (Gersten et al., 2009). The first study mentioned in the practice guide investigated timed, one-minute measures of whole number facts (e.g., 6 – 1 = ?) administered in both third and fifth grades as predictors of performance on a statewide mathematics test at the end of each grade (Jiban & Deno, 2007). A limitation of the study is the type of statistical analysis used by the researchers. They conducted multiple regression analyses to assess whether the whole number facts measures predicted later mathematics achievement. Although regression analyses are powerful for assessing the strength of a predictor measure for an overall sample, they do not provide data on the accuracy of a screener for placing *individual* students into one of two populations of interest: students who are at risk and students who are not at risk (Jordan et al., 2010). The second study cited in the practice guide (Gersten et al., 2009) for providing data on mathematics screeners beyond second grade is a review written by Foegen, Jiban, and Deno (2007). However, similar to the Jiban and Deno (2007) study, the few studies mentioned in the review assess predictors of later mathematics achievement rather than exploring the measures as screeners for predicting individual student performance (e.g., Foegen & Deno, 2001).

Beyond the studies cited by the *What Works Clearinghouse* practice guide, only two additional studies were detected that assess mathematics screeners in the intermediate grades. Both studies are drawn from the same project that took place during

the 2002-2003 school year. Data were collected from first through fifth grade in two school districts (Keller-Margulis, Shapiro, & Hintze, 2008; Shapiro, Keller, Lutz, Santoro, & Hintze, 2006). Rather than using regression analyses, the researchers used a statistical analysis called receiver operating characteristic (ROC) curve analyses. This analysis is better suited for assessing the strength of a screener, as it provides data on the accuracy of a screener for predicting individual student membership into the at-risk population or the not at-risk population (Jordan et al., 2010; Smolkowski & Cummings, 2015).

Both studies used ROC curve analyses to assess the accuracy of curriculum-based measurements (CBM) in mathematics for predicting overall mathematics performance on a statewide assessment measure (Keller-Margulis et al., 2008; Shapiro et al., 2006). CBM is intended to mirror grade-appropriate skills that students use during everyday instruction and to provide repeated samples of student performance over time, rather than a snapshot (Shapiro et al., 2006). The first CBM was the Monitoring Basic Skills Progress (MBSP)—Math Computation (Fuchs, Hamlett, & Fuchs, 1998), which assesses student progress in mathematics computation. The measure increases in complexity through the grades. The probe for first grade and second grade consists of addition and subtraction problems, and the probe for third grade includes multiplication and division of whole numbers. The fourth grade measure includes fractions and multi-digit multiplication (Fuchs et al., 1998). The second CBM probe used in both studies was the MBSP—Math Concepts and Applications (Fuchs, Hamlett, & Fuchs, 1999). Items assess concepts such as counting, names of numbers, measurement, and fractions, with items increasing in difficulty across the grades (Shapiro et al., 2006).

The researchers' examination of fourth-grade CBM performance on later fifth-grade mathematics achievement provides an example of using ROC curve analyses to explore screener measures in intermediate grades. As mentioned above, ROC curve analyses provide data on the accuracy of a screener for predicting student membership

into one of two populations of interest: students who are at risk and students who are not at risk (Jordan et al., 2010). A ROC curve statistic called the area under the curve (AUC) is the recommended index of accuracy when assessing a screener (Pepe, 2003; Smolkowski & Cummings, 2015). The AUC is the overall *diagnostic accuracy* of a screener; for example, an AUC of .70 indicates that the screener accurately predicts student membership 70% of the time. In the study assessing CBM probes, Keller-Margulis and colleagues (2008) report the AUC values of the probes administered in the spring of fourth grade for predicting mathematics achievement in the spring of fifth grade. The researchers report an AUC of .79 for the computation probe, which means that the measure accurately identified students as either at risk or not at risk 79% of the time. The concepts probe yielded a slightly lower AUC of .72, meaning that the concepts measure correctly identified students 72% of the time. Both AUC values hover close to the .75 threshold recommended in the educational literature as indicative of good screeners for determining risk status (Cummings & Smolkowski, 2015). However, the authors do not report the AUC values for the fall or winter CBM probes. This omission raises a concern because screener measures are typically administered in the fall or winter to provide information of students' understanding at the beginning or middle of the academic year (Gersten et al., 2009). Without the data for the earlier administrations of the measures, the true value of these measures as screeners in the classroom is questionable.

The earlier 2006 paper from the same data project also used ROC curve analyses to explore the validity of fourth- and fifth-grade CBM probes; unfortunately, this particular study does not report the AUC values for any of the ROC analyses (Shapiro et al., 2006). The researchers report other ROC curve statistics for specific cut scores (e.g., the rate of true positives and true negatives; see Appendix C Glossary), but they do not provide the AUC values that represent the diagnostic accuracy of the screeners. Thus,

without the AUC values reported, the overall strength of the screeners cannot be determined.

Overall, the research on mathematics screeners for the intermediate grades is limited, and the results reported are not consistent across studies. Another important difference to note is the categories of mathematics screeners assessed in the studies; these categories are described next.

Categories of Screener Measures

Some of the aforementioned studies on mathematics screeners in the intermediate grades explored single-proficiency screeners and others assessed multiple-proficiency screeners. A single-proficiency screener assesses only one aspect of number competence (Gersten et al., 2012). An example of a single-proficiency screener is the measure of whole number facts (e.g., 5 – 1 = ?) assessed by Jiban and Deno (2007). Alternatively, the CBM probe administered by Shapiro and colleagues (2006) that assessed names of numbers, measurement, and fractions is an example of a multiple-proficiency screener (Gersten et al., 2011; 2012). Researchers have suggested that multiple-proficiency screeners may be more fruitful than a screener that targets only one discrete skill (e.g., Foegen et al., 2007). Purpura, Reid, Eiland, and Baroody (2015) argue that mathematics skills develop as a sequence of concepts and skills and as such, a measure that cover a broader range of mathematics content may serve as a stronger screener than a single-proficiency measure.

Another possibility for assessing multiple proficiencies is to combine screener measures. For example, two single- or multiple-proficiency screener measures can be combined and assessed within a study; an interesting question that arises is whether the combined measure makes more accurate predictions than either single measure alone. The general hypothesis is that multiple measures combined tap into more aspects of

students' knowledge base and thus are likely to be more predictive of students' later performance (Gersten et al., 2012).

Jenkins, Hudson, and Johnson (2007) urged researchers to explore this option, proposing that combining multiple measures may improve classification accuracy. For example, in a study to predict students' reading difficulties, Speece and colleagues (2011) used logistic regression to assess the classification accuracy of three predictor measures combined: the Word Identification subtest of the Woodcook-Johnson Reading Test (Woodcock, 1998), teachers' rating of students' overall reading ability, and word identification fluency. They found that the combination of measures had a superior diagnostic accuracy as compared to each measure individually.

In a recent publication, Clemens and colleagues (2016) also recommend assessing combinations of measures. Although they discuss the importance of both mathematics and reading screeners, the example they provide is focused on screener measures for reading difficulties. They assessed three separate screener measures: letter-sound fluency, letter-naming fluency, and phoneme segmentation fluency. They also used logistic regression to combine the predictors and determine which were statistically significant in the prediction of their outcome variable, which was first-grade reading fluency. The researchers report that the best-fitting model included two screener measures, letter-sound fluency and letter-naming fluency. By assessing this new combination of measures using both logistic regression and a ROC curve analysis, the researchers reported that the combination of multiple measures yielded slightly higher diagnostic accuracy than either measure alone (Clemens et al., 2016). The researchers provided a helpful matrix for making interpretations of students' later reading achievement; the matrix shows ranges of scores along both screener measures and allows a teacher or researcher to make a prediction based on a student's scores on each screener.

As such, there is support for assessing a combination of multiple screener measures in hopes of improving diagnostic accuracy. However, even though many

researchers support this notion, they are simultaneously wary of its tradeoffs (e.g., Jenkins et al., 2007; Speece et al., 2012). For example, Jenkins and colleagues (2007) mention that although multiple measurements may improve screening precision, administering more than one measure is far less efficient than administering one single measure. Clemens et al. (2016) warn researchers and schools that the administration of multiple measures requires additional resources and time. Furthermore, the use of several different measures for screening can lead to confusion about how to interpret scores.

Fortunately, Clemens et al. (2016) provide a helpful solution for the concern of interpreting student scores from a combination of measures. They use logistic regression not only to evaluate the multiple measures combined, but also to provide a multiple-measure matrix that helps teachers interpret students' scores. As mentioned previously, the researchers assessed a combination of a letter-sound fluency measure and a letter-naming fluency measure for the prediction of meeting a first-grade reading criterion. The researchers report a multiple-measure matrix that provides predicted probabilities of meeting the first-grade reading criterion for different combinations of scores on each measure. A predicted probability indicates the likelihood of a student meeting the reading criterion and is determined by the linear combination of the two measures as computed within the logistic regression (Clemens et al., 2016). For example, a student who scores within a certain range on the letter-sound fluency measure (e.g., 20-24) and also scores within a certain range on the letter-naming fluency measure (e.g., 35-39) has a predicted probability ranging from 0.84 to 0.89. This predicted probability range indicates that the student has an 84%-89% chance of meeting the reading criterion. Teachers can therefore use the matrix to interpret any student's scores on the two measures.

While Clemens and colleagues (2016) provide a useful solution for interpreting students' scores when using a combination of screening measures, they do not provide a solution for the issue of needing additional time and resources to administer more than

one screener measure. The researchers conclude that combining multiple measures necessitates extra resources and is hence less efficient and practical for classroom use.

Improving the Efficiency of a Combination of Measures

A possible solution to the concern of multiple screener measures being less efficient than single screener measures is provided by Purpura and colleagues (2015). The researchers propose that the efficiency and practicality of a screener measure can be improved by reducing the number of items on the measure. For example, they assessed preschool-aged children on 25 measures of early numeracy skills with a total of 143 items. Using item response theory (IRT) analysis in Mplus (Muthen & Muthen, 2012), they reduced the combined measure to only 24 items while simultaneously retaining the strength of the full predictor measure. In other words, the researchers modified a long 143-item assessment to a brief 24-item assessment without sacrificing the screener's diagnostic accuracy. The administration of this new shorter measure would of course require less time to administer in the classroom and less time to score, hence saving valuable resources. This approach of reducing the amount of items on the screener measure can be applied to a single measure alone (e.g., reducing a 20-item measure to only 11-items) or to two measures combined (e.g., reducing the amount of items on Measure A and reducing the amount of items on Measure B). Importantly, Purpura and colleagues (2015) recommend that researchers assess the diagnostic accuracy of the shorter screener and compare it to the accuracy of the original measure(s); these steps can be achieved with logistic regression and ROC curve analyses (Wilson et al., 2016).

Overall, researchers have begun testing new approaches for improving screening measures and for considering classroom practicality. The present study follows suit by assessing shortened versions of screeners with high diagnostic accuracy and testing combinations of the shortened screener measures (when deemed appropriate by statistical thresholds). The analysis that is foundational to these goals of the present study is

receiver operating characteristic (ROC) curve analysis. The next section provides a detailed introduction to ROC curve analyses.

ROC Curve Analyses

ROC curve analyses are recognized as the state-of-the-art method for describing the accuracy of a diagnostic test (Weinstein, Obuchowski, & Lieber, 2005). The purpose of a diagnostic test is to screen for the presence or the absence of a certain condition, event, or risk-status. ROC curve analyses are widely used in research predicting reading difficulties (e.g., Cummings & Smolkowski, 2015; Wilson et al., 2016) and in research predicting mathematics difficulties (e.g., Jordan et al., 2010; Seethaler & Fuchs, 2010). The ROC method of analysis is also prevalent in clinical decision-making in medicine, such as confirming the presence of a disease and ruling out the disease in healthy individuals (e.g., Jeffries et al., 2015; Youngstrom, 2014).

Recently, two articles were published with a clear shared goal: to describe ROC curve analyses for the purpose of statistically evaluating diagnostic tests and to share recommendations for reporting ROC statistics for the purpose of applying the statistics to real-world situations. One of the publications is written from a medical standpoint (Youngstrom, 2014) and the other from an educational perspective (Cummings & Smolkowski, 2015). From the medical literature, Youngstrom (2014) provides recommendations for using ROC curve analyses in clinical situations. Youngstrom's recommendations are based on and add to the Standards for Reporting of Diagnostic Accuracy (STARD) published over a decade earlier by different researchers in the medical field (Bossuyt et al., 2003). On the other hand, Cummings and Smolkowski (2015) provide guidelines that center on educational-decision making, meaning the identification of risk for reading difficulties and need of additional support. Both studies not only focus on how to conduct ROC analyses but also seek to help others apply the ROC statistics in real-world settings. The foundation of both sets of guidelines is the

ROC analysis itself, which is inherently the same regardless of discipline. The subsequent sections address the following: (a) important statistical terms that are foundational to the ROC curve analysis regardless of discipline, (b) how to leverage ROC statistics for the selection of a screener cut score in both clinical and educational settings, and (c) a recommendation for improving how researchers report ROC statistics in the educational literature.

Important Terms associated with ROC Curve Analyses

Several statistical concepts are foundational to ROC curve analyses and for assessing the power of a diagnostic measure, regardless of the discipline or the type of diagnostic test. Two terms associated with ROC curve analyses across disciplines are a *positive* test result and a *negative* test result. The most well-known and familiar example of these terms is likely set in a clinical context. As mentioned above, clinical diagnostic tests are used to screen for two possible outcomes: the presence of a disease or the absence of the disease. When a diagnostic test confirms the presence of a disease, the patient is told that he/she screened "positive." If the test rules out the disease, the patient is told he/she screened "negative." To summarize, a positive test result indicates the likely presence of the disease and the negative test result indicates its likely absence. These terms also apply to an educational context. In the educational literature, a diagnostic test is called a screener. The purpose of a screener is to place students into one of two groups: students who are at risk for later difficulties and students who are not at risk. Although both student groups are hence important, the primary focus of a screener is to identify students who are at risk (Cummings & Smolkowski, 2015). As such, a positive screener result means that the student is likely at risk for later difficulties and a negative result means that the student is likely *not* at risk.

ROC curve analyses assess how accurately a screener predicts student membership into the two groups. Importantly, the two groups must be defined by a valid

criterion or outcome measure. For use in a ROC curve analysis, the outcome must be dichotomous (Youngstrom, 2014). The most widely-used outcome measure in education is an end-of-the-year state achievement test, for which each student is categorized as either meeting the proficiency standard for his/her grade or not meeting the standard (e.g., Cummings & Smolkowski, 2015; Jordan et al., 2010).

A powerful advantage of ROC curve analyses is the ability to select a specific screener cut score for predicting student performance on the outcome measure. A cut score is a selected score on the screener that separates positive test results (i.e., likely to *not* meet the mathematics standard at the end of the year) from negative test results (i.e., likely to meet the standard). For example, imagine a study for which researchers are assessing a ten-item mathematics screener to predict risk for later mathematics difficulties. They define risk status by whether or not students meet a proficiency standard on a mathematics achievement test at the end of the school year. A cut score of five points on the screener means that a student score that falls at or below five points would be categorized as a positive screener result, and a student score that exceeds the cut score would be categorized as a negative screener result.

In considering all possible students in the population, some of the predictions based on screener and outcome performance will be true and others will be false; no screener and/or cut score will have perfect predictive accuracy (Cummings & Smolkowski, 2015). For example, consider a student who scored below the screener cut score (i.e., positive screener result) but actually *met* the mathematics standard at the end of the year. This scenario indicates that the ROC analysis misidentified the student, placing him/her in the wrong student group. In other words, the positive screener result was false. Overall, since there are two possible screener results (i.e., positive screener result vs. negative screener result) and two possible outcome results (i.e., positive outcome result vs. negative outcome result), there is a total of four potential scenarios (See Appendix C for the 2x2 matrix): (a) true positive (b) false positive, (c) true negative

and (d) false negative. The values of these statistics change with each possible screener cut score. That is, a cut score of five points on a screener will have a different proportion of true positives than a cut score of six points on the same measure. Descriptions and examples of all four of the statistics are described next within an educational context.

True Positive

The first possible scenario is a *true positive,* meaning that a student who is truly at risk is correctly identified by a positive screener result. When considering the proportion of true positives in the total population, the word "fraction" is often added to the term; for example, *true positive fraction* (TPF) refers to the proportion of true positives in the total population. Another popular term for this statistic is the *sensitivity* value. For ease of understanding, the present study primarily relies on the terms *true positive* and *true positive fraction*. As an example, a .90 rate of true positives means 90% of students who are truly at risk are identified accurately by a positive screener result. In other words, 90% of students who failed to meet the mathematics standard scored at or below the screener cut score.

False Positive

When assessing the accuracy of a screener, a second possible scenario is a *false positive.* A false positive refers to a student who is *misidentified* by the ROC analysis. In particular, the student is not at risk but is misidentified by scoring below the screener cut score (i.e., a positive screener result). The proportion of false positives that occurs in a population is called the *false positive fraction* (FPF). A false positive fraction of .20 means that 20% of students who met the mathematics standard are misidentified by a positive screener result. These students would be identified as needing additional supports even though they are not truly at risk. This scenario is not ideal because schools or researchers would be using time and resources to help students who do not require

intervention. Since no screener or cut score can have perfect accuracy, a certain proportion of false positives is inevitable in any student population. Ideally, researchers select a screener cut score that minimizes false positives.

True Negative

A third possible scenario is that a student who is not at risk is correctly identified with a negative screener result; this scenario is called a *true negative*. The proportion of true negatives in the population is the *true negative fraction* (TNF). This statistic is also popularly known as the *specificity* value. For consistency and ease of interpretation, the present study primarily uses the terms *true negative* and *true negative fraction*. As an example, a true negative fraction of .60 means 60% of students who are not at risk are identified accurately by the screener cut score. That is, 60% of students who met the mathematics standard received a negative screener result.

False Negative

Out of the four possible scenarios, a *false negative* is often the most undesirable. A false negative means that a student who is truly at risk is *misidentified* by the screener cut score. In other words, the student scored above the screener cut score (i.e., negative screener result) but does *not* meet the mathematics standards at the end of the year. As such, the student is mistakenly categorized as not at risk and is not identified as needing the additional supports that he/she needs for future success. A likely explanation for this misidentification is that the student scored just above the screener cut score, such as scoring six points when the established cut score was five points. False negatives are highly undesirable in education because students who actually need additional supports are misidentified as on track for future mathematics success. A certain proportion of false negatives, called the *false negative fraction* (FNF), is inevitable for all screener measures and all cut scores, as no screener will have perfect accuracy. Fortunately, a main

advantage of ROC curve analyses is that researchers, schools, or clinicians can select a cut score that minimizes or maximizes certain values based on the setting, goals, and resources.

Selecting a Cut Score based on the Setting, Goals, and Resources

A main advantage of using ROC curve analyses is that the selection of a diagnostic test's cut score is based not only on true positives and false negatives, but also the constraints of an educational or medical situation. For example, if a school wants to make sure an intervention is provided to as many at-risk students in the population as possible, a cut score can be selected that "over-identifies" students as at risk. In other words, the school can cast a "wide net" for identifying students in need of intervention. To achieve this goal, the school selects a cut score that maximizes true positives (i.e., sensitivity). A second advantage of casting this wide net and maximizing the rate of true positives is that the likelihood of "missing" at-risk students decreases. In the educational literature surrounding mathematics screeners for young children, researchers often emphasize the importance of maximizing true positives when selecting a cut score because they want to identify as many at-risk students as possible and avoid misidentifying these students. For example, Jordan et al. (2010) suggest that a cut score correspond, at minimum, to an 85% rate of true positives. This value indicates that 85% of students who are at risk are correctly identified and only 15% of at-risk students are misidentified (i.e., false negatives).

Researchers who choose to maximize the proportion of true positives must accept certain tradeoffs. For example, as the rate of true positives increases, the rate of false positives also increases. A false positive is a student who is misidentified as at risk. In the scenario of an educational intervention, this student will be identified as a potential participant for the intervention. If a screener identifies 50 cases of false positives in a population, these 50 students will be identified as needing additional supports even

though they are not actually at risk. The school will spend unnecessary money and use resources such as time and money to provide supports for these children who are not struggling in mathematics. Thus, the selection of cut scores involves a balance of advantages and disadvantages; by considering all the possible tradeoffs in the context of one's available resources, the most advantageous cut score can be selected.

Another real-world possibility is that a school may not have the financial resources or time to cast such a wide net and to provide an intervention to a large amount of students. In response, a lower cut score associated with a lower proportion of true positives can be selected, meaning that less students overall will be identified as needing the intervention. In other words, by choosing a lower cut score, a smaller proportion of students will fall below the cut score and hence be identified as at risk. This "smaller net" unfortunately implies that the school will be missing a greater amount of students who truly are in need of intervention but are misidentified by the screener cut score (i.e., false negatives). However, this increase in false negatives may be an acceptable or necessary tradeoff for the school if they have finite resources for administering the intervention.

A diagnostic test in a clinical scenario may have much higher stakes than an educational screener. For example, a diagnostic test that screens for the presence of a life-threatening condition will certainly impact how the researchers or clinicians select a cut score. Patients who receive a positive result on the diagnostic test are identified as likely having the condition and as a result, may be identified as needing further testing and/or a lifesaving treatment. In this scenario, clinicians are likely to cast a wide net to save as many lives as possible and to avoid missing patients who truly do have the condition. Again, the tradeoff here when casting a wide net is that the clinicians will also identify a high proportion of false positives. This tradeoff is seemingly minor when considering the alternative. However, if a clinician casts too large a net, then many healthy patients will receive the terrifying news of a positive test result. High false positive rates may weaken the reputation of the diagnostic test and discourage patients from taking the test

altogether. As such, there is certainly a delicate balance when making cut score decisions based on ROC curve statistics. Overall, whether considering a medical context or an educational context, ROC curve statistics have important implications for selecting cut scores that will be applied in real-world scenarios.

A Limitation of the Four Primary ROC curve statistics

The four ROC curve statistics described thus far (i.e., true positive, false positive, true negative, false negative) are essential for determining the accuracy of a screener and for considering tradeoffs when selecting a screener cut score. As such, the values of the four statistics are vital for research purposes. However, the statistics are not as helpful for teachers or clinicians as they are for researchers. The reason for this distinction is that the interpretation of each statistic is *backward* rather than *predictive*. On a broad level, this distinction means that the statistics use the outcome measure to reason backwards about the screener measure that was administered months or years earlier.

For example, consider the interpretation of the true positive rate (i.e., sensitivity) for a mathematics screener. The true positive value is not calculated using the entire student population; rather, the statistic is calculated using the sub population of students who do *not* meet the mathematics standard at the end of the year. In other words, the true positive fraction is concerned only with the sub population of students who are truly at risk. Imagine a sub population of 100 students who are truly at risk. If only 80 out of the 100 students screened positive on the screener that was administered *before* the outcome measure, then the true positive fraction equals .80 (i.e., 80/100). Following this calculation, the interpretation of the true positive fraction is *backwards*; that is, 80% of students who are truly at risk had previously screened positive on the screener.

Considering only 80 out of the 100 at-risk students screened positive on the screener, this indicates that 20 of the at-risk students are misidentified by a negative screener result (i.e., 100 - 80 = 20). These 20 students are examples of false negatives.

Since 20 out of the 100 students are categorized as false negatives, the false negative fraction is .20 (i.e., 20/100). The amount of true positives and the amount of false negatives always equals the sub population of students who are truly at risk (e.g., 80 + 20 = 100). In other words, the proportion of true positives plus the proportion of false negatives will always equal one (e.g., .80 + .20 = 1.00). This relationship holds true for all ROC curve analyses. A parallel relationship exists between the proportion of false positives and true negatives; adding the two values together always yields a sum of one.

Overall, the rates of true positives, false positives, true negatives, and false negatives are essential for research purposes. However, the backwards interpretations are not as beneficial for teachers or clinicians who want to administer a diagnostic test and make a forward prediction based an individual's score on the screener. Fortunately, ROC curve analyses *do* yield such predictive statistics.

ROC Statistics for making Predictive Interpretations

Other ROC curve statistics reported in both the educational and clinical literature are the positive predictive power (PPV) and negative predictive power (NPV). These statistics are desirable because they offer *forward* interpretations of an individual's performance on a diagnostic test. There is a PPV associated with every cut score along a diagnostic test. For illustration, consider again the example in which a 10-item mathematics screener is predicting students' performance on a mathematics achievement outcome. Imagine a .75 positive predictive power is associated with a cut score of five. For a student scoring below the cut score, the teacher can say that he/she has a 75% chance of being at risk for later mathematics difficulties. Notice that the positive predictive power is a forward *predictive* interpretation of an individual student's performance on the screener.

There is also a statistic that allows for a predictive interpretation for scores above the cut score. This statistic is called the negative predictive power (NPV); this statistic is

not as discussed in the literature as the positive predictive power because it is concerned only with the subpopulation of students who are not at risk. For example, a negative predictive power of .80 indicates that a student scoring above the cut score has an 80% chance of meeting the later mathematics standards.

The predictive interpretations of a student's screener score makes the positive predictive power and negative predictive power desirable statistics for both education and clinical decision-making. However, the statistics share a major limitation: the statistics are dependent on the base rate (see Appendix C for equations). In an educational context, the base rate is the prevalence of students in the sample who are truly at risk. Thus, these two values cannot be generalized to different samples with different base rates (Youngstrom, 2014). Fortunately, both the educational literature and the clinical literature provide ways for other people to compute the positive predictive power (PPV) for their own sample.

Guidelines for Computing the Positive Predictive Power (PPV) for Different Samples

The educational literature provides teachers with guidelines for computing the positive predictive power (PPV) for their own samples of students, and the clinical literature provides clinicians with guidelines for predicting this statistic for their patients. Even though both guidelines yield the same statistic, the guidelines for computing the statistic vary greatly. The medical literature provides a simpler and more understandable approach that has not yet been recommended in studies of educational screeners.

In the educational guidelines for reporting ROC analyses, Cummings and Smolkowski (2015) provide teachers with intimidating formulas for computing the positive predictive power and the negative predictive power. The formulas are provided below, just as they are provided within the article:

$$PPV = \rho TPF / (\rho TPF + (1 - \rho)FPF)$$

$$NPV = (1 - \rho)(TNF) / ((1 - \rho)(TNF) + \rho(FNF))$$

Where ρ = *base rate*

TPF = *true positive fraction*

FPF = *false positive fraction*

TNF = *true negative fraction*

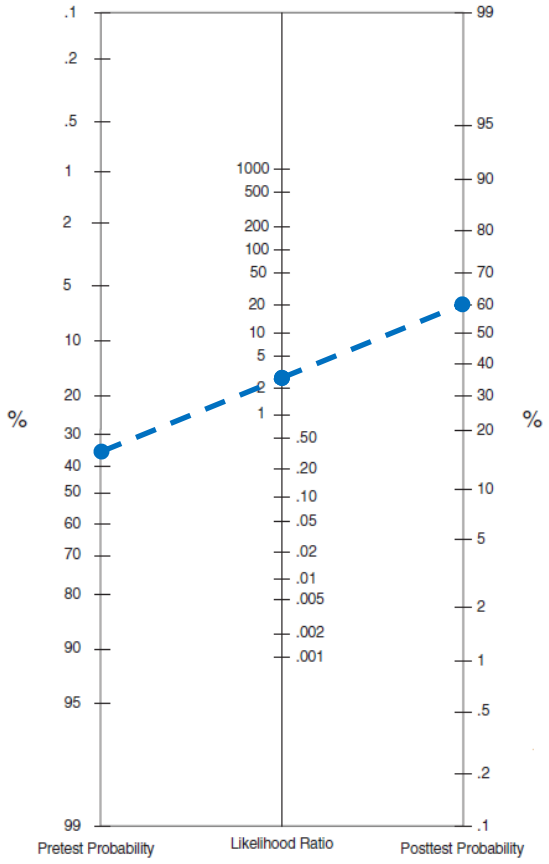
FNF = *false negative fraction*

Cummings and Smolkowski (2015) instruct teachers to first find a research article that reports ROC statistics for a screener measure of interest. Next, the teachers must identify and enter three ROC statistics reported within the journal article for a screener score of interest into the formula (i.e., true positive fraction, true negative fraction, and false positive fraction). Finally, they must then enter their own sample's base rate (e.g., the proportion of students who are at risk in their school, which can be estimated by student performance from the previous academic year). The complex appearance of the formula itself may create barriers for some individuals who are not familiar with research study statistics and/or have mathematics anxiety.

In the clinical guidelines, Youngstrom (2014) provides a different method for computing the positive predictive power for a cut score (and the negative predictive power). The clinical article recommends that research studies report diagnostic likelihood ratio (DLR) statistics that are independent of the sample's base rate. By providing DLRs, readers can then determine the positive predictive power by using one simple figure rather than a complex, intimidating mathematical formula. The simple figure is called a "probability nomogram (See Figure 1). An example is provided by following the dashed lines on Figure 1, starting on the left and extending to the right. On the left side of the nomogram, the teacher locates his/her sample's base rate (e.g., 35%) and "draws" a line that extends to the middle of the nomogram and matches the DLR+ from the published study (e.g., 3.00). The teacher continues extending the line to the right side of the

nomogram; the point at which the dotted line crosses the right side of the nomogram tells the teacher the positive predictive power for his/her own sample of students (e.g., 60%). DLRs and probability nomograms are praised in the clinical literature as a more intelligible way for clinicians to apply ROC statistics (Florkowski, 2008); yet, the practical benefits of DLRs and the nomograms have not been reported in the educational literature.

Figure 1 Probability Nomogram



Overall, to gain predictive statistics from ROC curve analyses, the clinical literature provides a helpful figure for clinicians to use in practice, while the educational literature asks teachers to plug three different values into a complex formula that likely appears daunting. The study demonstrates the importance of DLR statistics, along with the probability nomogram, being reported in educational research assessing screener

measures. For use in the classroom, ROC statistics must be transformed into findings that are feasible and useful for teachers. The study demonstrates how bringing DLR values and probability nomograms into the educational literature can improve the translation of ROC statistics into real-world educational contexts. To date, such analyses have not been applied to educational contexts.

The Present Study

The present study uses three years of longitudinal data (i.e., fourth through sixth grade) from the Center for Improving Learning of Fractions, funded by the Institute of Education Sciences (Professor Nancy C. Jordan, Principal Investigator). The study assesses three predictor measures of fraction understanding that were administered to students in the fall or winter of each grade (i.e., two fraction concepts measures and one fraction arithmetic measure) as potential screener measures for predicting later mathematics achievement. Specifically, the study assesses the strength of each fraction measure for predicting students' later performance on an end-of-the-year state mathematics achievement test. The following sections first provide the rationale for each fraction task being considered as a potential screener measure and then the rationale for the statistical analyses used for assessing these screener measures.

Basis for Fraction Measures Considered

As discussed in the previous section in the literature review, researchers have demonstrated the importance of both fraction concepts and fraction procedures for success with fractions (e.g., Hecht et al., 2003). In light of the importance of different types of mathematical knowledge, the present study assesses two measures of fraction concepts and one measure of fraction arithmetic. One measure of fraction conceptual understanding is a fraction number line estimation task; this task is a single proficiency measure that assesses students' accuracy when estimating fraction magnitudes on a

number line, from 0 to 1 and 0 to 2. The second measure of fraction conceptual understanding is considered a multiple-proficiency measure that assesses students on multiple fraction conceptual items, such as part-whole understanding, fractions as magnitudes, and fraction equivalency.

Rationale for Analyses

To assess the three fraction measures as potential screeners, the present study follows recommendations in the literature for conducting receiver operating characteristic (ROC) curve analyses. ROC curve analyses are promoted in multiple disciplines as the best method for determining the accuracy of a single diagnostic test in making predictions (Weinstein et al., 2005). For example, ROC curve analyses are employed to assess the accuracy of clinical diagnostic tests for confirming the presence of a disease in patients (e.g., Hajian-Tilaki, 2013), for assessing sensors for the detection of earthquakes (e.g., Faulkner et al., 2011) and for assessing educational screeners for the identification of students at risk for later difficulties (e.g., Cummings & Smolkowski, 2015).

If multiple fraction measures emerge as strong screener measures in the present study (as identified by certain thresholds and statistical tests outlined in the Data Analysis section), the present study will complete the following steps: (a) determine the best subset of items for each measure and (b) assess a combination of the best subset measures for predicting later mathematics achievement using both logistic regression analyses and ROC curve analyses (e.g., Clemens et al., 2016; Wilson et al., 2016).

Potential Implications for Education

The present study addresses an important gap in the literature: the need for mathematics screeners for the intermediate grades (e.g., Keller-Margulis et al., 2008; Shapiro et al., 2006). Fractions represent a major portion of the mathematics curriculum in fourth through sixth grades. By assessing different components of fraction

understanding, the study aims to identify an efficient and easy to use fraction screener measure (either a single measure or a combined-proficiency measure) that will help teachers identify students who are at risk for not meeting proficiency standards in mathematics. Many students struggle with mathematics in the intermediate grades, especially struggle with fractions (e.g., Hansen et al., in press). A powerful screener measure is the first step for addressing this educational concern. By identifying students who are likely to experience later mathematics difficulties, schools can then provide additional supports.

The fraction measure or combination of fraction measures that emerge as a powerful screener may also have implications for future research and for classroom instruction. For example, if the fraction number line estimation task emerges as most predictive of later performance, then future research should explore the underlying reasons for this strong predictive power. Also, this potential finding would imply that classroom instruction should support skills that help students' understanding of fraction magnitudes, such as the relation between the numerator and denominator (DeWolf, Grounds, Bassok, & Holyoak, 2013).

The present study also seeks to inform future research studies that assess mathematics screeners. In particular, the study aims to provide guidelines for educational research that draw from recommendations found within both the educational literature and the clinical literature. In particular, the present study urges educational researchers to report screener statistics in ways that help schools translate the findings into usable information. Importantly, by improving the ways in which researchers report ROC statistics, we may increase the likelihood of schools making educational decisions (e.g., recommending a student for intervention) that are driven by data rather than by intuition or judgement alone (Smolkowski & Cummings, 2015).

Research Questions

In summary, the present study addresses the following research questions:

1. What is the diagnostic accuracy of fraction measures (i.e., two measures of fraction concepts and one measure of fraction arithmetic) given in the fall/winter of fourth, fifth, and sixth grades for identifying students who did not meet the mathematics proficiency standard on a state test administered in the spring of each grade? In particular, which fraction measure or combination of measures holds the highest diagnostic power in each grade for identifying at-risk students? Based on the literature, it was predicted that the fraction number line estimation task in all three grades would emerge as the strongest predictor measure of students' later mathematics achievement. The second fraction concepts measure was also expected to emerge as a strong screener measure since it taps into multiple aspects of fraction conceptual understanding. The fraction arithmetic measure was hypothesized to be the weakest screener measure for each grade. These three hypotheses align with the integrated theory of numerical development (Siegler & Lortie-Forgues, 2014) and prior literature demonstrating the importance of fraction conceptual understanding for later mathematics achievement (e.g., Hallet et al. 2010; Siegler et al., 2012).
2. Can the measures identified in Research Question 1 be improved psychometrically by reducing the number of items while simultaneously retaining or improving the diagnostic accuracy of the measure?
3. How can the resulting findings help educational researchers improve the ways in which they report results so that the information is accessible and usable for schools and teachers?

Chapter 3

METHOD

Students were drawn from nine elementary schools within two Delaware school districts serving families of diverse socioeconomic backgrounds. As a part of a larger longitudinal study, data collection began in third grade. All third-grade students from participating schools were sent an IRB approved informed consent letter requesting their participation in the study. A total of 517 returned consent forms to participate in the study, of whom 36 opted out before the first assessment. Students were then followed through sixth grade. By the end of third grade, 23 students dropped out of the study, by the end of fourth grade an additional 68 students dropped out, and by the end of sixth grade an additional 39 children dropped out. Attrition was due to students moving to another school district (67%), a lack of information regarding students' transition into middle school for sixth grade (23%), and students withdrawing from the study (10%). The sample was replenished in fourth grade (n = 27 new children) and again in fifth grade (n = 28 new children). In total, the sample for the present study included 536 students. In fourth grade and again in fifth grade, the same informed consent letter was sent out to replenish the sample. Twenty-seven new children joined the study in fourth grade and 28 new children in fifth grade, resulting in a total sample of 536 students.

Student demographic information for the total 536 students is presented in Table 1. Attrition rates and missing data result in slightly different total students included in each of the ROC analysis; total students included in each analysis will be reported in the Results section. Reportedly, participating schools followed curriculum benchmarks aligned with the Common Core State Standards in Mathematics (NGACBP & CCSSO, 2010) starting in fourth grade.

Table 1 Demographic Information for Total Longitudinal Study

Characteristic	%
Gender	
Male	47.0
Female	53.0
Race	
White	51.8
Black	40.0
Asian/Pacific Island	5.7
American Indian/Alaskan Native	2.5
Hispanic	17.7
Low Income	60.9
English Learner	10.6
Special Education	10.6
Learning Disability	5.8
Mean Age in Months	105.9
Total N	536

Screening Measures Administered across Grades

The screening measures included students' performance on two measures of fraction concepts and one measure of fraction arithmetic.

Fraction Concepts

NAEP Fraction Concepts

A paper and pencil measure of released items from the National Assessment of Educational Progress (NAEP; U.S. Department of Education, 1990-2009) measured fraction concepts. The NAEP is administered across the United States in fourth and eighth grades, with items ranked from easy to hard.

NAEP items in the longitudinal study assessed part-whole understanding of area models (e.g., "Which shows 3/4 of the picture shaded"), set models (e.g., "What fraction of the group of umbrellas is closed?"), equivalence (e.g., "These three fractions are

equivalent. Write two more fractions that are equivalent to these"), fraction magnitude (e.g., "On the number line above, what number does *P* represent"), estimation (e.g., "Which fraction has a value closest to 1/2?"), and comparison and ordering (e.g., "In which of the following are the three fractions arranged from least to greatest?"). The measure included 18 total items in fourth grade; these items were consistent across all time points, but additional items were included in fifth and sixth grades to align with the instruction students were receiving in school and to avoid ceiling effects. In fifth grade, one item was added to the measure for a total of 19 items. In sixth grade, 5 items were added, making a total 24 items. See Appendix A for a list of all NAEP items included at each grade.

Items were read aloud in a group setting. Administration lasted for approximately 35 minutes in each grade; thus, students were allotted approximately 2 minutes per item in fourth and fifth grades and approximately 1.5 minutes per item in sixth grade. Students earned one point for each correct response. The measure had high internal reliability in each grade (α = .78 in fourth grade; .78 in fifth grade; .84 in sixth grade) Items are publically available through the NAEP website (https://nces.ed.gov/nationsreportcard/).

Fraction Number Line Estimation

A fraction number line estimation (FNLE) task adapted from Siegler et al. (2011) was administered on a laptop computer using DirectRT v2012. Each number line was 17.5cm long and presented in the middle of the laptop screen. Fractions were presented one at a time beneath the middle of the number line. For each item, the cursor was set at "0"; students used the arrow keys to slide the cursor along the number line and then pressed a different key to indicate their final estimation. After providing their response, a new blank number line and a new fraction were presented and the cursor was reset to "0". Students had no time constraints to make their individual estimates, but most students

responded with 5 seconds per trial. The total administration time was approximately 2-3 minutes.

The items assessed on the number line task were the same across the three grade levels. Students estimated the location of nine fractions (1/5, 13/14, 2/13, 3/7, 5/8, 1/3, 1/2, 1/19, and 5/6) on a 0-1 number line and 19 fractions and mixed numbers (1/3, 7/4, 12/13, 1 11/12, 3/2, 5/6, 5/5, 1/2, 7/6, 1 2/4, 1, 3/8, 1 5/8, 2/3, 1 1/5, 7/9, 1/19, 1 5/6, and 4/3) on a 0-2 number line. All estimations were combined to create a single score, which had high internal reliability (α = .91 in fourth grade; .98 in fifth grade; .95 in sixth grade). Scores were calculated as the mean percent absolute error (PAE). The mean PAE was calculated by dividing the absolute value of the difference between the estimated position and actual position by the numerical range of the number line (1 or 2), multiplying by one hundred for each item, and averaging across all trials). For example, if a child was asked to locate 3/2 on a 0 to 2 line and marked the location corresponding to 5/4, the PAE for this individual item would be 12.5% [|(1.5 – 1.25)|/2 x 100]. The computer program DirectRT provided the location of each estimation in pixels; these estimations were transformed into PAE using both Excel and the Statistical Package for the Social Sciences (SPSS). Higher percent absolute error indicates poorer performance.

Fraction Arithmetic

The paper and pencil fraction measure was adapted from Hecht (1998). In fourth grade, there were four addition (e.g., 2/5 + 1/5) and four subtraction (e.g., 3/4 - 1/4) computation items, all of which involved fractions with the same denominators. In fifth grade, two items were with unlike denominators (e.g., 5/6 + 2/3; 7/8 – 1/2) were included, making a total of 10 items. In the sixth grade, there was a total of 26 items. One addition (3/4 + 2/3), one subtraction (1 1/3 – 4/5), nine multiplication (e.g., 3 x 1/3), and five division (e.g., 1/6 ÷ 3) items were added. Administration in fourth and fifth grade lasted approximately 10 minutes; administration in sixth grade lasted approximately 30 minutes.

See Appendix B for a list of fraction arithmetic items at each grade. At each time point, the measure was reliable (α = .95 in fourth grade; .84 in fifth grade; .88 in sixth grade).

Outcome Mathematics Achievement Measure (DCAS)

In spring of each grade, students' performance was assessed on the mathematics section of Delaware Comprehensive Assessment System (DCAS; American Institutes for Research, 2012), a statewide test of mathematics achievement. The DCAS requires students to answer multiple choice questions that assess algebraic reasoning (e.g., find a given term in an arithmetic sequence), numeric reasoning (e.g., using and applying meanings of multiplication and division), geometric reasoning (e.g., analyze and classify two-dimensional shapes according to their properties), and quantitative reasoning (e.g., construct and use data displays) (American Institutes for Research, 2012). The DCAS does not report further information regarding the proportion of items assessing certain mathematics topics on each grade of the assessment. However, the algebraic reasoning and numeric reasoning categories likely involve fraction items in fourth, fifth, and sixth grades. The Common Core State Standards point to other mathematics topics and skills likely assessed on the DCAS in addition to fraction items, including: (a) find factor pairs for a whole number in fourth grade, (b) understand operations with decimals in fifth grade, and (c) divide multi-digit whole numbers in sixth grade (NGACBP & CCSSO, 2010). Published internal consistency at each time point of the DCAS was .86 (spring of fourth grade), .89 (spring of fifth grade), and .88 (spring of sixth grade) (American Institutes of Research, 2012).

Each student in the state is given an "accountability score" that is determined by his/her performance on items that measure grade level content only (American Institutes for Research, 2012). Accountability scores range from 0-1300. Based on these scores, students are classified with scores of 1 (well below standards), 2 (below standards), 3 (meets standards), or 4 (advanced). For the ROC curve analyses in the present study,

students' scores were further classified as a binary outcome, which is a requirement for ROC analyses: 1 (below and well below the mathematics standard) and 0 (meets the standard or advanced). Assessments of educational screeners using ROC curve analyses typically use statewide achievement tests as the binary outcome measure (e.g., Cummings & Smolkowski, 2015; Gersten et al., 2011), since such tests are used in school decision making.

The mathematics achievement measure has high criterion validity. The measure is highly correlated with the Wide Range Achievement Test (WRAT; Wilkinson & Robertson, 2006) fourth edition in mathematics, a standardized measure of general mathematics achievement. Bivariate correlations between concurrent administrations of the mathematics achievement outcome and WRAT in the present sample range from .71 to .76.

Procedure

Table 2 summarizes the assessment timeline. Students were given the NAEP fraction concepts measure in fall of fourth grade (Fall 2011), fall of fifth grade (Fall 2012), and winter of sixth grade (Winter 2014). The fraction number line estimation measure was administered in the winter of fourth grade (Winter 2012), the fall of fifth grade (Fall 2012) and the winter of sixth grade (Winter 2014). The fraction arithmetic measure was administered in fall of fourth grade (Fall 2011), fall of fifth grade (Fall 2012), and winter of sixth grade (Winter 2014). The DCAS mathematics achievement outcome measure was administered in the spring of each grade: fourth grade (2012), fifth grade (2013), and sixth grade (2014).

Trained assessors on the research team administered all measures, except for the mathematics achievement outcome measure, which was given by the school districts.

Table 2 Timeline of Predictor Measures and Outcome Measure

Measure	4th F	4th W	4th S	5th F	5th S	6th W	6th S
Predictor Measures							
NAEP fraction concepts	X			X		X	
FNLE		X		X		X	
Fraction arithmetic	X			X		X	
Outcome Measure							
Mathematics achievement			X		X		X

Note. F = Fall, W = Winter, S = Spring.

Assessors administered the NAEP fraction concepts measure in a whole-class setting. All problems were read aloud to students. The fraction number line estimation task and the fraction arithmetic measure were administered individually.

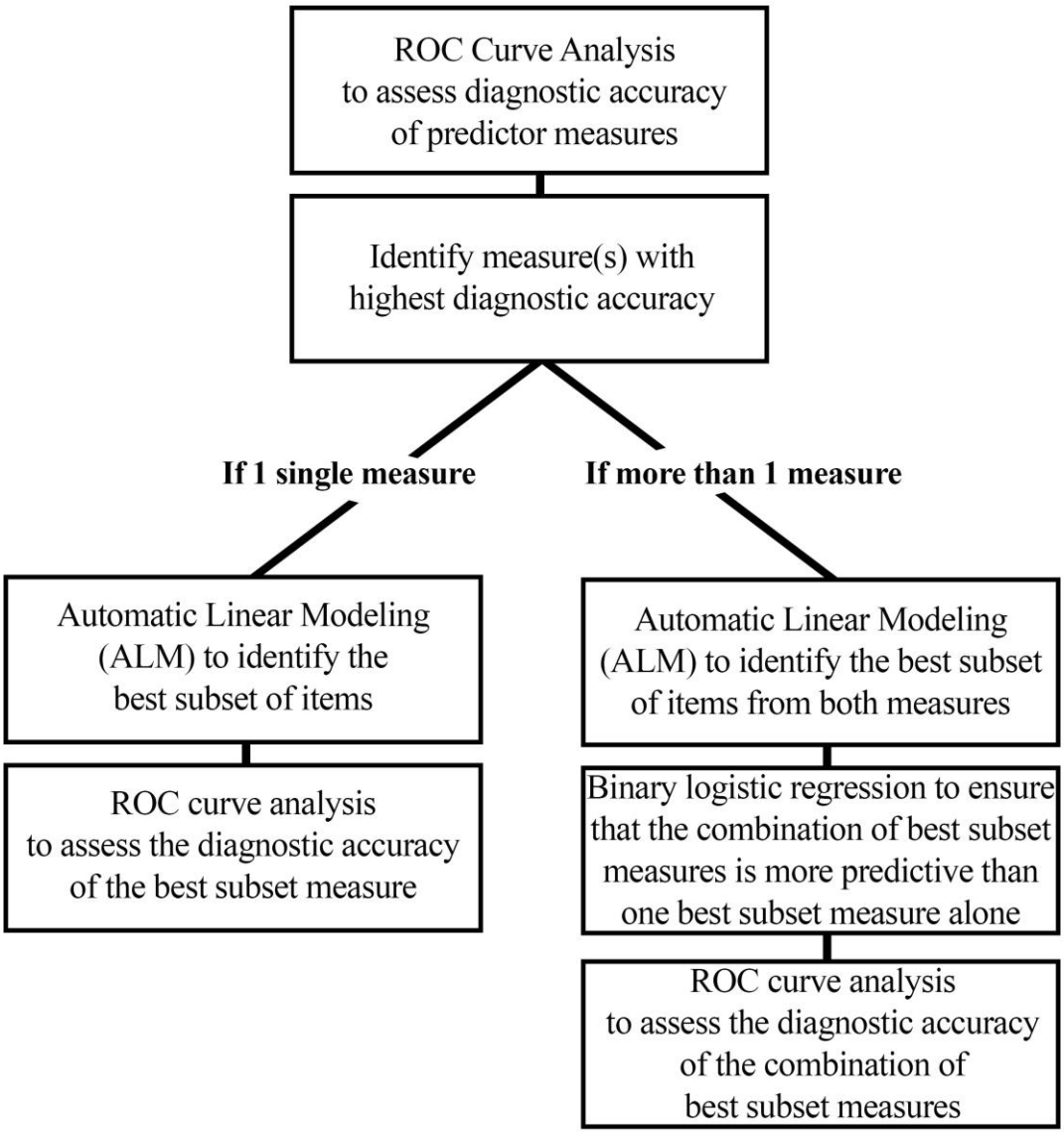
Data Analysis

The data analysis plan of the current study involved several steps. Figure 2 is a diagram that provides an overview of the main steps of the data analysis plan.

Selection of Predictor Measures for ROC Curve Analyses

First the researcher selects and describes the predictor measures and the outcome measure that will be entered in the ROC curve analysis. The researcher conducts unpaired *t*-tests to assess whether the two groups of interest (e.g., children who met and did not meet the mathematics standards) significantly differ in their performances on the predictor measure (Youngstrom, 2014). If a statistically significant difference is present, the predictor measure may be considered a potentially useful screener and appropriate for the ROC analysis. The researcher reports correlations for all screening measures across all time points, as well as means and standard deviations for each of these measures for students who met and did not meet the mathematics standards in each grade.

Figure 2 Data Analytic Plan of the Present Study



ROC Curve Analyses to Assess Potential Screeners

To assess the measures of the present study as potential screeners and to identify the measure(s) with high diagnostic accuracy, receiver operating characteristic (ROC) curve analyses were conducted using the statistical program SPSS Version 24.0 (IBM Corporation, 2016). Diagnostic accuracy, as applied to the present study, refers to a measure's ability to accurately predict student membership into one of two groups: students who are likely to meet the mathematics standard versus students who are not likely to meet the standard (See Appendix C for a glossary of key terms related to the ROC analyses of the present study, including "diagnostic accuracy"). Six different ROC curve analyses were conducted to assess the fraction measures administered in grades four through six, which are outlined in Table 3.

Overall, across the range of all possible cut scores, an accurate screener will yield high a high true positive fraction (i.e., more true positives and fewer false negatives) and a high true negative fraction (i.e., more true negatives and fewer false positives) (Smolkowski & Cummings, 2015). ROC curve plots are reported in the present study to allow for a visual interpretation of each measure's overall diagnostic accuracy (see Figure 3 for a sample ROC plot). If a measure accurately discriminates between students who meet and do not meet the end-of-the-year mathematics standard, its ROC curve will extend toward the upper left corner of the plot. That is, there are many cut scores along the measure that have *both* high rates of true positives and true negatives. As a ROC curve approaches the upper left corner, the area under the curve (AUC) increases. AUC is the most commonly used global index of diagnostic accuracy (Fluss, Faragii, & Reiser, 2005) and is easy to understand. For instance, if one student is randomly selected from the at-risk population and another student is randomly selected from the higher-achieving population, the AUC is the probability of distinguishing between those two students with the predictor measure (McFall & Treat, 1999). Thus, an AUC of .50 means that the measure correctly places students 50% of the time; this measure would not be considered

a powerful screener since it does not provide any discrimination between students who are at risk and students who are not at risk (Swets, Dawes, & Monahan, 2000).

Table 3 Six Receiver Operating Characteristic (ROC) Curve Analyses

Predictor Measures		Mathematics Achievement Outcome Measure
Fourth-grade fraction measures	→	Fourth grade spring
	→	Fifth grade spring
	→	Sixth grade spring
Fifth-grade fraction measures	→	Fifth grade spring
	→	Sixth grade spring
Sixth-grade fraction measures	→	Sixth grade spring

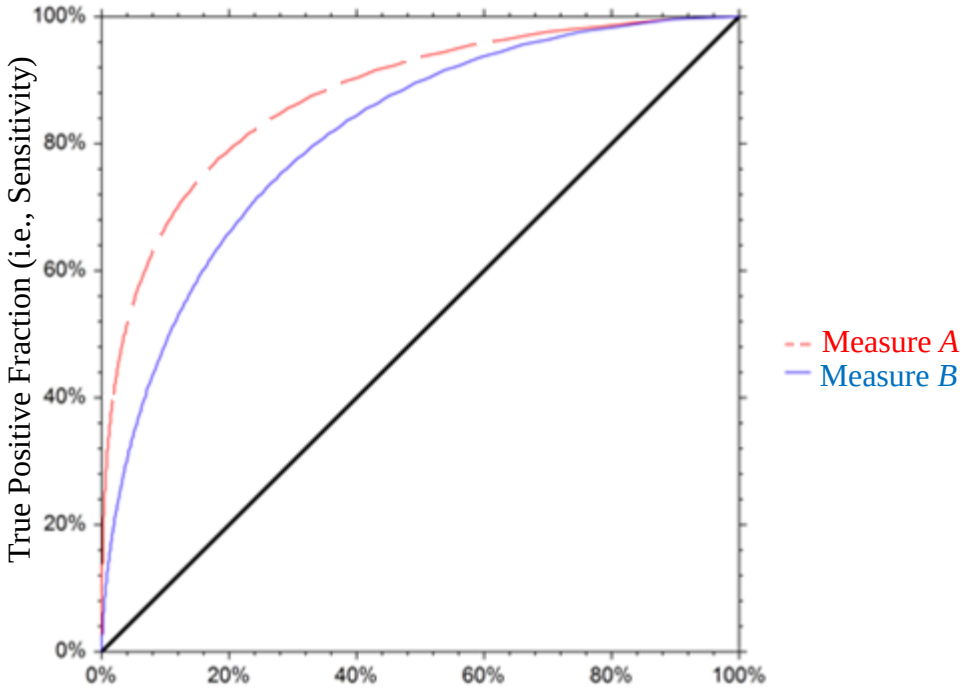
Overall, the ROC curve plot allows for a visual interpretation of the ROC curve results. The researcher can immediately see which measure holds the highest diagnostic accuracy by viewing the curve that extends closest to the upper left corner of the plot. However, additional steps are recommended below for further examination of the predictor measures.

Identify the Measure(s) with Highest Diagnostic Accuracy

Beyond looking at the ROC curve plot, a researcher must also report the actual AUC values and confidence intervals associated with each predictor measure. An AUC value of .750, for example, means that the measure correctly places students 75% of the time. The educational literature reports that AUC values ranging from .750 to .850 indicate good screeners for determining risk status, and AUC values ranging from .850 to .950 signify a very good screener (Cummings & Smolkowski, 2015). Whereas Cummings and Smolkowski report that an AUC value of .95 and above indicates an *excellent* screener, Youngstrom (2014) cautions about interpretations of such AUC values in the clinical literature; the researcher claims that AUC values greater than .90 are more likely to indicate design flaws of the predictor or outcome measure than exceptional

diagnostic accuracy. For example, such high AUC values may suggest that the predictor measure and the outcome measure are *too* correlated, such as using a state test of mathematics achievement in the fall of fifth grade to predict the state test in the spring of sixth grade. AUCs, along with corresponding 95% confidence intervals, are reported in the present study for all measures included in each ROC curve analysis. The present study uses the AUC threshold range of .750 to .850 as an indicator of a good screener measure; the threshold range of .850 to .950 is used as an indicator of a very good screener (Cummings & Smolkowski, 2015).

Figure 3 Sample ROC Curve Plot



For each ROC curve analysis, measures will be identified that hold an AUC value of .750 to .950. If only measure meets this requirement, this one measure will be recommended as the best screener measure.

If more than measure has an AUC value of .750 or higher, a method proposed by Hanley and McNeil (1983) is used for evaluating whether measures differ *significantly* from one another in terms of predictability. The method corrects for dependence in AUC values when both measures are assessed within the same sample. The method yields a critical ratio *z* value; when the value of $z = \pm 1.96$, the difference between the AUC values is statistically significant at $p < .05$. If Measure A, for example, significantly differs from Measure B, then Measure A will be suggested as the best screener measure. If Measure A does not significantly differ from Measure B, then a *combination* of the measures will be assessed.

Overall, there are two possible scenarios for each ROC curve analysis: (a) one single measure will emerge as a superior screener or (b) a combination of measures will be recommended as the best predictive screener. Subsequent sections of the data analytic procedure address different steps to follow for each scenario.

Automatic Linear Modeling to Improve the Screener

When a single measure or a combination of measures emerges as holding high diagnostic accuracy, automatic linear modeling (ALM) is used in the present study to improve the efficiency of the measure(s). ALM determines the best set of items for predicting students' later mathematics achievement. The analysis is run with SPSS Version 24.0 (IBM SPSS, 2016) and is an application of multiple linear regression modeling that identifies which predictor items are most influential in predicting a target outcome variable (Yang, 2013). Although the analysis has been available on SPSS since only 2010, it has been utilized in a variety of fields, including medical research (e.g., Ban et al., 2014) and marketing (e.g., Kadam & Nimbalkar, 2015). Whereas the nature of a ROC curve analysis necessitates a binary outcome (e.g., does not meet the mathematics standard vs. meets the mathematics standard), ALM allows for a continuous outcome (e.g., students' accountability scores on the DCAS).

The ALM selection method used in the present study was Best Subsets, which investigates all possible models for a given set of predictor items and determines the best set for predicting students' later mathematics achievement (Yang, 2013). The best model was identified using an Information Criterion; specifically, the Akaike Information Criterion (AIC), with lower AIC indicating better model fit (Meyers, Gamst, & Guarino, 2013). The model's adjusted R^2 was also computed, which is a measure of the proportion of variation in the target variable that is accounted for by the set of predictor items in the model. Thus, an R^2 of .25 would indicate that 25% of the variance in mathematics achievement (i.e., the target outcome variable) is accounted for by the linear combination of the predictor items. The adjusted R^2 includes a correction for the number of predictors in the model. The analysis pinpoints a combination of items that results in the best model fit; items that matter most in making the prediction are included in the model and predictors that matter least are excluded. For example, challenging items that most students answered incorrectly (e.g., an item that is beyond students' grade level) hold little variability and thus would not make strong predictions of students' later performance; this type of item is excluded from the screener. Likewise, easy items that all students answered correctly are eliminated, because the items do not discriminate well between students who are at risk for later difficulties and students who are not at risk. Retained items hold predictive power because some students gave the correct responses for the items while others missed the items (Meyers et al., 2013). The analysis provides a value of each item's importance that represents the sum of squares for the residual with the predictor removed from the model. An item deemed as important to be in the model indicates that leaving it out of the model would produce a substantial increase in the residual sum of squares (Meyers et al., 2013). Higher importance values indicate higher importance in the model. The importance values are relative; that is, the sum of the values for all predictors in a model is 1.0. Overall, the Best Subset ALM analysis

determines the best combination of screening items for making a prediction of students' later mathematics achievement.

When using Best Subset ALM for a combination of measures, all items from all measures are tested in the model. Using the model indices described above, the analysis determines the best combination of items across measures for predicting later achievement. For example, 11 items from Measure A combined with five items from Measure B may emerge as the best combination of items for predicting the target outcome variable.

Binary Logistic Regression to Justify the Combination of Screeners

Binary logistic regression was conducted to provide extra statistical support for combining best subset measures (e.g., Measure A and Measure B). Hierarchical block entry was used to empirically assess whether the addition of best subset Measure B statistically significantly improved prediction over best subset Measure A alone (e.g., Wilson et al., 2016). Order of entry was determined by the AUC values associated with the original ROC curve analysis; for example, if the original Measure A held a higher AUC value than the original Measure B, then Measure A would be entered in the model first.

Additional ROC Curve Analyses to Test the Best Subset Screeners

After determining the best subset of screening items, the present study conducted additional ROC curve analyses to compare the best subset measure with the original measure(s) with all items included. This final step was performed to ensure that the best subset screener performed better or equally as well as the original measure(s) for predicting students' later achievement. If the best subset measure was a *combined* best subset measure, then the ROC curve analysis also compared this measure to each

individual best subset measure. As described previously, AUC values were compared to assess the diagnostic accuracy of all measures.

For a single best subset measure, this final ROC curve analysis yields the necessary ROC statistics associated with certain cut scores for applying the data to other real-world settings. These statistics include the true positive fraction (i.e., sensitivity), the true negative fraction (i.e., specificity), and the positive predictive power (PPV). As recommended in the clinical literature, the present study also reports the diagnostic likelihood ratio for a positive screener result and a negative screener result (i.e., DLR+ and DLR-). The diagnostic likelihood ratios allow a researcher or practitioner to make empirically-driven predictions of his/her own students' mathematics achievement.

For a combined best subset measure, the researcher must address additional considerations. If combining screeners with discrete data (e.g., students receive one point per correct response), the researcher can compute a new variable that is a total score for all items combined. By doing so, the researcher no longer considers the measures as separate screeners but as one single measure. The new measure can then be entered in a ROC curve analysis, and cut scores and all ROC curve statistics can be reported (e.g., true positive fraction and the positive predictive power). If combining one screener with discrete data and a second screener with continuous data (e.g., measuring students' estimates on the fraction number line measure with percent absolute error), the researcher cannot as easily combine the two measures into one screener. Rather, the researcher must now consider a *combination* of scores rather than simply one score alone to make predictions of student performance. For this scenario, the researcher can use logistic regression to assess the combination of measures in a ROC curve analysis (e.g., Wilson et al., 2016). The regression analysis produces predicted probabilities for all students based on a linear combination of the measures (Clemens et al., 2016). The predicted probabilities can then be entered as a "measure" within the ROC curve analysis, allowing an AUC value to be computed for the combined best subset measure. Although the ROC

analysis provides the AUC of the combined measure, it does not provide helpful statistics associated with cut scores for the combined measure. Instead, the researcher can create a predicted probabilities matrix that allows a researcher or practitioner to make a prediction of a student's later performance based on his/her score on Measure A *along with* his/her score on Measure B.

Power Analysis and Consideration of Missing Data

A power analysis was performed using MedCalc Statistical Software version 16.4.3 (MedCalc Statistical Software, 2016) for calculating the required sample size for an AUC value to be significantly different from the null hypothesis. The null hypothesis for a ROC curve analysis is an AUC of .50, which signifies that a screener has zero discriminating power. The power analysis was conducted for an anticipated AUC value of .750, since this value is the AUC threshold recommended in the educational literature to assess measures as good screeners (Cummings & Smolkowski, 2015). Thus, the analysis was conducted with the following information: AUC of .75, null hypothesis of .50, power of .80, and alpha of .05. The analysis also requires a ratio representative of the expected amount of negative cases in the sample (students who meet the mathematics standard) as compared to the amount of positive cases (students who do not meet the standard). A ratio value of 2.5 was selected for the present study because the amount of negative cases across the time points of the DCAS outcome measure was approximately 2.5 times more prevalent than the amount of positive cases in the sample. Results from the power analysis revealed that a sample size of 49 students (with approximately 14 positive cases and 35 negative cases) would be sensitive to differences between an AUC of .75 and the null hypothesis.

Additional power analyses were conducted to determine the required sample size for the comparison of two ROC curves from the same sample; specifically, analyses were conducted to assess the sample size required to detect a significant difference between an

AUC value of .75 and a second AUC of .85. The first analysis was conducted with the following information: AUC of .75 for one ROC curve, AUC of .85 for the second ROC curve, power of .80, alpha of .05, and a ratio value of 2. The total sample size required to compare the two ROC curves is 285 students. The second analysis was conducted for the same AUC values but with a ratio value of 3; the total sample size required is 356 students. Thus, in order to compare these two AUC values that differ by no less than .10 (e.g., AUC = .75, 85; respectively), the required sample size is between 285 - 356 students.

In the present study, six initial ROC curve analyses were conducted (see Table 3) using one large longitudinal dataset from grades four through six. As is common for longitudinal datasets, missing data is observed and must be considered (Martinez-Camblor, 2013). The most frequently reported remedy for missing data in ROC curve analyses is to assess AUCs only from subjects who have compete information, called the available-case analysis (Martinez-Camblor, 2013). In the present study, the available-case analysis yields a total of over 300 students per each ROC curve, which is much greater than the sample size determined by the first power analysis (N = 72). The available cases in the present study also allow for AUC comparisons. Since the available-case analysis is the most commonly used treatment of missing data in ROC research and is expected to yield sample sizes that far exceed or meet the sample size determined by the power analyses, this treatment of missing data is used in the present study.

Chapter 4

RESULTS

Correlations among all variables across grades are shown in Table 4, with the DCAS mathematics achievement outcome entered as a binary variable to align with the ROC curve analyses. All variables are significantly correlated.

Fourth-Grade Measures Predicting Later Mathematics Achievement in Fourth, Fifth, and Sixth Grades

Table 5 presents means and standard deviations for all fourth-grade predictor measures, separated for students who met the end-of-the-year mathematics standard and students who did not meet the standard in fourth, fifth, and sixth grades, respectively. Independent samples t –tests revealed that each predictor differentiated students who met the standard from those who did not meet the standard, regardless of the grade of the outcome measure (p = .001). Thus, all fourth-grade predictor measures qualified for use in ROC curve analyses.

ROC Curve Analyses with Fourth-Grade Predictor Measures

Three ROC curve analyses were conducted to assess the diagnostic accuracy of the fourth-grade predictor measures on later mathematics achievement. The first analysis assessed the measures as potential screeners for predicting the fourth-grade outcome. The base rate of the first ROC curve analysis was .21, meaning that 21% of the students received a positive outcome result. In other words, 21% of the sample did not meet the mathematics standard. The second ROC curve analysis assessed the ability of the same measures for predicting the fifth-grade outcome (base rate = .21). The third ROC curve analysis assessed the same measures for predicting the sixth-grade outcome

Table 4 Correlations Among all Predictor and Outcome Variables

	1	2	3	4	5	6	7	8	9	10	11	12
1. FNLE–4th	--											
2. FNLE–5th	.674	--										
3. FNLE–6th	.575	.697	--									
4. NAEP Concepts–4th	-.631	-.602	-.527	--								
5. NAEP Concepts–5th	-.597	-.695	-.633	.675	--							
6. NAEP Concepts–6th	-.566	-.679	-.735	.571	.723	--						
7. Fraction Arithmetic–4th	-.450	-.482	-.392	.521	.500	.423	--					
8. Fraction Arithmetic–5th	-.428	-.502	-.492	.507	.621	.559	.393	--				
9. Fraction Arithmetic–6th	-.498	-.494	-.491	.470	.538	.575	.392	.460	--			
10. DCAS Outcome–4th	.353	.390	.487	-.405	-.516	-.482	-.294	-.395	-.354	--		
11. DCAS Outcome–5th	.369	.452	.543	-.424	-.557	-.608	-.323	-.424	-.428	.596	--	
12. DCAS Outcome–6th	.421	.524	.613	-.457	-.627	-.662	-.283	-.460	-.491	.608	.644	--

Note. All correlations are significant at the .01 level. Fraction Number Line Estimation (FNLE) is measured in percent absolute error; higher scores indicate poorer performance.

Table 5 Mean Differences for Fourth-Grade Predictor Measures between Students Who Did and Did Not Meet the Mathematics Standard in Fourth, Fifth, and Sixth Grade

	Met the Math Standard 4th Grade: (n = 326) 5th Grade: (n = 264) 6th Grade: (n = 203)	Did Not Meet the Math Standard 4th Grade: (n = 85) 5th Grade: (n = 98) 6th Grade: (n = 101)		
4th-Grade Predictor Measure	M (SD)	M (SD)	*t(df)*	*p*
Predicting 4th-Grade Outcome				
FNLE	22.82 (8.49)	30.26 (5.28)	-10.04(211)	.001
NAEP Fraction Concepts	10.70 (3.42)	7.15 (2.62)	10.39(167)	.001
Fraction Arithmetic	3.34 (3.37)	0.95 (2.01)	8.37(224)	.001
Predicting 5th-Grade Outcome				
FNLE	22.36 (8.67)	29.41 (5.37)	-9.26(279)	.001
NAEP Fraction Concepts	11.02 (3.38)	7.66 (2.64)	9.92(220)	.001
Fraction Arithmetic	3.50 (3.39)	1.11 (2.27)	7.70(259)	.001
Predicting 6th-Grade Outcome				
FNLE	21.92 (8.64)	29.49 (5.36)	-9.38(288)	.001
NAEP Fraction Concepts	10.97 (3.33)	7.53 (2.82)	9.42(232)	.001
Fraction Arithmetic	3.41 (3.40)	1.42 (2.55)	5.72(256)	.001

Note. FNLE = Fraction Number Line Estimation. FNLE is measured in percent absolute error (PAE), meaning higher scores indicate poorer performance. List-wise deletion was utilized to correspond with the cases included in each ROC curve analysis per each grade.

(base rate = .33). Base rates for each analysis are presented in Table 6, along with the total count of positive outcome results and negative outcome results per grade.

Table 6 Positive Outcomes, Negative Outcomes, and Base Rates for ROC Analyses with Fourth-Grade Measures predicting the Mathematics Achievement Outcome in Fourth, Fifth, and Sixth Grades

Grade of Outcome	Positive	Negative	Base Rate
4th	85	326	21%
5th	98	264	27%
6th	101	203	33%

ROC curve plots for all three analyses provide a visual interpretation of the ROC curve data (Figure 4). On all plots, the NAEP fraction concepts curve extended furthest to the top left corner. As such, the fourth-grade NAEP fraction concepts measure held the highest area under the curve and thus the highest diagnostic accuracy for predicting mathematics achievement in all three grades.

ROC curve statistics for each analysis are presented in Table 7. The AUC values associated with the NAEP fraction concepts measure in each ROC analysis exceed .750, indicating that the measures met the minimum acceptable value to be effective for determining risk status (Cummings & Smolkowski, 2015).

In all three ROC curve analyses, the fourth-grade fraction number line estimation measure also emerged as a powerful screener for predicting the outcome. The fraction number line estimation measure met the AUC threshold of .750 for predicting the fourth-grade outcome and for predicting the sixth-grade outcome. The AUC approached the threshold of .750 for predicting the fifth-grade outcome (AUC = .745). However, the

Figure 4 ROC Curve Plots Showing Fourth-Grade Predictor Measures

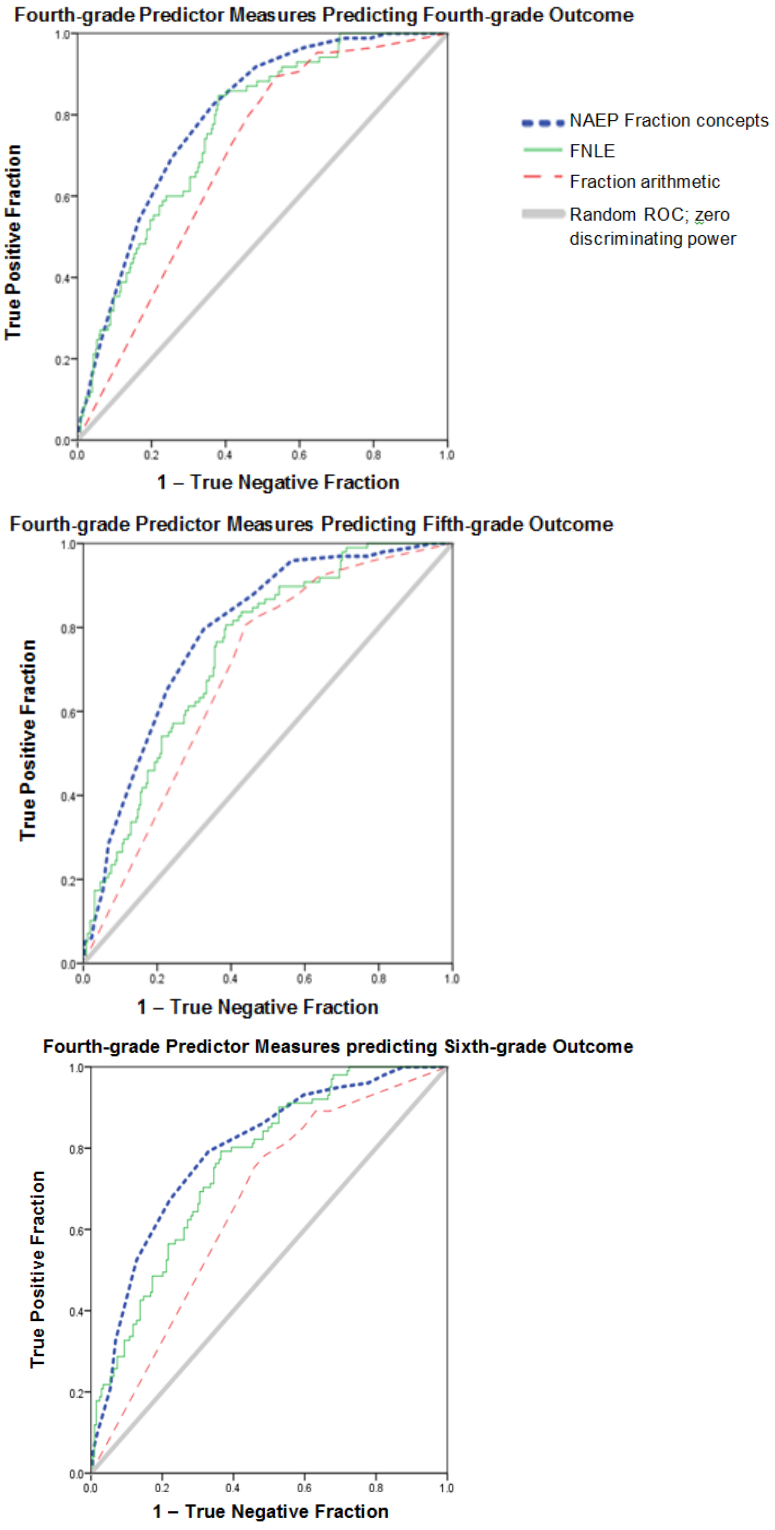


Table 7 ROC Area Under the Curve (AUC) Statistics for Fourth-Grade Predictor Measures Predicting the Mathematics Achievement Outcome in Fourth, Fifth, and Sixth Grades

			95% Confidence Interval	
4th-Grade Predictor Measure	AUC	SE	Lower	Upper
Predicting 4th-Grade Outcome				
NAEP Fraction Concepts	.796[1]	.024	.749	.843
FNLE	.766[1]	.026	.715	.817
Fraction Arithmetic	.693	.028	.637	.749
Predicting 5th-Grade Outcome				
NAEP Fraction Concepts	.789[1]	.025	.740	.838
FNLE	.745	.027	.693	0798
Fraction Arithmetic	.692	.029	.636	.749
Predicting 6th-Grade Outcome				
NAEP Fraction Concepts	.791[1]	.027	.738	.844
FNLE	.760[1]	.028	.706	.814
Fraction Arithmetic	.659	.032	.596	.722

[1]AUC > .750, indicating that the measure meets the minimum acceptable value to be effective for determining risk status (Cummings & Smolkowski, 2015).

AUC value for the fraction number line estimation measure in each analysis does not exceed those associated with the NAEP fraction concepts measure.

The fraction arithmetic measure yielded the lowest AUC values in each analysis (AUC = 693, .692, and .659, respectively). The AUC values do not meet the .750 AUC threshold for effective screeners. As such, the fourth-grade fraction arithmetic measure is not a powerful screener measure for predicting the mathematics achievement outcome measure given fourth, fifth and sixth grades.

Fifth-Grade Measures Predicting Later Mathematics Achievement in Fifth and Sixth Grades

Table 8 presents means and standard deviations for all fifth-grade predictor measures, separated for students who met the end-of-the-year mathematics standard and

students who did not meet the standard in fifth and sixth grades. Independent samples *t* – tests revealed that each predictor differentiated students who met the standard from those who did not meet the standard in both fifth and sixth grade ($p = .001$). All fifth-grade predictor measures thus qualified for the ROC curve analyses.

ROC Curve Analyses with Fifth-Grade Predictor Measures

Two separate ROC curve analyses were run to assess the diagnostic accuracy of the fifth-grade predictor measures. Table 9 presents the base rate for each analysis. The first analysis assessed the accuracy of the fifth-grade measures for predicting the fifth-grade outcome. The second analysis assessed the ability of the same fifth-grade measures for predicting the sixth-grade outcome. The ROC curve plot for each analysis shows the NAEP fraction concepts curve extending furthest to the top left corner of the plot, meaning that the NAEP fraction concepts measure held the highest diagnostic accuracy for predicting the DCAS mathematics achievement outcome in both grades (Figure 5).

ROC curve statistics for each analysis are presented in Table 10. The AUC values associated with the NAEP fraction concepts measure in each ROC analysis exceeded .750, the minimum acceptable AUC value to be effective for determining risk status Cummings & Smolkowski, 2015). Although the NAEP fraction concepts measure held the highest AUC for both administrations of the outcome, the other predictor measures also held high diagnostic accuracy. The AUC of the fraction number line estimation measure exceeded .750 in each analysis. The AUC of the fraction arithmetic measure exceeded the threshold in each analysis but was considerably lower than those associated with the other predictor measures. Overall, all fifth-grade measures met the minimum AUC threshold for determining risk status at the end of fifth grade and the end of sixth grade, with the NAEP fraction concepts measure still emerging as the most accurate screener.

Table 8 Mean Differences for Fifth-Grade Predictor Measures between Students Who Did and Did Not Meet the Mathematics Standard in Fifth and Sixth Grades

5th-Grade Predictor Measure	Met the Standard 5th Grade: (n = 279) 6th Grade: (n = 211) M (SD)	Did not Meet the Standard 5th Grade: (n = 105) 6th Grade: (n = 107) M (SD)	$t(df)$	p
Predicting 5th-Grade Outcome				
FNLE	16.15 (10.30)	26.74 (6.08)	-12.37(313)	.001
NAEP Concepts	14.51 (3.09)	9.98 (2.81)	13.11(382)	.001
Fraction Arithmetic	6.75 (2.95)	3.47 (3.60)	8.35(159)	.001
Predicting 6th-Grade Outcome				
FNLE	15.33 (9.92)	26.87 (6.34)	-12.57(300)	.001
NAEP Concepts	14.82 (2.85)	9.99 (2.82)	14.31(316)	.001
Fraction Arithmetic	7.01 (2.73)	3.70 (3.62)	8.41(169)	.001

Note. List-wise deletion was utilized to correspond with the cases included in each ROC curve analysis per each grade.

Table 9 Positive Outcomes, Negative Outcomes, and Base Rates for ROC Analyses with Fifth-Grade Measures predicting the Mathematics Achievement Outcome in Fifth and Sixth Grades

Grade of Outcome	Positive	Negative	Base Rate
5th	105	279	27%
6th	107	211	34%

Figure 5 ROC Curve Plots Showing Fifth-Grade Predictor Measures

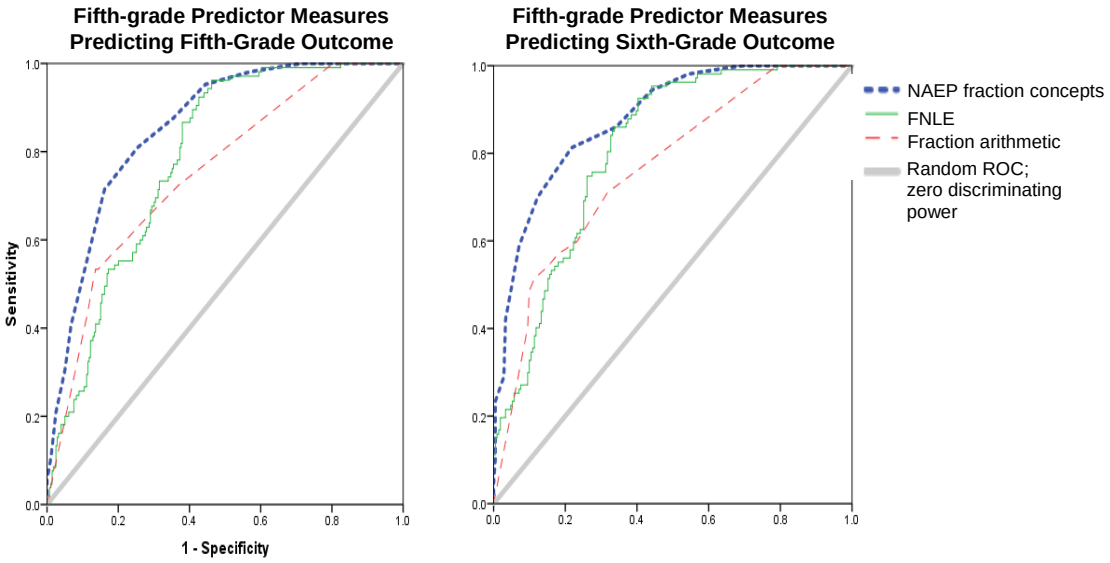


Table 10 ROC Area Under the Curve (AUC) Statistics for Fifth-Grade Predictor Measures Predicting the Mathematics Achievement Outcome in Fifth and Sixth Grades

			95% Confidence Interval	
5th-Grade Predictor Measure	AUC	SE	Lower	Upper
Predicting 5th-Grade Outcome				
NAEP Fraction Concepts	.856[1]	.019	.818	894
FNLE	.784[1]	.023	.738	.829
Fraction Arithmetic	.753[1]	.027	.699	.806
Predicting 6th-Grade Outcome				
NAEP Fraction Concepts	.880[1]	.019	.842	.917
FNLE	.810[1]	.024	.764	.856
Fraction Arithmetic	.764[1]	.028	.710	.818

[1]AUC > .750, indicating that the measure meets the minimum acceptable value to be effective for determining risk status (Cummings & Smolkowski, 2015).

Sixth-Grade Measures Predicting Later Mathematics Achievement in Sixth Grade

Table 11 presents means and standard deviations for all sixth-grade predictor measures, separated for students who met the end-of-the-year mathematics standard and students who did not meet the standard at the end of sixth grade. An independent samples *t* –test revealed that each predictor differentiated students who met the standard from those who did not meet the standard in sixth grade ($p = .001$). All three of the sixth-grade predictor measures thus qualified for inclusion in the ROC curve analysis.

ROC Curve Analysis with Sixth-Grade Predictor Measures

One ROC curve analysis assessed the diagnostic accuracy of the sixth-grade predictor measures for predicting the end-of-the-year sixth-grade outcome. The base rate was .33, meaning that 33% of the sample did not meet the sixth-grade mathematics standard (Table 12). Similar to the other analyses, the NAEP fraction concepts curve yet

Table 11 Mean Differences for Sixth-Grade Predictor Measures between Students Who Did and Did Not Meet the Mathematics Standard in Sixth Grade

	Met the Math Standard (n = 220)	Did Not Meet the Math Standard (n = 107)		
6th-Grade Predictor Measure	M (SD)	M (SD)	*t(df)*	*p*
Predicting 6th-Grade Outcome				
FNLE	8.92 (7.08)	21.26 (8.62)	-12.84(178)	.001
NAEP Fraction Concepts	20.34 (2.98)	14.23 (3.88)	14.36(169)	.001
Fraction Arithmetic	13.41 (4.65)	8.30 (3.52)	11.05(269)	.001

Note. List-wise deletion was utilized to correspond with the cases included in each ROC curve analysis per each grade.

Table 12 Positive Outcomes, Negative Outcomes, and Base Rate for ROC Analysis with Sixth-Grade Measures Predicting the Mathematics Achievement Outcome in Sixth Grade

Grade of Outcome	Positive	Negative	Base Rate
6th	107	220	33%

again extended furthest to the top left corner of the ROC curve plot. The plot indicates that the sixth-grade NAEP fraction concepts measure held the highest diagnostic accuracy for predicting the outcome at the end of sixth grade (Figure 6).

ROC curve statistics for the sixth-grade analysis are presented in Table 13. The AUC value for the NAEP fraction concepts measure exceeded .750 and was the highest AUC across all ROC analyses in the present study (AUC = .895). The diagnostic accuracy of the fraction number line estimation measure (AUC = .864) and fraction arithmetic measure (AUC = .817) also exceeded the .750 AUC threshold. Overall, all sixth-grade measures met the minimum AUC threshold for determining risk status at the end of sixth grade, with the NAEP fraction concepts measure demonstrating the highest diagnostic accuracy.

Figure 6 ROC Curve Plot showing Sixth-Grade Predictor Measures Predicting the Sixth-Grade Mathematics Achievement Outcome

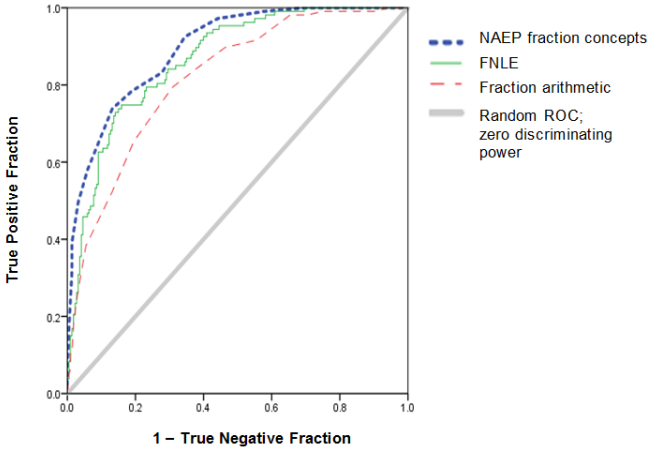


Table 13 ROC Curve Area Under the Curve (AUC) Statistics for Sixth-Grade Predictor Measures Predicting Sixth-Grade Mathematics Achievement Outcome

6th-Grade Predictor Measure	AUC	SE	95% Confidence Interval	
			Lower	Upper
Predicting 6th-Grade				
NAEP Fraction Concepts	.895[1]	.017	.861	.929
FNLE	.864[1]	.020	.826	.905
Fraction Arithmetic	.817[1]	.024	.770	.864

[1]AUC > .750, indicating that the measure meets the minimum acceptable value to be effective for determining risk status (Cummings & Smolkowski, 2015).

Comparing AUC Values

Table 14 shows a summary of AUC values across all six ROC curve analyses. Table 15 shows *p* values of the three AUC comparisons analyzed per ROC curve analysis (i.e., NAEP fraction concepts vs. FNLE, FNLE vs. fraction arithmetic, and NAEP fraction concepts vs. fraction arithmetic). For the present study, it was of particular interest to assess if the measure with the highest AUC value in each analysis (i.e., NAEP fraction concepts) was significantly superior to the measure with the next highest AUC value (i.e., fraction number line estimation).

Comparison of Fourth-Grade Predictor Measures

The NAEP fraction concepts measure did not significantly outperform the fraction number line measure as a screener for any year of the outcome measure ($p > .05$; see Table 15). The NAEP fraction concepts measure performed significantly better than the fraction arithmetic measure in each analysis ($p < .05$). The fraction number line measure outperformed the fraction arithmetic measure at fourth and sixth grade ($p < .05$). Overall, the results suggest that a combination of the NAEP fraction concepts items and fraction number line estimation items would yield an improved screener measure.

Table 14 Summary of Area under the Curve (AUC) Statistics and 95% Confidence Intervals (CI) for Predictors Administered in Fourth, Fifth, and Sixth Grades

Predictor	AUC[1] [95% CI] for Prediction of the Outcome		
	4th-Grade Outcome	5th-Grade Outcome	6th-Grade Outcome
4th grade			
NAEP Fraction Concepts (18 items)	.796[1] [.749, .843]	.789[1] [.740, .838]	.791[1] [.738, .844]
FNLE	.766[1] [.715, .817]	.745 [.693, .798]	.760[1] [.706, .814]
Fraction Arithmetic (8 items; +,-)	.693 [.637, .749]	.692 [.636, .749]	.659 [.596, .722]
5th grade			
NAEP Fraction Concepts (19 items)		.856[1] [.818, .894]	.880[1] [.842, .917]
FNLE		.784[1] [.738, .829]	.810[1] [.764, .856]
Fraction Arithmetic (10 items; +,-)		.753[1] [.699, .806]	.764[1] [.710, .818]
6th grade			
NAEP Fraction Concepts (24 items)			.895[1] [.861, .929]
FNLE			.864[1] [.826, .905]
Fraction Arithmetic (26 items; +,-,x,÷)			.817[1] [.770, .864]

[1]AUC > .750

Table 15 Comparing the Diagnostic Accuracy of Predictor Measures

Predictor Measure Comparisons	Statistical Significance Level (p value) of AUC Comparisons		
	4th-Grade Outcome	5th-Grade Outcome	6th-Grade Outcome
4th grade			
NAEP Fraction Concepts vs. FNLE	.290	.129	.334
NAEP Concepts vs. Fraction Arithmetic	.001*	.003*	.001*
FNLE vs. Fraction Arithmetic	.033*	.130	.007*
5th grade			
NAEP Fraction Concepts vs. FNLE	---	.002*	.005*
NAEP Concepts vs. Fraction Arithmetic	---	.001*	.001*
FNLE vs. Fraction Arithmetic	---	.304	.161
6th grade			
NAEP Fraction Concepts vs. FNLE	---	---	.129
NAEP Concepts vs. Fraction Arithmetic	---	---	.002*
FNLE vs. Fraction Arithmetic	---	---	.086

*$p < .05$

Comparison of Fifth-Grade Predictor Measures

The AUC of the fifth-grade NAEP fraction concepts measure was significantly better than both fraction number line estimation ($p < .05$ for both grades of the outcome measure) and fraction arithmetic ($p < .05$ for both grades of the outcome measure). The AUC values of fraction number line estimation and fraction arithmetic were not significantly different ($p > .05$). The results indicate that the fifth-grade NAEP fraction concepts measure alone is the most predictive screener of later mathematics performance.

Comparison of Sixth-Grade Predictor Measures

The AUC of the NAEP fraction concepts measure did not significantly differ from the AUC of the fraction number line estimation measure ($p > .05$). However, the NAEP fraction concepts measure performed significantly better than fraction arithmetic ($p < .05$). Fraction number line estimation did not differ significantly from the fraction arithmetic measure ($p > .05$). The results are similar to the fourth-grade predictor results

and suggest that a combination of NAEP fraction concepts items and fraction number line estimation items would yield an improved screener measure.

Summary of AUC Comparisons

Results of the AUC comparisons differed by grade. For the fourth-grade and sixth-grade predictor measures, the NAEP fraction concepts measure performed equally well as the fraction number line estimation measure in all analyses. A combination of the two measures in fourth grade and sixth grade will hence be assessed. For the fifth-grade predictor measures, the NAEP fraction concepts measure significantly outperformed the fraction number line estimation in both analyses. As such, the NAEP fraction concepts measure alone will be assessed as the best-performing fifth-grade screener.

Determining Best Subset Measures

Fourth-Grade Best Subset Measure

ROC curve results indicated that both the fourth-grade NAEP fraction concepts measure and the fourth-grade fraction number line estimation measure were strong screeners of students' later mathematics performance at the end of fourth grade, fifth grade, and sixth grade. As such, items from both measures were assessed with best subset automatic linear modeling (ALM). The ALM analyses determined the most predictive combination of items from the 18-item NAEP fraction concepts measure and the 28-item fraction number line estimation measure. Three separate ALM analyses were run for the three grades of the mathematics achievement outcome variable: fourth, fifth, and sixth grade. ALM analyses allow for a continuous outcome measure; accountability scores on the DCAS mathematics achievement test served as the outcome.

For the prediction of the fourth-grade outcome, the final model had an adjusted R^2 of .59. Thus, 59% of the variance in the fourth-grade mathematics achievement outcome was accounted for by the linear combination of the selected predictor items. Six NAEP

items and 10 fraction number line items were included in the model. Table 16 shows the specific NAEP items included in the final model, and Table 17 shows the fraction number line items included in the model. The tables report the importance values of each item included in the final model, with higher importance values indicating higher importance in the model and hence higher predictive power for predicting the mathematics achievement outcome (Meyers et al., 2013). The item with the highest importance value per time point is bolded. Items with low importance values that were excluded from the final model have blank cells. Items that were not included in any of the six models were excluded altogether from the tables (i.e., NAEP items 16, 20, 21, 22; FNLE 0-1 items 1/5, 13/14, 5/8, 1/3; and FNLE 0-2 items 111/12, 5/6, 5/5, 1 5/8, and 1 1/5).

The two best subset measures (i.e., the six-item NAEP fraction concepts best subset measure and the 10-item fraction number line best subset measure) were entered into binary logistic regression for two reasons: (a) to empirically assess whether both measures made significant improvements to the model when predicting the fourth-grade mathematics achievement outcome and (b) to provide extra support for combining the two best subset measures. Hierarchical entry was used with the NAEP fraction concepts best subset measure entered in the first block and the fraction number line best subset measure entered in the second block; the NAEP measure was entered first because the original measure had a slightly higher AUC value than the fraction number line estimation measure. Regression diagnostics were performed to evaluate whether the model met underlying assumptions (Meyers, Gamst, & Guarino, 2006). The analyses revealed no univariate or multivariate outliers. A further evaluation of assumptions was satisfactory for the absence of influential cases, multicollinearity, and violations regarding the expected frequencies per cell for a logistic regression analysis. The Hosmer-Lemeshow goodness-of-fit test showed good model fit with the data (p = .702). Table 18 presents regression coefficients (B), Wald statistics and significance levels for

each best subset predictor measure in the model. The Wald test revealed that both best subset measures were statistically significant (p = .001), providing further rationale for combining the two measures in subsequent analyses.

When predicting the fifth-grade mathematics outcome, the final model of fourth-grade predictor items accounted for 47% of the variance in the outcome. The best subset model included five NAEP items and 11 fraction number line items (See Tables 16 and 17). The two best subset measures were entered into a binary logistic regression to empirically assess whether both measures made significant improvements to the model for the prediction of the fifth-grade mathematics achievement outcome. The model met the underlying assumptions of the regression analysis. The Hosmer-Lemeshow goodness-of-fit test showed good model fit with the data (p = .654). Table 19 presents regression coefficients. Both best subset measures were statistically significant (p = .001).

When predicting the sixth-grade mathematics outcome, the best subset final model accounted for 59% of the outcome variance. The model included six NAEP items and 10 fraction number line items (See Tables 16 and 17). Binary logistic regression with hierarchical entry revealed that both subset models significantly improved the model prediction of the sixth-grade outcome. Again the model met underlying assumptions and showed good model fit with the data as indicated by the Hosmer-Lemeshow goodness-of-fit test (p = .269). Table 20 presents the regression coefficients.

Table 16 NAEP Fraction Concepts Items included on Best Subset Measures per Time Point and Associated Importance Values

	4th-Grade Screeners			5th-Grade Screeners		6th-Grade Screener
Predictor Item	4th-Grade	5th-Grade	6th-Grade	5th-Grade	6th-Grade	6th-Grade
NAEP 1		.037				
NAEP 2		.044			.032	
NAEP 3	.031	.100		**.321**	.050	
NAEP 4				.039		.020
NAEP 5			.048	.042	.071	
NAEP 6	.034		.022			
NAEP 7	.095	.073	.056		.057	.056
NAEP 8	.054			.039	.127	
NAEP 9	.030					
NAEP 10				.040		
NAEP 11						.040
NAEP 12	.032		.033		.062	.014
NAEP 13			.068	.058		.076
NAEP 14			.082	.227	.144	
NAEP 15				.049	.076	.022
NAEP 17				.051	**.215**	.036
NAEP 18				.047	.115	.056
NAEP 19		.034		.090	.052	.035
NAEP 23						.221
NAEP 24						.137

Table 17 Fraction Number Line (FNLE) Items Included on Best Subset Measures per Time Point and Associated Importance Values

	4th-Grade Screeners			5th-Grade Screeners		6th-Grade Screener
Predictor Item	4th-Grade	5th-Grade	6th-Grade	5th-Grade	6th-Grade	6th-Grade
FNLE (0-1): 2/13	.022	.024	.029	x	x	
FNLE (0-1): 3/7				x	x	.035
FNLE (0-1): 1/3	**.222**	**.159**	.103	x	x	
FNLE (0-1): 1/19	.034	.034		x	x	
FNLE (0-1): 5/6	.133	.065	.048	x	x	
FNLE (0-2): 1/3			.030	x	x	.021
FNLE (0-2): 7/4	.036			x	x	**.230**
FNLE (0-2): 12/13		.025	.031	x	x	
FNLE (0-2): 3/2		.092		x	x	
FNLE (0-2): 1/2	.039		.047	x	x	
FNLE (0-2): 7/6		.109		x	x	
FNLE (0-2): 1 2/4		.037		x	x	
FNLE (0-2): 1	.051	.069	**.296**	x	x	
FNLE (0-2): 3/8	.083	.061	.041	x	x	
FNLE (0-2): 2/3				x	x	
FNLE (0-2): 7/9		.038		x	x	
FNLE (0-2): 1/19	.042			x	x	
FNLE (0-2): 1 5/6	.062		.031	x	x	
FNLE (0-2): 4/3			.034	x	x	

Table 18 Regression Coefficients of Fourth-Grade Best Subset Measures predicting the Fourth-Grade Mathematics Achievement Outcome

Predictor	*B*(SE)	Wald
Best Subset NAEP Fraction Concepts	-0.34(.09)	16.37
Best Subset FNLE	0.09(.02)	20.84
Constant	-2.57(.76)	11.36

Table 19 Regression Coefficients of Fourth-Grade Best Subset Measures Predicting the Fifth-Grade Mathematics Achievement Outcome

Predictor	*B*(SE)	Wald
Best Subset NAEP Fraction Concepts	-0.52(.12)	18.58
Best Subset FNLE	0.10(.02)	22.63
Constant	-2.45(.77)	10.26

Table 20 Regression Coefficients of Fourth-Grade Best Subset Measures Predicting the Sixth-Grade Mathematics Achievement Outcome

Predictor	*B*(SE)	Wald
Best Subset NAEP Fraction Concepts	-0.36(.10)	12.09
Best Subset FNLE	0.15(.03)	30.80
Constant	-3.92(.84)	21.85

Overall, the total items included in each best subset model differed by the grade of the outcome measure. One NAEP item (multiple-choice NAEP item 7; "Luis had two apples and he cut each apple into fifths. How many pieces of apple did he have?") and five fraction number line items (2/13, 1/3, and 5/6 on the 0-1 number line; 3/8 and the whole number 1 on the 0-2 number line) were included in all three models. Each fourth-grade combined best subset measure included a total of 16 items, which is fewer total items than the original 18-item NAEP and the original 28-item fraction number line estimation measure.

Fifth-Grade Best Subset Measure

ROC curve analyses assessing fifth-grade predictor measures indicated that the NAEP fraction concepts measure alone was the strongest screener measure for the prediction of the fifth-grade and sixth-grade outcome. The 19 items of the fifth-grade NAEP fraction concepts measure were entered into best subset ALM analyses to assess

the best subset of items for predicting later mathematics achievement. Separate ALM analyses were run for the two grades of the mathematics achievement outcome variable: fifth and sixth grade.

For the prediction of the fifth-grade outcome, the final model of predictor items accounted for 37% of the variation in the outcome variable. This amount of explained variance is noticeably lower than the variance explained in the models with fourth-grade predictor measures. A subsequent section of the Results section (i.e., ROC Curve Analyses to Test the Best Subset Measure) provides further analysis of this best subset measure and compares its diagnostic accuracy with the diagnostic accuracy of the full 19-item NAEP measure. The best subset model included 11 items from the full 19-item measure (See Tables 16 and 17).

For the prediction of the sixth-grade outcome, the best subset final model accounted for 59% of the variance in the outcome measure. The final model included 11 items, eight of which were also included in the final model predicting the fifth-grade outcome (NAEP items 3, 5, 8, 14, 15, 17, 18, and 19; See Tables 16 and 17).

Sixth-Grade Best Subset Measure

ROC curve results indicated that both the sixth-grade NAEP fraction concepts measure and the sixth-grade fraction number line estimation measure were strong predictor measures of students' later mathematics achievement. Items from both measures were assessed in a best subset automatic linear modeling (ALM) analysis to predict the sixth-grade mathematics achievement outcome. The ALM analysis determined the most predictive combination of items from the 24-item NAEP fraction concepts measure and the 28-item fraction number line estimation measure.

The final best subset model yielded an adjusted R^2 of .69, meaning that the subset of predictor items accounted for 69% of the variance in the outcome. The model included 11 NAEP items and three fraction number line items (See Tables 16 and 17). The 14-item

combined best subset measure had 10 fewer items than the original 24-item measure and 14 fewer items than the original 28-item fraction number line measure. Binary logistic regression with hierarchical entry revealed that combining the two best subset measures significantly improved the model prediction of the sixth-grade outcome. Again the model met underlying assumptions and showed good model fit with the data as indicated by the Hosmer-Lemeshow goodness-of-fit test (p = 454). Table 21 presents the regression coefficients.

Table 21 Regression Coefficients of Sixth-Grade Best Subset Measures Predicting the Sixth-Grade Mathematics Achievement Outcome

Predictor	*B*(SE)	Wald
Best Subset NAEP Fraction Concepts	-0.74(.10)	55.52
Best Subset FNLE	0.06(.02)	15.17
Constant	2.78(.75)	13.75

ROC Curve Analyses to Test the Best Subset Measures

Additional ROC curve analyses were conducted to assess the diagnostic accuracy of the best subset measures. For example, the diagnostic accuracy of the fourth-grade combined best subset measure was included in a ROC curve analysis to examine whether it performed better or equally as well as the original measures (i.e., measures with all items included) and the individual best subset measures (i.e., the best subset NAEP fraction concepts measure and the best subset fraction number line estimation measure) for predicting the fourth-grade mathematics achievement outcome. For each ROC curve analysis, the AUC values of the original NAEP measure and original fraction number line measure match the AUC values reported previously; the values were reported again for ease of comparison to the best subset measures.

Fourth-Grade Measures Predicting Mathematics Achievement in Fourth, Fifth, and Sixth Grades

The following five predictor measures were included in ROC curve analyses of fourth-grade measures predicting each grade of the outcome: (a) the original fourth-grade NAEP fraction concepts measure with all 18 items included, (b) the best subset NAEP fraction concepts measure, (c) the original fraction number line estimation measure with all 28 items included, (d) the best subset fraction number line estimation measure, and (e) the combination of the best subset NAEP fraction concepts measure and the best subset fraction number line estimation measure. In each analysis with the fourth-grade predictor measures, the combined best subset measure included 16 total items. Three separate ROC curve analyses were conducted for each grade of the outcome: fourth, fifth, and sixth grade.

Table 22 presents a summary of AUC values associated with the fourth-grade predictor measures in each ROC curve analysis. The combined best subset measure met the .750 AUC threshold for being a good screener for each grade of the outcome (Cummings & Smolkowski, 2015). AUC comparisons between the combined best subset measures and original measures in each analysis yielded consistent results across analyses: The combined best subset measure with 16 items performed just as well as the original NAEP fraction concepts measure with 18 items and the original fraction number line estimation measure with 28 items. In other words, the fourth-grade combined best subset measure in each analysis did not perform significantly better or significantly worse than the highest-performing original measure ($p > .05$). As such, the combined best subset measure performed equally as well but with fewer items.

The AUC of the combined best subset measure was also compared to the AUC of each individual best subset measure. The results of these comparisons were consistent across the ROC analyses of fourth-grade measures predicting later mathematics

achievement in fourth, fifth, and sixth grades. The fourth-grade combined best subset measure did not perform significantly better or significantly worse than the individual best subset NAEP fraction concepts measure or the individual best subset fraction number line estimation measure ($p > .05$). These results indicate that the combined best subset measure performed just as well as each individual best subset measure. The results further suggest that the combined best subset measure or either individual best subset measure as the preferred fourth-grade screener could be identified. However, the present study considers four reasons for recommending the combined best subset measure over the individual best subset measures for predicting later achievement. First, the results of the AUC comparisons contradict the results of the binary logistic regression analyses reported in the previous section. The regression analyses indicated that including both best subset measures significantly improved model fit when predicting all grades of the outcome measure. Second, although the AUC comparisons were not significantly different, the AUC value of the combined best subset measure in each analysis was consistently slightly higher than the AUC value of each individual best subset measure (see Table 22). Third, combining the best subset measures consistently yielded a reasonable total number of items (i.e., 16 total items) that would not require extensive time to administer to students. Fourth, the small amount of additional items on the combined best subset measure means that teachers would gain additional information of their students' strengths and/or misconceptions by looking at performance on an item-level without requiring too much extra time to administer the screener during class time. The fourth-grade combined best subset measures per each grade of the outcome are hence

recommended as the preferred screening measures over the original measures and the individual best subset measures.

Fifth-Grade Measures Predicting Mathematics Achievement in Fifth and Sixth Grades

The ROC analyses of the fifth-grade measures predicting the fifth- and sixth-grade outcome included the best subset of the NAEP fraction concepts measure rather than a combination measure. As such, the analysis of fifth-grade measures included the following: (a) the original fifth-grade NAEP fraction concepts measure with 19 items, (b) the original fraction number line measure with 28 items, and (c) the best subset NAEP fraction concepts measure with only 11 items. Table 22 reports the AUC values for all predictor measures. The best subset NAEP measure in fifth and sixth grades met the .850 AUC threshold for being a very good screener of risk status (Cummings & Smolkowski, 2015). AUC comparisons revealed that the best subset measure did not perform significantly better or significantly worse than the original measures ($p > .05$). The best subset NAEP measures performed just as well as the other measures but with fewer items. Thus, the present study recommends the best subset NAEP measures as the preferred fifth-grade screening measures for predicting the mathematics achievement outcome in both fifth and sixth grades.

Sixth-Grade Measures Predicting Mathematics Achievement in Sixth Grade

The following five measures were included in the ROC analysis of the sixth-grade measures predicting the sixth-grade mathematics achievement outcome: (a) the original sixth-grade NAEP fraction concepts measure with all 24 items included, (b) the best subset NAEP fraction concepts measure with 11 items, (c) the original fraction number line estimation measure with all 28 items included, (d) the best subset fraction number line estimation measure with three items, and (e) the combination of the best subset NAEP fraction concepts measure and the best subset fraction number line estimation

measure with 14 items. Since the sixth-grade measures predict out to only one grade of the outcome in the present study (i.e., sixth grade), only one ROC analysis was conducted for the sixth-grade measures. Again, the AUC values of the original NAEP fraction concepts measure and fraction number line measure match the values reported earlier in the present study.

Table 22 reports the AUC values for all sixth-grade predictor measures. The AUC of the combined best subset measure met the .85 threshold for being considered a very good screener (Cummings & Smolkowski, 2015). AUC comparisons revealed that the combined best subset measure with 14 items performed just as well as the original NAEP fraction concepts measure with 24 items and the original fraction number line estimation measure with 28 items. In other words, the sixth-grade combined best subset measure in each analysis did not perform significantly better or significantly worse than the highest-performing original measure ($p > .05$). As such, the combined best subset measure performed equally as well but with fewer items.

The AUC of the combined best subset measure in sixth grade was also compared to the AUC of each individual sixth-grade best subset measure. The combined best subset measure performed significantly better than the sixth-grade best subset fraction number line estimation measure ($p < .05$) but did not perform significantly better than the sixth-grade best subset NAEP fraction concepts measure ($p > .05$). These results suggest that the present study could recommend the combined best subset measure or the best subset NAEP fraction concepts measure alone as the preferred sixth-grade screener. The present study uses the same four reasons described previously to recommend the combined best subset measure in sixth grade instead of the best subset NAEP fraction concepts measure. Furthermore, combining the sixth-grade best subset fraction number line measure with the best subset NAEP fraction concepts measure is an addition of only three fraction number line items (see Table 17).

Table 22 Area Under the Curve (AUC) Statistics and 95% Confidence Intervals (CI) for Best Subset and Original Predictor Measures with All Items Included

Predictor Measure	AUC[1] [95% CI] 4th-Grade Outcome	5th-Grade Outcome	6th-Grade Outcome
4th grade			
All items NAEP Concepts	.796[1] [.749, .843]	.789[1] [.740, .838]	.791[1] [.738, .844]
Best Subset NAEP Concepts	.749 [.701, .796]	.757[1] [.703, .811]	.771[1] [.717, .826]
All items FNLE	.766[1] [.715, .817]	.745 [.693, .798]	.760[1] [.706, .814]
Best Subset FNLE	.779[1] [.728, .828]	.732 [.679, .784]	.800[1] [.750, .850]
Combined Best Subset Measure	.780[1] [.729, .828]	.785[1] [.736, .834]	.808[1] [.762, .854]
5th grade			
All items NAEP Concepts		.856[1] [.818, .894]	.880[1] [.842, .917]
All items FNLE		.784[1] [.738, .829]	.810[1] [.764, .856]
Best Subset NAEP Concepts		.850[1] [.810, .890]	.879[1] [.842, .915]
6th grade			
All items NAEP Concepts			.895[1] [.861, .929]
Best Subset NAEP Concepts			.881[1] [.844, .917]
All items FNLE			.864[1] [.826, .905]
Best Subset FNLE			.800[1] [.751, .849]
Combined Best Subset Measure			.899[1] [.866, .933]

Summary of ROC Curve Analyses Testing the Best Subset Measures

The present study identified fraction screening measures for fourth, fifth, and sixth grades that are supported by empirical data for predicting later mathematics achievement with high diagnostic accuracy. A combination best subset screener measure

is recommended for fourth and sixth grades; a best subset NAEP fraction concepts measure is recommended for fifth grade. As shown previously in Tables 16 and 17, the items selected for each grade of the best subset measures differed by grade of the outcome. Thus, a total of six separate screeners were identified: (a) a fourth-grade combined best subset screener for predicting the fourth-grade outcome, (b) a fourth-grade combined best subset screener for predicting the fifth-grade outcome, (c) a fourth-grade combined best subset screener for predicting the sixth-grade outcome, (d) a fifth-grade best subset NAEP fraction concepts screener for predicting the fifth-grade outcome, (e) a fifth-grade best subset NAEP fraction concepts screener for predicting the sixth-grade outcome, and (f) a sixth-grade combined best subset screener for predicting the sixth-grade outcome. Next, each screener is explored further for use in real-world settings.

Translating Screener Statistics into Usable Information for Researchers and Practitioners

Different statistics are computed and reported depending on whether the screener is a single best subset measure or a combined best subset measure. For a single best subset screener, the typical ROC curve statistics were reported (e.g., true positive fraction and true negative fraction; Cummings & Smolkowski, 2015), along with the statistics necessary for using a probability nomogram (Youngstrom, 2014). A different method was used in the present study for combining a measure with continuous data (i.e., the fraction number line estimation measure) with a measure with discrete data (i.e., the NAEP fraction concepts measure); a predicted probabilities matrix was designed that allows a researcher or teacher to make predictions of a student's performance by using his/her scores on *both* best subset measures (Clemens et al., 2016).

Fourth-Grade Combined Best Subset Screener for Predicting Fourth-Grade Achievement

Earlier analyses in the previous section indicated that a combined best subset measure was the preferred screener for predicting fourth-grade mathematics achievement.

Table 23 presents the predicted probabilities matrix of not meeting the fourth-grade mathematics standard associated with the best subset six-item NAEP fraction concepts scores and the best subset fraction number line estimation scores. Due to the continuous nature of the fraction number line task measured in percent absolute error (PAE), the number line scores were grouped into six categories based on percentiles (see Clemens et al., 2016). The categories were determined by the following percentiles: (a) less than or equal to the 10^{th} percentile, (b) greater than the 10^{th} percentile and less than or equal to the 25^{th}, (c) greater than the 25^{th} and less than or equal to the 50^{th}, (d) greater than the 50^{th} and less than or equal to the 75^{th}, and (e) greater than the 75^{th} and less than the 90^{th}, and (f) greater than or equal to the 90^{th} percentile. The percent absolute error (PAE) scores were recoded (i.e., 100 – PAE) for the matrix so that lower scores indicate poorer performance, just as lower scores on the NAEP measure indicate poorer performance. The recoded number line scores can be thought of as "percent absolute accuracy" rather than percent absolute error. In contrast, the matrix presents all values of the six-item best subset NAEP measure.

The matrix allows for interpretation of any combination of scores. The matrix was designed so that poorer performance scores fall into cells at the top left corner of the matrix; these cells are expected to have high predicted probabilities for not meeting the mathematics standard at the end of the year. For example, a student with a score of zero on the best subset NAEP fraction concepts measure *and* a percent absolute accuracy equal to or less than 63.40% has a predicted probability range of .71-.80. This range indicates that the student has a high probability between 71%-80% of not meeting the fourth-grade mathematics standard at the end of the year.

In contrast, higher performance scores fall into cells at the lower right corner; these cells are expected to have lower predicted probabilities for not meeting the mathematics standard. For example, a student with a score of six on the six-item best

subset NAEP measure *and* a fraction number line score equal to or greater than 91.90% has a probability of only 1% of not meeting the mathematics standard.

Table 23 Combination Matrix Summarizing Predicted Probabilities of Fourth-Grade Screeners Associated with Not Meeting the Fourth-Grade Mathematics Standard

Best Subset FNLE	Best Subset NAEP Fraction Concepts						
	0	1	2	3	4	5	6
≤ 63.40	.71-.80	.61-.85	.57-.83	.52-.73	.49-.75	.38-.61	.33-.44
63.41-68.33	.62-.65	.48-.60	.41-.53	.35-.48	.29-.41	.25-.35	.26-.27
68.34-74.24	--	--	.27-.40	.21-.31	.17-.28	.13-.21	.12-.18
74.25-81.44	--	--	.14-.22	.11-.19	.10-.16	.07-.13	.04-.09
81.45-91.89	--	--	.07-.11	.04-.10	.02-.07	.02-.06	.01-.03
≥ 91.90	--	--	--	--	--	.01-.02	.01

Note. Empty cells indicate that too few students achieved that combination of scores to generate a probability value
FNLE = Fraction number line estimation (measured in percent absolute accuracy)

Fourth-Grade Combined Best Subset Screener for Predicting Fifth-Grade Achievement

A combined best subset screener in fourth grade was recommended for predicting fifth-grade mathematics achievement. Table 24 shows the predicted probabilities matrix for the combination screener for the prediction of the fifth-grade outcome. The matrix presents fraction number line scores grouped into the same six categories described above based on percentiles and all possible scores on the five-item NAEP best subset measure. The lowest performing students (i.e., score of zero on the NAEP and percent absolute accuracy less than or equal to 66.25% on the fraction number line measure) have a predicted probability of 77%-84%. This range of predicted probabilities indicates that a student has a probability of 77% to 84% of not meeting the fifth-grade mathematics standard.

Table 24 Combination Matrix Summarizing Predicted Probabilities of Fourth-Grade Screeners Associated with Not Meeting the Fifth-Grade Mathematics Standard

Best Subset FNLE	Best Subset NAEP Fraction Concepts					
	0	1	2	3	4	5
≤ 66.25	.77-.84	.65-.73	.51-.58	.37-.44	.26-.32	--
66.26-75.74	.68-.73	.58-.61	.40-.50	.28-.37	.19-.26	.13-.17
70.32-75.75	.63-.64	.39-.52	.28-.39	.18-.29	.12-.18	--
75.76-81.48	.40-.49	.27-.31	.25-.26	.17-.18	.11-.12	.05-.07
81.49-90.17	--	--	.16-.17	.07-.10	.05-.07	.03-.04
≥ 90.18	--	--	--	.04	.02-.03	.01-.02

Note. Empty cells indicate that too few students achieved that combination of scores to generate a probability value

Fourth-Grade Combined Best Subset Screener for Predicting Sixth-Grade Achievement

For the prediction of sixth-grade mathematics achievement, a combination of the NAEP fraction concepts screener and the fraction number line screener emerged once again as the superior option. Table 25 shows the predicted probabilities matrix for the combination screener for the prediction of the sixth-grade outcome. The matrix presents fraction number line scores grouped into the six percentile categories and all possible scores on the six-item NAEP best subset measure. The lowest performing students (i.e., score of zero on the NAEP and fraction number line score less than or equal to 66.23%) have a predicted probability of 68%. This predicted probability indicates that a student has a probability of 68% of not meeting the mathematics standard at the end of sixth grade.

Table 25 Combination Matrix Summarizing Predicted Probabilities of Fourth-Grade Screeners Associated with Not Meeting the Sixth-Grade Mathematics Standard

Best Subset FNLE	Best Subset NAEP Fraction Concepts						
	0	1	2	3	4	5	6
≤ 66.23	.68	.60-.80	.51-.78	.46-.65	.34-.54	.38-.43	.19-24
66.24-71.69	.60-.68	.48-.59	.38-.44	.31-.37	.23-.38	.19-.25	.17-.24
71.70-75.80	.50-.55	.33-.39	.26-.30	.20-.22	.15-.17	.12-.14	.10-.12
75.81-81.00	--	.18-.30	.15-.25	.11-.14	.07-.10	.05-.07	.04-.07
81.01-89.29	--	.07-.16	.07-.12	.04-.09	.02-.04	.02-.03	.01-.02
≥ 89.30	--	--	--	.02-.03	.01-.02	.01-.02	.00-.01

Note. Empty cells indicate that too few students achieved that combination of scores to generate a probability value
FNLE = Fraction number line estimation (measured in percent absolute accuracy)

Fifth-Grade Best Subset NAEP Fraction Concepts Screener for Predicting Fifth-Grade Achievement

Analyses of the present study assessing fifth-grade measures indicated that the best subset NAEP fraction concepts screener was the most predictive and practical screener option, rather than a combination of measures. Table 26 presents all possible cut scores and associated ROC statistics (i.e., true positive fraction, true negative fraction, diagnostic likelihood ratios, positive predictive power, and negative predictive power) for the best subset 11-item NAEP fraction concepts screener.

The diagnostic likelihood ratio for a positive screener result (DLR+) allows a researcher or teacher to translate the results for his/her own sample of students by using the probability nomogram. To minimize "missing" at-risk students, teachers may desire a cut score with an 85% true positive threshold (Jordan et al., 2010); this value indicates that 85% of truly at-risk students were accurately identified by the cut score. A teacher can identify the cut score associated with an 85% true positive fraction, find the DLR+ associated with the cut score, and finally use the probability nomogram to reveal the positive predictive power (PPV). The positive predictive power indicates the likelihood

of a student not meeting the mathematics standard at the end of the year. In other words, the positive predictive power allows a teacher to make a forward prediction of a student's later mathematics achievement.

Table 26 ROC Curve Statistics Associated with all Possible Cut Scores on the Fifth-Grade NAEP Best Subset Screener for Predicting the Fifth-Grade Outcome

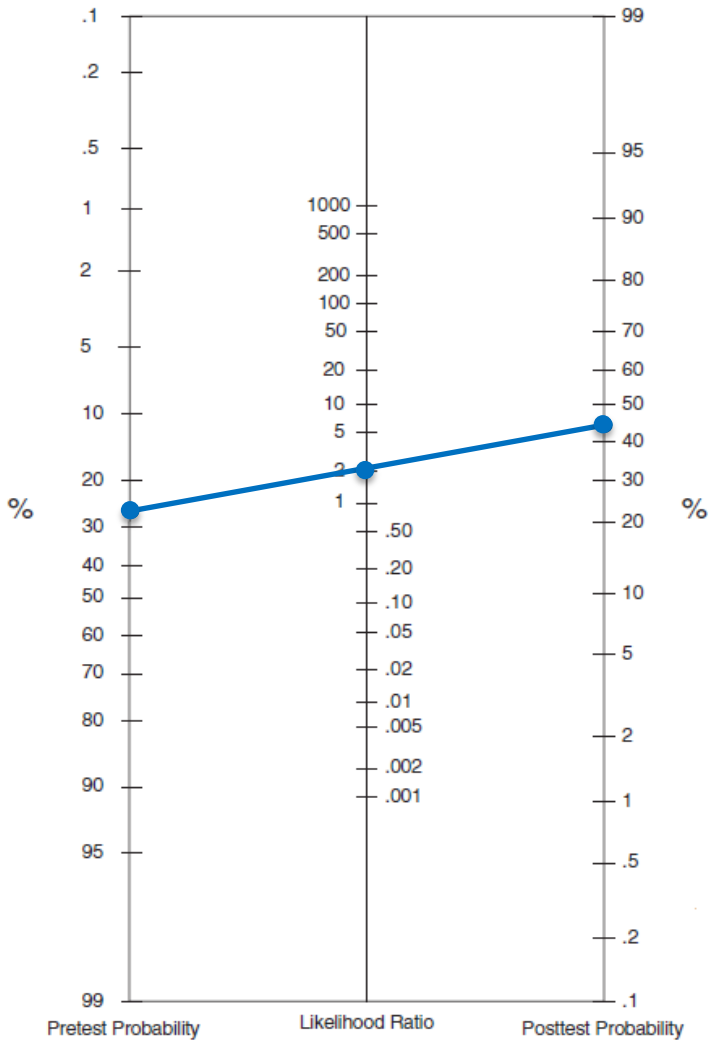
	Independent of Base Rate				Dependent on Base Rate of 27%	
Cut Score	True Positive Fraction	True Negative Fraction	DLR+	DLR-	PPV	NPV
≤ 1	.00	1.00	--	1.00	--	.73
≤ 2	.03	.99	3.00	.98	.53	.73
≤ 3	.15	.98	7.50	.87	.72	.75
≤ 4	.28	.96	7.08	.75	.73	.78
≤ 5	.47	.91	5.19	.59	.66	.82
≤ 6	.68	.84	4.20	.39	.61	.87
≤ 7	.83	.75	3.30	.23	.55	.92
≤ 8	.92	.58	2.21	.13	.45	.95
≤ 9	.97	.48	1.88	.06	.41	.98
≤ 10	.98	.34	1.49	.06	.36	.98
≤ 11	1.00	.00	1.00	--	.27	--

Note. The cell highlighted in gray indicates the cut score associated with the 85%-true positive fraction threshold.
DLR = diagnostic likelihood ratio
PPV = positive predictive power
NPV = negative predictive power

Figure 7 presents an example for using the probability nomogram to translate the ROC statistics. A dot is placed on the left line of the nomogram that corresponds to the base rate (e.g., 27%). Next, a line is extended from the dot to the location on the middle line that corresponds to the DLR+ reported in the present study (i.e., 2.21). Last, the line is further extended to the right side of the nomogram, revealing the positive predictive power (PPV) that is specific to the sample base rate (i.e., 45%). The interpretation would

be as follows: 45% of students who fall below the cut score are at risk for not meeting the fifth-grade mathematics standard. If a higher probability was desired, the researcher or teacher could select a lower cut score associated with higher positive predictive power.

Figure 7 Probability Nomogram Example of Fifth-Grade Screener Predicting the Fifth-Grade Outcome



Fifth-Grade Best Subset NAEP Fraction Concepts Screener for Predicting Sixth-Grade Achievement

The present study identified the best subset NAEP fraction concepts screener as the most predictive and practical screener option in fifth grade for the prediction of the sixth-grade outcome. Table 27 presents all possible cut scores and associated ROC curve statistics for the best subset 11-item NAEP fraction concepts screener.

Table 27 ROC Curve Statistics Associated with all Possible Cut Scores on the Fifth-Grade NAEP Best Subset Screener for Predicting the Sixth-Grade Outcome

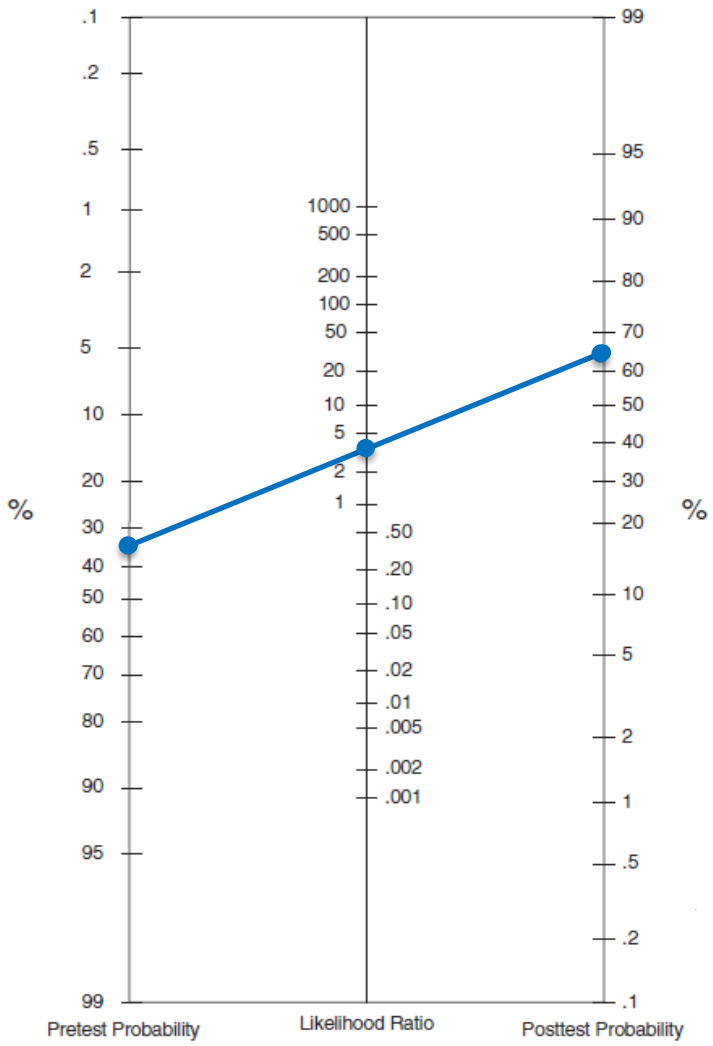
	Independent of Base Rate				Dependent on Base Rate of 34%	
Cut Score	True Positive Fraction	True Negative Fraction	DLR+	DLR-	PPV	NPV
≤ 1	.00	1.00	--	1.00	--	.66
≤ 2	.01	1.00	--	.99	--	.67
≤ 3	.08	1.00	--	.93	--	.68
≤ 4	.26	.99	18.71	.75	.90	.72
≤ 5	.46	.97	13.88	.56	.88	.78
≤ 6	.63	.86	4.41	.44	.69	.82
≤ 7	.85	.77	3.66	.20	.65	.91
≤ 8	.94	.59	2.29	.10	.54	.95
≤ 9	.99	.45	1.79	.02	.48	.99
≤ 10	1.00	.28	1.40	.00	.41	1.00
≤ 11	1.00	.00	1.00	--	.34	--

Note. The cell highlighted in gray indicates the cut score associated with the 85%-true positive fraction threshold.

Figure 8 presents an example for using the probability nomogram to translate the ROC statistics into usable information for any sample of students. First, a dot is placed on the left line of the nomogram that corresponds to the base rate (e.g., 34%). Second, a line is extended from the dot to the location on the middle line that corresponds to the DLR+ reported in the present study for a specific screener cut score (i.e., 3.66). Last, the line is

further extended to the right side of the nomogram, revealing the positive predictive power (PPV) that is specific to the sample base rate (i.e., 65%). A researcher or teacher would make the following interpretation of the PPV: 65% of students who fall below the cut score are at risk for not meeting the sixth-grade mathematics standard.

Figure 8 Probability Nomogram Example of Fifth-Grade Screener Predicting the Sixth-Grade Outcome



Sixth-Grade Combination Screener for Predicting Sixth-Grade Achievement

Earlier analyses of the present study indicated that a combination of screeners was the superior sixth-grade screener option when predicting sixth-grade mathematics achievement. Table 28 presents the predicted probabilities matrix of not meeting the sixth-grade mathematics standard associated with all possible scores on the best subset 11-item NAEP fraction concepts scores and the best subset 3-item fraction number line estimation scores. At this time point, the lowest score received on the best subset NAEP screener was a score of one point. A student with a score of one on the best subset NAEP screener and a fraction number line score less than or equal to 63.38% has a predicted probability of .98. This value means that the student has a 98% chance of not meeting

Table 28 Combination Matrix Summarizing Predicted Probabilities of Sixth-Grade Screeners Associated with Not Meeting the Sixth-Grade Mathematics Standard

Best Subset FNLE	Best Subset NAEP Fraction Concepts											
	0	1	2	3	4	5	6	7	8	9	10	11
≤ 63.38	--	.98	.97-.98	.95-.96	.88-.90	.78-.82	.62-.71	.45-.57	.28-.35	--	--	--
63.39-70.03	--	--	--	.91-.93	.84-.88	.71-.75	.53-.61	.36-.42	.21-.25	.11-.14	--	--
70.04-82.15	--	.96-.98	.91-.94	.83-.90	.72-.83	.55-.63	.37-.43	.23-.34	.11-.20	.06-.10	.03-.05	--
82.16-91.00	--	--	--	.82-.83`	.61-.69	.42-.52	.25-.35	.14-.20	.07-.09	.03-.05	.02-.03	.01
91.01-96.41	--	--	--	--	--	.33-.39	.19-.24	.11-.12	.05-.06	.03	.01-.02	.01
≥ 96.42	--	--	--	--	--	--	--	--	--	.02-.03	.01	.01

Note. Empty cells indicate that too few students achieved that combination of scores to generate a probability value
FNLE = Fraction number line estimation (measured in percent absolute accuracy)

the sixth-grade standard at the end of the school year. In contrast, a higher-achieving student with a score of 11 on the best subset NAEP screener and a fraction number line score equal to or greater than 96.42% has only a 1% chance of not meeting the mathematics standard.

Summary of Results

Different screeners emerged for each time point of the present study as the preferred screener for predicting later risk status. The selected *fourth-grade* screener for the prediction of fourth-grade mathematics achievement was a combined best subset measure with 16 total items (i.e., six NAEP fraction concepts items and 10 fraction number line items). The selected fourth-grade screener for the prediction of fifth-grade mathematics achievement was also a combined best subset measure with 16 total items (i.e., five NAEP fraction concepts items and 11 fraction number line items). Similarly, the selected fourth-grade screener for the prediction of the sixth-grade outcome was a combined best subset measure with 16 items (i.e., six NAEP fraction concepts items and 10 fraction number line items). All three of the combined best subset screeners met the AUC threshold for determining students' risk status for later mathematics difficulties. The present study reported a matrix that allows researchers or teachers to make interpretations of students' scores. For example, a teacher can administer the two best subset measures to his/her class and make the following interpretation: Student A received a score of two points on the NAEP fraction concepts screener and a score of 64% on the fraction number line measure. Student A has a 41%-53% chance of not meeting the mathematics standard at the end of the year.

The best *fifth-grade* screener for predicting the fifth-grade outcome was a best subset NAEP fraction concepts measure with 11 items. Similarly, the best fifth-grade screener for predicting the sixth-grade outcome was a best subset NAEP fraction concepts measure also with 11 items that differed slightly from the items included in the

other best subset NAEP measure. The two screeners met the thresholds for being very good screeners of risk status. A table of ROC curve statistics and a probability nomogram were reported to help researchers and teachers translate the ROC results into usable information. For example, a teacher can administer the screener to his/her own students and leverage the statistics reported in the present study to make interpretations of a student's later achievement. For example, a teacher could make the following interpretation: Student A received a score of two points on the screener and thus has a 75% chance of not meeting the mathematics standard at the end of the year.

The selected *sixth-grade* screener for the prediction of the sixth-grade outcome was a combined best subset measure of 14 items (i.e., 11 NAEP fraction concepts items and three fraction number line items). The screener meets the threshold for being a very good screener of students' later achievement. The present study again reported a matrix that allows researchers or teachers to make interpretations based on both best subset measures. For example, a teacher could make the following interpretation: Student A received a score of three points on the NAEP fraction concepts screener and a score of 70% on the fraction number line measure. Student A thus has a 91%-93% chance of not meeting the mathematics standard at the end of the year.

Chapter 5

DISCUSSION

The development of mathematics screeners for the intermediate grades is a priority. Students who are at risk for later mathematics difficulties may not receive the additional mathematics help they need in the absence of valid screening measures. Previous research has focused on mathematics screeners for younger children, neglecting the importance of screeners for third grade and beyond (Gersten et al., 2012). Students who are not identified as at risk in the early grades (i.e., kindergarten through second grade) may start to encounter difficulties in later grades when abstract topics are introduced in the curriculum, such as fractions. Fractions are a challenging topic for many elementary and middle school students (e.g., Hansen et al., in press; Ni & Zhou, 2005). As such, the present study assessed fraction measures as potential screeners for fourth, fifth, and sixth grades for the prediction of later mathematics achievement. More specifically, the study assessed three different measures at each grade: (a) NAEP fraction concepts, (b) fraction number line estimation, and (c) fraction arithmetic. The outcome variable was the end-of-the-year state mathematics achievement test (i.e., DCAS). The test was assessed as a binary outcome: (a) those who did not meet the established state mathematics standard versus (b) those who met the standard. The present study used data from a larger longitudinal project that followed students from fourth through sixth grade on a variety of fraction measures.

Screeners with High Diagnostic Accuracy

The three fraction measures were assessed in fourth, fifth and sixth grades, respectively. Separate analyses were run for each grade of the mathematics achievement

outcome measure. The fourth-grade measures predicted out to all three grades of mathematics achievement (i.e., the spring of fourth, fifth and sixth grades), the fifth-grade measures predicted out to two grades of the outcome measure (i.e., the spring of fifth and sixth grades), and the sixth-grade measures predicted out to the spring of sixth grade. The present study sought to identify one "best" screener for each of the six ROC curve analyses.

Fourth-Grade and Sixth-Grade Screeners

Analyses of the fourth-grade measures predicting out to all grades of the outcome (i.e., fourth, fifth, and sixth grades) and analyses of the sixth-grade measures predicting out to the end of sixth grade yielded similar results. As hypothesized, the two fraction concepts measures in fourth and sixth grades emerged as accurate screeners of students' performance on the mathematics achievement outcome. The NAEP fraction concepts screener consistently held higher (but not significantly better) predictive power than the fraction number line screener, which contradicts the hypothesis that the number line measure would be the most predictive fraction measure. However, researchers have suggested that multiple-proficiency screeners may be more powerful and accurate than a screener that targets only one discrete skill (e.g., Foegen et al., 2007; Purpura et al., 2015). This line of thinking supports the result of the present study, as the NAEP fraction concepts measure includes multiple types of fraction concepts items (e.g., part-whole items, fraction comparisons items, and number line items) and is hence categorized as a multiple-proficiency screener.

The fraction arithmetic measure in fourth grade did not meet the statistical threshold for being a good screener of students' later performance. Thus, further analyses of fourth-grade screeners excluded the fraction arithmetic measure and combined the two fraction concepts measures. In sixth grade, the fraction arithmetic measure did show high diagnostic accuracy for predicting the outcome. Fraction arithmetic may be more

important to mathematics achievement in sixth grade than in earlier grades since sixth-graders have had more instruction with different fraction operations. However, the fraction arithmetic measure performed significantly worse than the NAEP fractions concepts measure. Subsequent analyses of sixth-grade screeners thus excluded the fraction arithmetic measure and combined the two fraction concepts screeners.

Fifth-Grade Screeners

Fifth-grade screeners for NAEP fraction concepts and number line estimation both met the threshold for being good screeners for each grade of the outcome. However, the multiple-proficiency NAEP fraction concepts measure emerged as the most predictive screener, this time significantly outperforming fraction number line measure as well as fraction arithmetic. These findings also contradict the hypothesis that the number line measure would be the most predictive measure. Once again, the results of the present study provide support of multiple-proficiency screeners (e.g., the NAEP measure) outperforming single-proficiency screeners (e.g., the fraction number line measure; Foegen et al., 2007; Purpura et al., 2015). As such, subsequent explorations of the fifth-grade screeners focused on the NAEP fraction concepts measure.

Summary of Screener Results

Overall, the two fraction concepts screeners emerged as accurate screeners of students' mathematics achievement, with fraction arithmetic being less predictive. These findings support previous research suggesting that conceptual understanding of key concepts (e.g., fraction magnitude) may be more important than arithmetic skill for mathematics achievement (e.g., Hecht et al., 2003). Although both mathematics concepts and procedures are recognized in the literature as important competencies (e.g., Hallet et al., 2010; Hecht et al., 2003), fraction concepts seem to have a stronger relation to overall mathematics achievement. Conceptual knowledge allows students to make sense of

procedures and to notice procedural errors in their own work (Hecht, 1998). For example, students who understand fraction concepts may realize they cannot add across denominators for a fraction addition problem with unlike denominators (2/3 + 2/6 = 4/9) (a common mistake; Newton, Willard, & Teufel, 2014) because thirds and sixths are different sizes and the fractions cannot be added without first identifying like denominators. Moreover, numerical magnitude knowledge, including knowledge of fraction magnitudes, provides a supporting structure for learning different mathematical concepts (Siegler et al., 2011).

The finding that NAEP fraction concepts significantly outperformed fraction number line estimation in fifth grade is of interest, since this result did not occur in the analyses of the fourth-grade and sixth-grade screeners. The finding is likely affected by procedural issues. That is, in fifth grade the number line measure was administered in the fall rather than the winter as in fourth and sixth grades. Thus, students completed the number line measure earlier in the school year, before completing the fraction curriculum in fifth grade. Less time between administrations often increases or inflates the predictive value of a screener (Cummings & Smolkowski, 2015). That is, a screener is likely to appear more accurate when administered one month before the outcome measure instead of five months before an outcome measure. Thus, it is important to consider whether the fraction number line measure would perform similarly to the NAEP measure if administered in the fall of fourth grade and the fall of sixth grade instead of the winter (which is in closer proximity to the outcome).

Best Subset Screeners

Teachers frequently lament the amount of instructional time lost in the classroom due to testing (Cobb, 2003). Researchers Clemens and colleagues (2016) also discuss the concerns of testing time, especially when using more than one screener. Multiple-measure screening batteries can waste both precious resources and instructional time. In

response, the present study sought to limit the amount of items on each screener by identifying the most predictive subset of items and removing the least predictive items while ensuring that the best subset of items maintained the high diagnostic accuracy of the original measures with all items included. Excluding items that did not discriminate well between students who are at risk for later difficulties and students who are not at risk minimizes the amount of time required for administration and for scoring, making the screener much more practical for classroom use. Furthermore, examining retained items on each best subset screener provided information about the type of items and/or concepts that are most important for predicting students' later success.

Fourth-Grade Combined Best Subset Screeners Predicting Mathematics Achievement in Fourth, Fifth, and Sixth Grades

At fourth grade, the NAEP fraction concepts screener was combined with the fraction number line screener. Without identifying the best subset of items on each measure, the combined screener would have a total of 46 items (i.e., 18 items on the NAEP screener and 28 items on the fraction number line screener). Instead, the combined best subset screeners each resulted in 16 predictive items for the different grades of the overall mathematics achievement outcome. Importantly, the estimated time required to administer these shortened screeners is 11-13 minutes (i.e., an estimated 2 minutes per NAEP item and an estimated 5 seconds per number line item) as compared to approximately 38 minutes for the longer screener. Analyses demonstrated that the shorter 16-item best subset screeners in fourth grade all performed just as well in predicting overall mathematics achievement at the end of fourth, fifth, and sixth grades as the original measures for the prediction of later mathematics achievement.

The number of fraction number line items retained on each fourth-grade combined screener always exceeded the number of NAEP items. Thus, although the full NAEP measure was slightly more predictive than the full number line measure, the best

subset number line screener actually outperformed the best subset NAEP measure. This finding indicates that a small subset of number line items in fourth grade held especially high predictive power. Specifically, proper fractions emerged as most consistently predictive over most NAEP items, mixed numbers, and improper fractions. This finding is not surprising since mixed numbers and improper fractions are hard for students at this level. Students' estimates of three proper fractions on the 0-1 number line (i.e., 2/13, 1/3 and 5/6) and one proper fraction on the 0-2 number line (i.e., 3/8) emerged as highly predictive on all fourth-grade screeners. The whole number 1 as estimated on the 0-2 number line also emerged on all three screeners.

The importance of students' estimations of proper fractions early is in keeping with previous research showing that students have greater understanding of proper fractions than improper fractions (Resnick et al., 2016), most likely because early instruction in mathematics classrooms typically emphasizes proper fractions (i.e., fractions less than one) rather than improper fractions (i.e., fractions equal to or greater than one; Vosniadou, Vamvakoussi, & Skopeliti, 2008). It is important to consider that the fourth-grade predictor measures were administered before students typically receive targeted fraction instruction in the classroom (NGACBP & CCSSO, 2010). As such, the majority of fourth-grade students likely struggled with the improper fractions; in other words, the improper fraction items likely were not predictive because they were challenging for most students.

Fifth-Grade NAEP Best Subset Screeners Predicting Mathematics Achievement in Fifth and Sixth Grades

The best subset of NAEP fraction concepts items was identified for the prediction of each grade of the mathematics achievement outcome. Both fifth-grade NAEP best subset screeners retained 11 of the original 19 items, eight of which were consistent across the two screeners. Importantly, the shorter best subset NAEP screener performed

just as well as the original 19-item measure. Whereas the administration of the original 19-item NAEP measure required approximately 35 minutes, the 11 items would require approximately 22 minutes. A look at the eight consistent items shows a variety of fraction concepts: (a) NAEP items 3 and 15 assess part-whole understanding, (b) item 5 asks students to identify a fraction that is equivalent to one-half, (c) item 8 is a fraction computation question with common denominators, (d) item 14 targets fraction estimation, (e) item 17 asks students to compare the sizes of two fractions, (f) item 18 requires students to order three fractions from least to greatest, and (g) item 19 shows students three different measuring cups and asks them to identify how to use the cups to measure one and one-third cups of sugar. (See Appendix A for a list of all NAEP items.)

The finding that five fourth-grade level NAEP items discriminated well between fifth-grade students at risk for later difficulties and students not at risk suggests that the lower-achieving fifth-graders were still struggling with certain fraction concepts below their grade level. For example, NAEP item three shows students a rectangle with two of five parts shaded. Students are asked to identify the fraction of the figure that is shaded (i.e., 2/5). NAEP categorizes this item as an easy fourth-grade item (U.S. Department of Education, 1990-2009). As such, the majority of fifth-grade students would be expected to answer this item correctly. However, its emergence on both best subset screeners indicates that the item is a strong fifth-grade predictor of students who do not meet the end-of-the-year mathematics standard, most likely because it reflects weak conceptual knowledge in general.

Sixth-Grade Combined Best Subset Screener Predicting Mathematics Achievement in Sixth Grade

A combination of the NAEP fraction concepts and fraction number line items was assessed for the sixth-grade prediction of the sixth-grade outcome. The original NAEP measure in sixth grade included 24 items; the best subset NAEP retained 11 of these

items. The original fraction number line measure with 28 items was reduced to only three items. Thus, the combined best subset screener for sixth grade included a total of 14 items and was just as powerful as the original, longer measures for the prediction of the sixth-grade outcome. The estimated administration time of the 14-item screener is 17 minutes (i.e., an estimated 1.5 minutes per NAEP item and an estimated 5 seconds per number line item) as compared to the 38 minutes required of the full measures. The subset of NAEP items held more predictive power than the best subset number line items; this finding is inconsistent with the fourth-grade results that showed higher predictive power among the number line items.

A look at the retained NAEP items provides insight into the type of concepts that are predictive in sixth grade of students' end-of-the-year mathematics achievement. The retained NAEP items point to a variety of fraction concepts, with three particular concepts emerging more than once throughout the screener: (a) items 4 and 24 assess fraction equivalency; (b) items 11, 15, and 23 target part-whole understanding; and (c) items 13 and 17 ask students to compare fraction magnitudes. (Appendix A presents a list all NAEP items.)

Fewer items were retained from the fraction number line measure. Placing the fraction 3/7 on the 0-1 number line emerged as a predictive item, indicating that lower-achieving students struggled with this particular item and higher-achieving students made more accurate estimates. One possible explanation for lower-achieving students struggling to locate 3/7 on the number line is the uncommon denominator of seven. Typical mathematics curriculum focuses on halves, fourths, eighths, thirds, and sixths (e.g., Wittenberg et al., 2012); some students may struggle when confronted with a less familiar denominator. The other two retained number line items were on the 0-2 number line: 1/3 and 7/4. The predictive power of the proper fraction 1/3 on a 0-2 number line is especially interesting, because students also estimated 1/3 on the 0-1 number line. However, only the estimate on the 0-2 number line was retained on the final screener,

suggesting that the 0-1 item was simple for most sixth-grade students and the 0-2 item was more challenging. As discussed previously, typical classroom instruction focuses on proper fractions on 0-1 number lines (Vosniadou et al., 2008). The 0-2 number line is not only less familiar to many students but also presents new conceptual challenges, requiring students to first recognize the location of one whole in the middle of the 0-2 line. The other retained item was the improper fraction 7/4 on the 0-2 number line. Previous research has suggested that a major limitation of focusing fraction instruction on proper fractions is that students often view all fractions as "small" numbers between zero and one (Resnick et al., 2016; Vosniadou et al., 2008). Whereas higher-achieving students may immediately recognize that 7/4 is greater than one whole because the numerator (i.e., 7) is greater than the denominator (i.e., 4), struggling students may hold the misconception that all fractions are less than one whole and incorrectly place the fraction close to zero on the number line (Rodrigues et al., 2017).

Reporting Usable Screener Information for Researchers and Practitioners

Although the identification of powerful and predictive screeners is informative for research purposes, it is not sufficient for classroom application. Educational researchers must also report the screening information in ways that translate the statistics into accessible and usable information for real-world contexts. The present study demonstrates two ways of reporting screener information that translates statistics into understandable information for researchers and practitioners. The end goal is for teachers to use the screeners with their own sample of students and make predictive interpretations of student performance. For example, a practitioner should be able to administer the screener and make the following interpretation: Student A scored 14 points on the screener and thus has an 80% risk of not meeting the mathematics standard at the end of the year.

Next, the two methods for reporting screener statistics are discussed and examples from the present study are provided. The first method discussed is specific for single screeners and is demonstrated through the fifth-grade NAEP best subset screeners. The second method is specific for combined screeners that bring together two different types of data: discrete and continuous; this particular method is demonstrated through the fourth-grade and sixth-grade combined screeners of the present study. This method may also have been applicable for fifth grade had the number line screener been administered in the fall rather than the winter.

Reporting ROC Curve Statistics for Single Screeners

The present study provided one table and one nomogram for each single screener (i.e., the fifth-grade best subset NAEP screeners) that are essential for translating the screener results into usable information for researchers and practitioners. Importantly, the table and nomogram provide readers with the power to leverage the ROC statistics into understandable information that is specific to their own sample of students. In comparison, the current educational literature often reports a table with ROC statistics for only one specific cut score and hence limited information for readers (e.g., Seethaler & Fuchs, 2010) and requires readers to use a complicated formula to translate the ROC statistics into usable information (Cummings and Smolkowski, 2015).

Table of ROC Statistics for all Cut Scores

First, educational researchers must report a table that presents important ROC statistics: (a) true positive fraction, (b) true negative fraction, (c) diagnostic likelihood ratio for a positive screener result, (d) diagnostic likelihood ratio for a negative screener result, (e) positive predictive power, and (f) negative predictive power. Educational research of mathematics screeners for younger ages often report these ROC statistics but only for certain cut scores along the screener that are associated with high true positive fractions (e.g., Jordan et al., 2010; Seethaler & Fuchs, 2010). In other words, researchers

typically report only one cut score that is associated with a certain statistical threshold. For example, Seethaler and Fuchs (2010) report cut scores that are associated with true positive fractions of 90%; this value means that 90% of students who did not meet proficiency on the outcome were accurately identified by the cut score. The rationale for reporting a cut score with a high true positive fraction is to avoid "missing" at-risk students. When researchers prioritize a high true positive fraction, they cast a wide net for identifying students. However, readers may not have the available resources for casting such a wide net. Thus, the one cut score reported in the Seethaler and Fuchs (2010) study may not be practical for all researchers and practitioners. This possible scenario is a huge educational concern; researchers are conducting excellent research but are not reporting the research in ways that are most helpful to their readers.

The present study presents a solution to this educational concern by providing a table that reports ROC curve statistics for *all* possible cut scores on the fifth-grade NAEP screeners. By doing so, the present study does not pick and choose certain cut scores based on the present sample. Rather, all cut scores and associated ROC statistics are reported, allowing a researcher or practitioner to select a cut score based on available resources and goals.

Probability Nomogram

Second, educational researchers can include a probability nomogram that helps consumers translate the ROC statistics into usable information for their sample of students. Probability nomograms are suggested in the medical literature as an understandable way for clinicians to apply ROC statistics and to interpret patients' results on diagnostic tests (Florkowski, 2008; Youngstrom, 2014). However, the probability nomogram has not yet been proposed in the educational literature for interpreting students' screener scores. The present study urges educational researchers to include a

probability nomogram when reporting ROC curve statistics. The nomogram allows researchers and practitioners to make predictive interpretations of students' performance.

To demonstrate the power of the probability nomogram, imagine a teacher who administers the best subset 11-item NAEP screener to his fifth-grade students. The teacher is interested in identifying all students in his classroom who are at risk of not meeting the fifth-grade mathematics standard at the end of the school year. First, he refers to the ROC statistics table reported in the present study to select the cut score associated with an 85% true positive fraction; he also identifies the diagnostic likelihood ratio for a positive screener that is reported in table for his selected cut score. Next, he determines the base rate of students who did not meet the mathematics standard in the previous school year. He now has the two values necessary for using the probability nomogram: (a) base rate for his sample and (b) the diagnostic likelihood ratio for a positive screener result (DLR+). He draws a straight diagonal line on the probability nomogram, matching the base rate on the left side of the nomogram to the DLR+ on the middle line. He then extends the line to the right side of the nomogram, revealing the positive predictive power that is unique to his sample of students. A positive predictive power (PPV) of .87, for example, means that a student scoring below the cut score has an 87% chance of not meeting the fifth-grade mathematics standard at the end of the year. Importantly, once an individual is familiar with the probability nomogram, the simple process of identifying the PPV value requires less than one minute.

In summary, the following information should be provide when reporting ROC statistics: (a) a helpful table that presents ROC statistics associated with all possible screener cut scores and (b) a probability nomogram for translating the reported ROC statistics into predictive interpretations of students' mathematics performance. The fifth-grade NAEP best subset screeners reported in the present study provide demonstrations for researchers.

Reporting a Combination Matrix for Combined Screeners

Combined screeners for fourth grade and sixth grade are proposed in the present study; the method for reporting statistics associated with the combined screeners was modeled after a method used in a recent publication by Clemens and colleagues (2016). Using logistic regression, the researchers reported a predicted probabilities matrix that allows a researcher or practitioner to use a student's scores on both measures for predicting later mathematics achievement. For measures with continuous data, the matrix groups student scores into ranges. The present study grouped all scores on the fraction number line screener and presented the ranges vertically on the left side of the matrix. All scores on the NAEP screener were presented horizontally on the top row of the matrix. A practitioner can locate a students' fraction number line score on the left and match it with the student's NAEP score on the top of the matrix, revealing the range of predicted probabilities associated with the combination of these two scores. For example, a predicted probabilities range of 30%-35% indicates that the student has a 30%-35% chance of not meeting the mathematics standard at the end of the year. While this process requires fewer steps than the method for translating ROC statistics described previously, it also has disadvantages. Next, advantages and disadvantages of the two methods are discussed.

Comparing the Methods for Reporting Screener Statistics: Advantages and Disadvantages

Each method of reporting screener statistics and translating the statistics into usable information has one major disadvantage. The disadvantage of the combined matrix is that the predicted probabilities are specific to the current sample of students and do not allow for a practitioner to compute the probabilities for his/her own base rate. In contrast, the probability nomogram using ROC statistics allows a practitioner to compute statistics based on his/her own sample. Thus, the ROC statistics table and probability nomogram allow readers to make more accurate predictions of their own students' mathematics

achievement. Unfortunately, this method does not provide interpretable statistics when assessing a combination of two measures with different types of data (i.e., discrete and continuous data). When combining two measures, the probabilities matrix is the best available method for providing researchers and practitioners with empirically supported interpretations of students' scores (Clemens et al., 2016).

The disadvantage of the ROC statistics and probability nomogram is the dichotomous nature of the ROC curve analysis. ROC statistics provide information about cut scores that split students into two groups: students who score below the cut score and students who score above the cut score. Thus, for a cut score of five points, a student who scores one point will receive that same predictive interpretation as a student who scores four points (Clemens et al., 2016). In contrast, the probabilities matrix for combined screeners offers interpretations of scores that are more specific to each individual student.

Overall, data from the present study favor ROC curve statistics over the predicted probabilities matrix. The ROC curve analysis is considered the best method for assessing the diagnostic accuracy of a test (Weinstein et al., 2005) and even allows readers to compute ROC statistics that are specific to their samples (Cummings & Smolkowski, 2015; Youngstrom, 2014). However, ROC curve statistics are not interpretable for combined measures with different types of data; for combined measures, the present study recommends a probability matrix. In the present study, specific methods along with examples using screeners proposed in the study are provided. Educational researchers can follow the guidelines to help bridge the gap between research and classroom practice.

Limitations of the Present Study

Several limitations must be kept in mind when interpreting the results of the present study. First, there are limitations regarding the student sample. First, the generalizability of the results may be limited because students were recruited from two school districts in only one geographic location. Second, each initial ROC curve analysis

included a slightly different set of students from the total sample which may have impacted the results.

Another limitation is the small scope of measures assessed as potential screeners. Although the assessment of three fraction measures yielded informative findings about screening measures for the intermediate grades, there are several other potential measures that may also be predictive of mathematics achievement during these grades. For example, screeners that assess students' understanding of other mathematics topics might also emerge as strong predictors of later achievement. Furthermore, some of the measures (i.e., number line estimation in fourth and sixth grades) were administered in the winter of the school year rather than in the fall, which may have increased their accuracy for the prediction of the state mathematics test administered in the spring of each grade.

The continuous nature of the fraction number line estimation measure poses certain obstacles for practitioners. Computing percent absolute error (PAE) for every estimate on the number line is more challenging than scoring a multiple-choice NAEP item as correct or incorrect. In the present study, the number line measure was administered on the computer. PAE was calculated by dividing the absolute difference between the estimated and actual magnitudes by the numerical range of the number line (1 or 2), and then multiplying by 100 for each estimate. This method of scoring is far less practical for classroom application than a simple paper and pencil measure. Potential solutions to this limitation are discussed below in the Future Directions section.

Another concern of the present study is that certain fraction arithmetic items may have emerged as predictive items if assessed in the best subset analyses. Even though the original fraction arithmetic measures with all items included were outperformed by the fraction concepts measures, it is possible that one or two fraction arithmetic items may have been retained on the final best subset screener. Additionally, the fraction arithmetic items may have been more predictive if they had been better matched to the curriculum at each grade level assessed. Better calibration of a fraction arithmetic measure is needed.

The state mathematics test as the outcome variable may also raise some concerns when interpreting the results. The test was administered by the school district and the specific items included on the test at each grade are unknown. The proportion of fraction items on each test is also unknown. However, districts and teachers rely on these state standardized tests for making decisions in the classroom and for assessing students' understanding. Furthermore, previous research also relies on these state tests for predicting student achievement (e.g., Siegler et al., 2011). Future research should test the screeners with other state or national assessments as the outcome measure. For example, 15 states currently administer the Smarter Balanced assessments, and seven states administer the Partnership for Assessment of Readiness for College and Careers (PARCC) exams. Another possibility is using student performance on the NAEP mathematics assessment as the outcome measure. The NAEP is administered to students across the nation and thus is preferable to a state-administered exam. The NAEP assessment is administered in fourth, eighth, and twelfth grades.

Insufficient information regarding the specific type of instruction that students received in the classroom during the span of the study is another potential limitation. Reportedly, participating schools followed curriculum benchmarks aligned with the Common Core State Standards in Mathematics (NGACBP & CCSSO, 2010). However, the specific type of instruction that students received in the classroom that targeted fraction concepts and procedures is not known.

Educational Implications

The present study fills a gap in the literature by assessing fraction screeners for the intermediate grades. Measures of fractions concepts consistently emerged as highly predictive screening tools and can help practitioners identify students who are at risk for later mathematics difficulties. Using predictive screeners to identify at-risk students allows schools to make data-driven decisions, such as providing interventions to lower-

achieving students. Without valid mathematics screeners in the intermediate grades, struggling students may never receive the supports they need to reach their full mathematics potential. Such screeners may help identify students who otherwise may have fallen through the cracks in the educational system.

The findings address broader impacts for quality of life. Students who struggle with fractions may progress through the educational system and enter the workforce without foundational mathematics knowledge. Lack of basic mathematics skills impedes workplace success (McNamara, 2009). According to the recent Skills, Technology, and Management Practices (STAMP) survey, 94% of workers in the U.S. use basic mathematics skills on the job (Handel, 2016). Approximately two-thirds of the workers reported using fractions in their day-to-day workplace activities (Handel, 2016).

The importance of mathematics skills for success in today's workforce emphasizes the necessity of screening for mathematics difficulties in the intermediate grades. Students identified with fraction difficulties in fourth grade or later can be targeted for additional mathematics supports to bolster their mathematics understanding. Overall, improving mathematics screening tools for the identification of students at risk has both short-term impacts (e.g., increasing the likelihood of data-driven decisions in the classroom) and potential long-term, broader impacts (e.g., increasing students' preparedness for the workforce).

A promising finding of the present study is that fraction screeners have the power to predict students' mathematics achievement *years* later. In general, the goal of educational screeners is to identify children who are at risk as early as possible (Gersten et al., 2012); schools can then provide supports for these struggling students. Thus, the finding of the present study that screeners administered in fourth grade can predict sixth-grade performance has particular importance. Fourth grade may be a crucial grade for administering mathematics screeners and for identifying students who are likely to struggle with abstract topics such as fractions.

The results of the present study have potential implications for classroom instruction. The measures of fractions concepts were consistently more predictive of risk status than the current fraction arithmetic measure, which suggests that classroom instruction might give additional attention to students' conceptual understanding of fractions (although fraction arithmetic should not be ignored). For example, students would benefit from instruction that highlights learning about the relation between the numerator and denominator in multiple contexts (e.g., DeWolf et al., 2013). Importantly, previous research suggests that fraction conceptual knowledge can actually support students' arithmetic skill and vice versa (e.g., Hecht et al., 2003).

The present study also provides detailed recommendations and demonstrations for reporting mathematics screeners in the literature. In particular, educational researchers are urged to report screener statistics in ways that help researchers and practitioners translate the findings into usable information. By refining the ways in which researchers report screener statistics, the likelihood of helping practitioners make data-driven educational decisions increases (Smolkowski & Cummings, 2015).

Future Directions

Future research should continue to evaluate potential mathematics screeners for the intermediate grades. Although the present study emphasizes the importance of fraction concepts to other areas of the mathematics curriculum, further investigation is warranted. In particular, research should examine the predictability of other mathematics topics besides fractions and perhaps examine whether different topics could be combined to create an even more accurate screening tool. A related avenue for future research is the possibility of leveraging mathematics screeners *and* measures of domain general processes (e.g., working memory) for the prediction of later mathematics difficulties. For example, working memory is a predictor of general mathematics (e.g., Geary, 2011) and of fraction knowledge (Jordan et al., 2013). Thus, future research might explore the

predictive power of combination screeners that assess both mathematical competencies and domain general processes.

A modified version of the fraction number line measure might be explored in future research that improves its feasibility for use in the classroom. One possible solution is a computerized program or application that automatically computes students' percent absolute error for each item and mean percent absolute error for the total measure. A second option is a paper and pencil measure that requires students to answer multiple-choice number line items. If future research creates a valid and reliable fraction number line measure with multiple-choice items, the new measure would combine more easily with other multiple-choice measures.

Another related goal should be the design of applications for tablets and smartphones that help researchers or practitioners quickly and easily translate screener statistics. For example, a simple "app" could replace the probability nomogram for translating ROC statistics. A practitioner could enter values of the base rate and the diagnostic likelihood ratio into the app, and the technology could immediately provide the positive predictive power. Although using the probability nomogram is not a complex process, an app or computer program would be even less intimidating for new users. A major advantage of such an app or program is that it could be used for *any* single screener measure, regardless of the concepts being assessed or the grade level of the students. As long as the research provides the diagnostic likelihood ratios associated with all possible cut scores, then any reader could translate the statistics for his/her own sample.

Summary and Conclusions

The present study investigated three fraction measures (i.e., NAEP fraction concepts, fraction number line estimation, and fraction arithmetic) administered in fourth, fifth, and sixth grades for the prediction of later mathematics achievement at the end of each school year. The study revealed that the two fraction concepts measures (i.e., NAEP

fraction concepts and fraction number line estimation) were good screeners of risk status across the grades. In fourth and sixth grades, a combination of the two measures was assessed. In fifth grade, the NAEP measure emerged as especially predictive and was hence assessed alone as the best screener in that specific grade. To improve practicality for classroom use, the length of each screener was reduced by eliminating least predictive items; the final screeners are called "best subset" screeners.

Screener statistics were provided for each screener that allow researchers or practitioners to interpret students' scores. Readers of the present study can use the reported statistics to determine a student's chances of not meeting the end-of-the-year mathematics standard. Importantly, a probability nomogram is recommended for reporting ROC curve screener statistics; this nomogram is recommended in the medical literature but has not yet been introduced in the educational literature.

Overall, the present study fills an important gap in the literature by identifying and validating useful mathematics screeners in the intermediate grades. While the present study highlights the importance of fraction understanding for at-risk students, it simultaneously points to the importance of fraction knowledge for all children. The study also demonstrates methods for reporting screener statistics that allow researchers and practitioners to make accurate predictions of students' later mathematics achievement. The work represents a major step toward translating key research findings in ways that are interpretable to diverse audiences.

REFERENCES

American Institutes for Research (2012). DCAS 2011-2012 Technical Report. Retrieved from http://www.doe.k12.de.us/cms/lib09/DE01922744/Centricity/Domain/111/Vol1_Annual_TechRep.pdf

Bailey, D. H., Hansen, N., & Jordan, N. C. (2017). The codevelopment of children's fraction arithmetic skill and fraction magnitude understanding. *Journal of Educational Psychology*, *109*(4), 509. doi:10.1037/edu0000152

Bailey, D. H., Hoard, M. K., Nugent, L., & Geary, D. C. (2012). Competence with fractions predicts gains in mathematics achievement. *Journal of Experimental Child Psychology*, *113*, 447-455. doi:10.1016/j.jecp.2012.06.004

Ban, D., Tanabe, M., Ito, H., Otsuka, Y., Nitta, H., Abe, Y., ... & Kaneko, H. (2014). A novel difficulty scoring system for laparoscopic liver resection. *Journal of hepato-biliary-pancreatic sciences*, *21*(10), 745-753. doi:10.1002/jhbp.166

Booth, J. L., & Newton, C. J. (2012). Fractions: could they really be the gatekeeper's doorman? *Contemporary Educational Psychology*, *37*, 247-253. doi:10.1016/j.cedpsych.2012.07.001

Booth, J. L., Newton, K. J., & Twiss-Garrity, L. K. (2014). The impact of fraction magnitude knowledge on algebra performance and learning. *Journal of experimental child psychology*, *118*, 110-118. doi:10.1016/j.jecp.2013.09.001

Bossuyt, P. M., Reitsma, J. B., Bruns, D. E.,Gatsonis, C. A., Glasziou, P. P., Irwig, L. M., … de Vet, H. C. W. (2003). Towards complete and accurate reporting of studies of diagnostic accuracy: The STARD initiative. *British Medical Journal*, *326*, 41–44.

Bryant, D. P., Bryant, B. R., Gersten, R., Scammacca, N., & Chavez, M. M. (2008). Mathematics intervention for first-and second-grade students with mathematics difficulties: The effects of tier 2 intervention delivered as booster lessons. *Remedial and special education*, *29*(1), 20-32. doi:10.1177/0741932507309712

Byrnes, J. P., & Wasik, B. A. (1991). Role of conceptual knowledge in mathematical and procedural learning. *Developmental Psychology*, *27*, 777-786. doi:10.1037/0012-1649.27.5.777

Clarke, B., Baker, S., Smolkowski, K., & Chard, D. J. (2008). An analysis of early numeracy curriculum-based measurement: Examining the role of growth in student outcomes. *Remedial and Special Education*, *29*(1), 46-57. doi:10.1177/0741932507309694

Cobb, C. (2003). Effective instruction begins with purposeful assessments. *The Reading Teacher*, *57*(4), 386.

Cummings, K. D., & Smolkowski, K. (2015). Bridging the gap: Selecting students at risk of academic difficulties. *Assessment for Effective Intervention*, *41*(1). doi:10.1177/1534508415590396

DeWolf, M., Grounds, M. A., Bassok, M., & Holyoak, K. J. (2013). Magnitude comparison with different types of rational numbers. *Journal of Experimental Psychology: Human Perception and Performance,* 40, 53-72. doi:10.1037/a0032916

DeWolf, M., & Vosniadou, S. (2011). The whole number bias in fraction magnitude comparisons with adults. In *Proceedings of the 33rd annual conference of the cognitive science society* (pp. 1751-1756). Cognitive Science Society Austin, TX.

Duncan, G. J., Dowsett, C. J., Claessens, A., Magnuson, K., Huston, A. C., Klebanov, P., ... & Sexton, H. (2007). School readiness and later achievement. *Developmental psychology*, *43*(6), 1428. doi:10.1037/0012-1649.43.6.1428

Faulkner, M., Olson, M., Chandy, R., Krause, J., Chandy, K. M., & Krause, A. (2011, April). The next big one: Detecting earthquakes and other rare events from community-based sensors. In *Information Processing in Sensor Networks (IPSN), 2011 10th International Conference*. IEEE.

Fazio, L. K., Bailey, D. H., Thompson, C. A., & Siegler, R. S. (2014). Relations of different types of numerical magnitude representations to each other and to mathematics achievement. *Journal of Experimental Child Psychology*, *123*, 53-72. doi:10.1016/j.jecp.2014.01.013

Florkowski, C. M. (2008). Sensitivity, specificity, receiver-operating characteristic (ROC) curves and likelihood ratios: communicating the performance of diagnostic tests. *Clinical Biochemist Reviews*, *29*(1), S83-S87.

Fluss, R., Faraggi, D., & Reiser, B. (2005). Estimation of the Youden Index and its associated cutoff point. *Biometrical Journal*, *47*(4), 458-472.

Foegen, A., & Deno, S. L. (2001). Identifying growth indicators for low-achieving students in middle school mathematics. *The Journal of Special Education*, *35*(1), 4-16. doi:10.1177/002246690103500102

Foegen, A., Jiban, C., & Deno, S. (2007). Progress monitoring measures in mathematics: A review of the literature. *The Journal of Special Education, 41*(2), 121-139. doi:10.1177/00224669070410020101

Fuchs, L. S., Hamlett, C. L., & Fuchs, D. (1998). *Monitoring Basic Skills Progress—Basic math computation* (2nd ed., Blackline Masters). Austin, TX: Pro-Ed.

Fuchs, L. S., Hamlett, C. L., & Fuchs, D. (1999). *Monitoring Basic Skills Progress—Basic math concepts and applications* (2nd ed., Blackline Masters). Austin, TX: Pro-Ed.

Fuchs, L. S., Schumacher, R. F., Long, J., Namkung, J., Hamlett, C. L., Cirino, P. T., ... & Changas, P. (2013). Improving at-risk learners' understanding of fractions. *Journal of Educational Psychology*, *105*(3), 683. doi:10.1037/a0032446

Geary, D. C. (2004). Mathematics and learning disabilities. *Journal of Learning Disabilities*, *37*, 4-15. doi:10.1177/00222194040370010201

Geary, D. C. (2006). Development of mathematical understanding. In W. Damon & R. M. Lerner (Series Eds.) & D. Kuhn & R. S. Siegler (Vol. Eds.), Handbook of child psychology: Vol 2. Cognition, perception, and language (6th ed., pp. 777 – 810). New York: Wiley.

Geary, D. C. (2011). Cognitive predictors of achievement growth in mathematics: a 5-year longitudinal study. Developmental Psychology, 47(6), 1539. doi:10.1037/a0025510

Gelman, R., & Williams, E. (1998). Constraints on cognitive development and learning. In W. Damon (Series Ed.), D. Kuhn, & R. Siegler (Vol. Eds.), *Handbook of child psychology: Vol. 2. Cognition, language, and perception* (5th ed., pp. 575-630). New York: Wiley.

Gersten, R., Beckmann, S., Clarke, B., Foegen, A., Marsh, L., Star, J. R., & Witzel, B. (2009). Assisting students struggling with mathematics: Response to Intervention (RtI) for elementary and middle schools. NCEE 2009-4060. *What Works Clearinghouse*. Retrieved from http://files.eric.ed.gov/fulltext/ED504995.pdf

Gersten, R., Clarke, B. S., Haymond, K., & Jordan, N. C. (2011). Screening for mathematics difficulties in K-3 students. *Center on Instruction*. Retrieved from http://files.eric.ed.gov/fulltext/ED524577.pdf

Gersten, R., Clarke, B., Jordan, N. C., Newman-Gonchar, R., Haymond, K., & Wilkins, C. (2012). Universal screening in mathematics for the primary grades: Beginnings of a research base. *Exceptional Children*, *78*(4), 423-445. doi:10.1177/001440291207800403

Grove, W. M., Zald, D. H., Lebow, B. S., Snitz, B. E., & Nelson, C. (2000). Clinical versus mechanical prediction: a meta-analysis. *Psychological assessment*, *12*(1), 19. doi:10.1037//1040-3590.12.1.19

Hajian-Tilaki, K. (2013). Receiver operating characteristic (ROC) curve analysis for medical diagnostic test evaluation. *Caspian Journal of Internal Medicine*, *4*(2), 627.

Hallett, D., Nunes, T., & Bryant, P. (2010). Individual differences in conceptual and procedural knowledge when learning fractions. *Journal Of Educational Psychology*, *102*(2), 395-406. doi:10.1037/a0017486

Hallett, D., Nunes, T., Bryant, P., & Thorpe, C. M. (2012). Individual differences in conceptual and procedural fraction understanding: The role of abilities and school

experience. *Journal of Experimental Child Psychology, 113*(4), 469-486. doi:10.1016/j.jecp.2012.07.009

Handel, M. J. (2016). What do people do at work?. *Journal for Labour Market Research, 49*(2), 177-197. doi:10.1007/s12651-016-0213-1

Hanley, J. A., & McNeil, B. J. (1983). A method of comparing the areas under receiver operating characteristic curves derived from the same cases. *Radiology, 148*(3), 839-843. doi:10.1148/radiology.148.3.6878708

Hansen, N., Jordan, N. C., & Rodrigues, J. (in press). Identifying persistent learning difficulties in fractions: A longitudinal study of student growth from third through sixth grade. *Contemporary Educational Psychology.* doi:10.1016/j.cedpsych.2015.11.002

Hecht, S. (1998). Toward an information-processing account of individual differences in fraction skills. *Journal of Educational Psychology, 90*, 545-59. doi:10.1037//0022-0663.90.3.545

Hecht, S., Close, L., & Santisi, M. (2003). Sources of individual differences in fraction skills. *Journal of Experimental Child Psychology, 86*, 277-302. doi:10.1016/j.jecp.2003.08.003

Hecht, S. A., & Vagi, K. J. (2012). Patterns of strengths and weaknesses in children's knowledge about fractions. *Journal of Experimental Child Psychology, 111*, 212–29. doi:10.1016/j.jecp.2011.08.012

Hurst, M., & Cordes, S. (2017). A systematic investigation of the link between rational number processing and algebra ability. *British Journal of Psychology*. doi:10.1111/bjop.12244

IBM Corporation (2016). IBM SPSS Statistics for Windows, Version 24.0. Armonk, NY: IBM Corp.

Jeffries, H. E., Soto-Campos, G., Katch, A., Gall, C., Rice, T. B., & Wetzel, R. (2015). Pediatric index of cardiac surgical intensive care mortality risk score for pediatric cardiac critical care. *Pediatric Critical Care Medicine*, *16*(9), 846-852. doi:10.1097/PCC.0000000000000489

Jenkins, J. R., Hudson, R. F., & Johnson, E. S. (2007). Screening for at-risk readers in a response to intervention framework. *School Psychology Review*, *36*(4), 582.

Jiban, C. L., & Deno, S. L. (2007). Using math and reading curriculum-based measurements to predict state mathematics test performance: Are simple one-minute measures technically adequate?. *Assessment for Effective Intervention*, *32*(2), 78-89. doi:10.1177/15345084070320020501

Jordan, N. C., Glutting, J., & Ramineni, C. (2008). A number sense assessment tool for identifying children at risk for mathematical difficulties. In A. Dowker (Ed.), Mathematical difficulties: Psychology and intervention (pp.45–58). San Diego, CA: Academic Press.

Jordan, N. C., Glutting, J., Ramineni, C., & Watkins, M. W. (2010). Validating a number sense screening tool for use in kindergarten and first grade: Prediction of mathematics proficiency in third grade. *School Psychology Review, 39*(2), 181-195.

Jordan, N. C., Hansen, N., Fuchs, L. S., Siegler, R. S., Gersten, R., & Micklos, D. (2013). Developmental predictors of fraction concepts and procedures. *Journal of Experimental Child Psychology, 116*(1), 45-58. doi:10.1016/j.jecp.2013.02.001

Jordan, N. C., Kaplan, D., Ramineni, C., & Locuniak, M. N. (2009). Early math matters: Kindergarten number competence and later mathematics outcomes. *Developmental Psychology, 45*(3), 850. doi:10.1037/a0014939

Kadam, A. V., & Nimbalkar, U. M. (2015). Automatic assembly modeling for product variants using parametric modeling concept. *International Journal of Engineering Research and Technology, 4*(4), 79-89.

Keller-Margulis, M. A., Shapiro, E. S., & Hintze, J. M. (2008). Long-term diagnostic accuracy of curriculum-based measures in reading and mathematics. *School Psychology Review, 37*(3), 374.

Lembke, E., & Foegen, A. (2009). Identifying early numeracy indicators for kindergarten and first-grade students. *Learning Disabilities Research & Practice, 24*(1), 12-20. doi:10.1111/j.1540-5826.2008.01273.x

Libertus, M. E., Feigenson, L., & Halberda, J. (2011). Preschool acuity of the approximate number system correlates with school math ability. *Developmental Science, 14*(6), 1292-1300. doi:10.1111/j.1467-7687.2011.01080.x

Martínez-Camblor, P. (2013). Area under the ROC curve comparison in the presence of missing data. *Journal of the Korean Statistical Society, 42*(4), 431-442. doi:10.1016/j.jkss.2013.01.004

Mazzocco, M. M., Feigenson, L., & Halberda, J. (2011). Impaired acuity of the approximate number system underlies mathematical learning disability (dyscalculia). *Child Development, 82*(4), 1224-1237. doi:10.1111/j.1467-8624.2011.01608.x

McFall, R. M., & Treat, T. A. (1999). Quantifying the information value of clinical assessments with signal detection theory. *Annual Review of Psychology, 50,* 215–241. doi:10.1146/annurev.psych.50.1.215

McNamara, B. R. (2009). The skill gap: will the future workplace become an abyss. *Techniques: Connecting Education and Careers, 84*(5), 24-27.

Medcalc Statistical Software (2016). MedCalc Statistical Software version 16.4.3. Ostend, Belgium. Retrieved from https://www.medcalc.org

Methe, S. A., Hintze, J. M., & Floyd, R. G. (2008). Validation and decision accuracy of early numeracy skill indicators. *School Psychology Review, 37*(3), 359.

Meyers, L. S., Gamst, G., & Guarino, A. J. (2006). *Applied multivariate research: Design and interpretation*. NY: Sage.

Meyers, L. S., Gamst, G. C., & Guarino, A. J. (2013). Performing data analysis using IBM SPSS. NY: John Wiley & Sons.

Morgan, P. L., Farkas, G., & Wu, Q. (2009). Five-year growth trajectories of kindergarten children with learning difficulties in mathematics. *Journal of Learning Disabilities*, *42*(4), 306-321. doi:10.1177/0022219408331037

Muthén, L. K., & Muthén, B. O. (2012). *Mplus User's Guide.* Los Angeles, CA: Muthén & Muthén.

National Governors Association Center for Best Practices & Council of Chief State School Officers. (2010). *Common Core State Standards for Mathematics*. Washington DC: Author.

National Mathematics Advisory Panel (NMAP) (2008). *Foundations for success: The final report of the National Mathematics Advisory Panel.* Washington, DC: U.S. Department of Education.

Newton, K. J., Willard, C., & Teufel, C. (2014). An examination of the ways that students with learning disabilities solve fraction computation problems. *The Elementary School Journal, 115,* 1-21. doi:10.1086/676949

Ni, Y., & Zhou, Y. D. (2005). Teaching and learning fraction and rational numbers: The origins and implications of whole number bias. *Educational Psychologist, 40,* 27–52. doi:10.1207/s15326985ep4001_3

Peck, D. M., & Jencks, S. M. (1981). Conceptual issues in the teaching and learning of fractions. *Journal for Research in Mathematics Education*, 339-348. doi:10.2307/748834

Pepe, M. S. (2003). *The statistical evaluation of medical tests for classification and prediction*. Oxford University Press, USA.

Purpura, D. J., Reid, E. E., Eiland, M. D., & Baroody, A. J. (2015). Using a brief preschool early numeracy skills screener to identify young children with mathematics difficulties. *School Psychology Review*, *44*(1), 41-59. doi:10.17105/SPR44-1.41-59

Resnick, I., Jordan, N. C., Hansen, N., Rajan, V., Rodrigues, J., Siegler, R. S., & Fuchs, L. S. (2016). Developmental growth trajectories in understanding of fraction magnitude from fourth through sixth grade. *Developmental Psychology*, *52*(5), 746-757. doi:10.1037/dev0000102

Rittle-Johnson, B., Siegler, R. S., & Alibali, M. W. (2001). Developing conceptual understanding and procedural skill in mathematics: an iterative process. *Journal of Educational Psychology, 93,* 346–362. doi:10.1037/0022-0663.93.2.346

Rodrigues, J., Dyson, N., Hansen, N., & Jordan, N. C. (2016). Preparing for algebra by building fraction sense. *Teaching Exceptional Children, 49*(2), 134-141.

Sadler, P. M., & Tai, R. H. (2007). The two high-school pillars supporting college science. *Science. 317*, 457-458. doi:10.1126/science.1144214

Seethaler, P. M., & Fuchs, L. S. (2010). The predictive utility of kindergarten screening for math difficulty. *Exceptional Children, 77*(1), 37-59. doi:10.1177/001440291007700102

Seethaler, P. M., Fuchs, L. S., Star, J. R., & Bryant, J. (2011). The cognitive predictors of computational skill with whole versus rational numbers: An exploratory study. *Learning and Individual Differences, 21*(5), 536-542. doi:10.1016/j.lindif.2011.05.002

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*(1), 19-35. doi:10.1177/0734282905285237

Siegler, R. S., Duncan, G. J., Davis-Kean, P. E., Duckworth, K., Claessens, A., Engel, M., Susperreguy, M. I., & Chen, M. (2012). Early predictors of high school mathematics achievement. *Psychological Science, 23*(7), 691-697. doi:10.1177/0956797612440101

Siegler, R. S., Fazio, L. K., Bailey, D. H., & Zhou, X. (2013). Fractions: The new frontier for theories of numerical development. *Trends in Cognitive Science, 17*, 13-19. doi:10.1016/j.tics.2012.11.004

Siegler, R. S., & Lortie-Forgues, H. (2014). An integrative theory of numerical development. *Child Development Perspectives, 8*(3), 144-150. doi:10.1111/cdep.12077

Siegler, R. S., Thompson, C. A., & Schneider, M. (2011). An integrated theory of whole number and fractions development. *Cognitive Psychology, 62*, 273-296. doi:10.1016/j.cogpsych.2011.03.001

Smolkowski, K., & Cummings, K. D. (2015). Evaluation of diagnostic systems: The selection of students at risk of academic difficulties. *Assessment for Effective Intervention, 41*(1), 41-54. doi:10.1177/1534508415590386

Speece, D. L., Schatschneider, C., Silverman, R., Case, L. P., Cooper, D. H., & Jacobs, D. M. (2011). Identification of reading problems in first grade within a response-to-intervention framework. *The Elementary School Journal, 111*(4), 585-607. doi:10.1086/659032

Swets, J. A., Dawes, R. M., & Monahan, J. (2000). Psychological science can improve diagnostic decisions. *Psychological Science in the Public Interest, 1*, 1–26.

Tai, R. H., Sadler, P. M., & Mintzes, J. J. (2006). Factors influencing college science success. *Journal of College Science Teaching, 36*(1), 52.

Torbeyns, J., Schneider, M., Xin, Z., & Siegler, R. S. (2015). Bridging the gap: Fraction understanding is central to mathematics achievement in students from three different continents. *Learning and Instruction, 37*, 5-13. doi:10.1016/j.learninstruc.2014.03.002

U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP). (1990-2009). Mathematics assessment. Retrieved from http://nces.ed.gov/nationsreportcard

VanDerHeyden, A. M., Broussard, C., Snyder, P., George, J., Lafleur, S. M., & Williams, C. (2011). Measurement of kindergartners' understanding of early mathematical concepts. *School Psychology Review, 40*(2), 296.

Van Hoof, J., Janssen, R., Verschaffel, L., & Van Dooren, W. (2015). Inhibiting natural knowledge in fourth graders: towards a comprehensive test instrument. *ZDM, 47*(5), 849-857. doi:10.1007/s11858-014-0650-7

Vosniadou, S., Vamvakoussi, X., & Skopeliti, I. (2008). The framework theory approach to the problem of conceptual change. In S. Vosniadou (Ed.), *International handbook of research on conceptual change* (pp. 3–34). New York, NY: Routledge.

Vukovich, R. K., Fuchs, L. S., Geary, D. S., Jordan, N. C., Gersten, R., & Siegler, R. S. (2014). Sources of individual differences in children's understanding of fractions. *Child Development, 85*(4), 1461-1476. doi:10.1111/cdev.12218

Weinstein, S., Obuchowski, N. A., & Lieber, M. L. (2005). Clinical evaluation of diagnostic tests. *American Journal of Roentgenology, 184,* 14–19. doi:10.2214/ajr.184.1.01840014

Wilkinson, G. S., & Robertson, G. J. (2006). Wide Range Achievement Test 4 professional manual. Lutz, FL: Psychological Assessment Resources.

Wilson, J., Olinghouse, N. G., McCoach, D. B., Santangelo, T., & Andrada, G. N. (2016). Comparing the accuracy of different scoring methods for identifying sixth graders at risk of failing a state writing assessment. *Assessing Writing, 27,* 11-23. doi:10.1016/j.asw.2015.06.003

Wittenberg, L., Economopoulos, L., Bastable, V., Bloomfield, K. H., Cochran,K., Earnet, D., Hollister, A., Horowitz, N., Leidl, E., Murrayr, M., Oh, Y., Perrfy, B. W., Russell, S. J., Schifter, D., & Sillman, K. (2012). *Investigations in Number, Data, and Space (Grade 3)*. Illinois: Pearson Education.

Woodcock, R. W. (1988). *Woodcock Reading Mastery Tests—Revised.* Circle Pines, MN: American Guidance Service.

Yang, H. (2013). The case for being automatic: Introducing the automatic linear modeling (LINEAR) procedure in SPSS statistics. *Multiple Linear Regression Viewpoints, 39*(2), 27-37.

Ye, A., Hansen, N., Resnick, I., Rodrigues, J., Rinne, L., & Jordan, N. C. (2016). Pathways to fraction learning: numerical abilities mediate the relation between early cognitive competencies and later fraction knowledge. *Journal of Experimental Child Psychology*, *152*, 242-263. doi:10.1016/j.jecp.2016.08.001

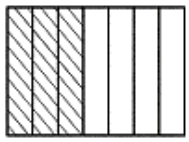
Youngstrom, E. A. (2014). A primer on receiver operating characteristic analysis and diagnostic efficiency statistics for pediatric psychology: We are ready to ROC. *Journal of Pediatric Psychology*, *39*(2), 204-221. doi:10.1093/jpepsy/jst062

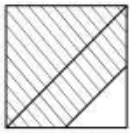
Appendix A

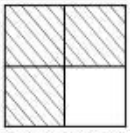
NAEP FRACTION CONCEPTS MEASURE

Item	***Grade(s)***

1. Which shows $\frac{3}{4}$ of the picture shaded? *4, 5, 6*

A. 

B. 

C. 

D. 

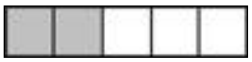
2. What fraction of the group of umbrellas is closed? *4, 5, 6*

A. $\frac{1}{3}$

B. $\frac{3}{7}$

C. $\frac{4}{7}$

D. $\frac{3}{4}$



3. What fraction of the figure above is shaded? *4, 5, 6*

Answer: _____

4-5. These three fractions are equivalent. Write <u>two</u> more fractions that are equivalent to these. *4, 5, 6*

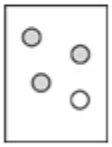
Answer: _____ , _____

6. Which picture shows that $\frac{3}{4}$ is the same as $\frac{6}{8}$? *4, 5, 6*

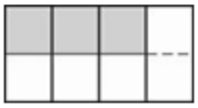
A.

B.

C.



D.



7. Luis had two apples and he cut each apple into fifths. How many pieces of apple did he have? *4, 5, 6*

A. $\frac{2}{5}$

B. 2

C. 5

D. 10

8. $\frac{4}{6} - \frac{1}{6} =$ *4, 5, 6*

A. 3

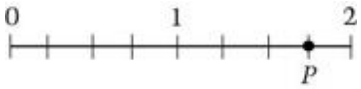
B. $\frac{3}{6}$

C. $\frac{3}{0}$

D. $\frac{5}{6}$

9. How many fourths make a whole? *4, 5, 6*

Answer: _____



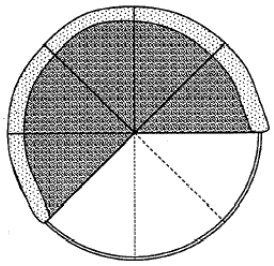
10. On the number line above, what number does <u>P</u> represent? *4, 5, 6*

A. $\frac{2}{3}$

B. $\frac{3}{4}$

C. $1\frac{2}{3}$

D. $1\frac{3}{4}$



11. The figure above shows that part of a pizza has been eaten. What part of the pizza is still there? *4, 5, 6*

A. $\frac{3}{8}$

B. $\frac{3}{5}$

C. $\frac{5}{8}$

D. $\frac{5}{3}$

4, 5, 6

12. On the portion of the number line above, a dot shows where 1/2 is. Use another dot to show where 3/4 is.

4, 5, 6

13. Students in Mrs. Johnson's class were asked to tell why $\frac{4}{5}$ is greater than $\frac{2}{3}$. Whose reason is best?

A. Kelly said, "Because 4 is greater than 2."

B. Keri said, "Because 5 is larger than 3."

C. Kim said, "Because $\frac{4}{5}$ is closer than $\frac{2}{3}$ to 1."

D. Kevin said, "Because 4 + 5 is more than 2 + 3."

14. Which fraction has a value closest to $\frac{1}{2}$? *4, 5, 6*

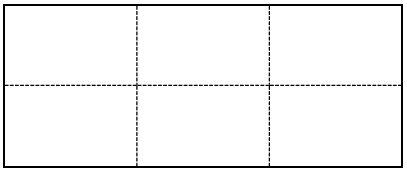
A. $\frac{5}{8}$

B. $\frac{1}{6}$

C. $\frac{2}{2}$

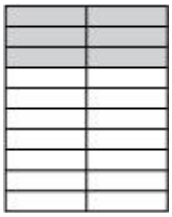
D. $\frac{1}{5}$

15. *4, 5, 6*



Shade $\frac{1}{3}$ of the rectangle above.

4, 5, 6



16. What fraction of the figure above is shaded?

A. $\frac{1}{4}$

B. $\frac{3}{10}$

C. $\frac{1}{3}$

D. $\frac{3}{7}$

Mark says $\frac{1}{4}$ of his candy bar is smaller than $\frac{1}{5}$ of the same candy bar. Is Mark right?

4, 5, 6

17.

A. Yes, Mark is right.

B. No, Mark is NOT right.

Draw a picture or use words to explain why you think Mark is right or wrong.

18. In which of the following are the three fractions arranged from least to greatest?

4, 5, 6

A. $\frac{2}{7}, \frac{1}{2}, \frac{5}{9}$

B. $\frac{1}{2}, \frac{2}{7}, \frac{5}{9}$

C. $\frac{5}{9}, \frac{1}{2}, \frac{2}{7}$

D. $\frac{5}{9}, \frac{2}{7}, \frac{1}{2}$

5, 6

19. A recipe requires $1\frac{1}{3}$ cups of sugar. Which of the following ways describes how the measuring cups shown can be used to measure $1\frac{1}{3}$ cups of sugar accurately?

A. Use the $\frac{1}{2}$ cup three times.

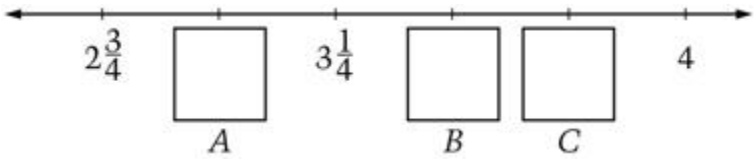
B. Use the $\frac{1}{4}$ cup three times.

C. Use the $\frac{1}{2}$ cup twice and the $\frac{1}{3}$ cup once.

D. Use the $\frac{1}{3}$ cup twice and the $\frac{1}{2}$ cup once.

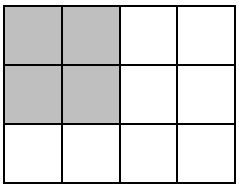
E. Use the $\frac{1}{4}$ cup once, the $\frac{1}{3}$ cup once, and the $\frac{1}{2}$ cup once.

6



20-22. Jorge left some numbers off the number line above. Fill in the numbers that should go in *A, B,* and *C.*

6



23. In the figure above, what fraction of rectangle *ABCD* is shaded?

6

24. If $\frac{2}{25} = \frac{n}{500}$, then n =

A. 10
B. 20
C. 30
D. 40
E. 50

Appendix B

FRACTION ARITHMETIC ITEMS

	4th F	5th F	6th W
3/6 + 1/6 =	X	X	X
2/5 + 1/5 =	X	X	X
3/4 + 2/4 =	X	X	X
3 3/8 + 1 2/8 =	X	X	X
3/4 - 1/4 =	X	X	X
5/6 - 2/6 =	X	X	X
1 3/4 - 1/4 =	X	X	X
2 2/3 - 1 1/3 =	X	X	X
5/6 + 2/3 =		X	X
7/8 -1/2 =		X	X
1 1/3 – 4/5 =			X
3/4 + 2/3 =			X
3 x 1/3 =			X
40 x 1/2 =			X
4 x 4/5 =			X
6 x 3/4 =			X
7/8 x 2/5 =			X
5/6 x 3/4 =			X
2 2/3 x 1/2 =			X
1 3/8 x 2/3 =			X
2 1/3 x 3 3/8 =			X
1/3 ÷ 4 =			X
1/6 ÷ 3 =			X
2 ÷ 3/4 =			X
7 ÷ 1/2 =			X
3/4 ÷ 1/8 =			X

Note. F = Fall, W = Winter.

Appendix C

GLOSSARY OF KEY TERMS

		Outcome Result	
		Positive (*Did not meet mathematics standard*)	**Negative** (*Met the mathematics standard*)
Screener Result	**Positive** (*"At risk on screener; ≤ cut score*)	True Positive	False Positive
	Negative (*Not "at risk" on screener; > cut score*)	False Negative	True Negative

Base Rate (ρ), also known as prevalence rate and prior probability: The proportion of students in the sample who did not meet the standards; calculated by dividing the amount of students who did not meet the standards by the total amount of students included in the analysis.

Area under the curve (AUC): The most commonly used global index of diagnostic accuracy (Fluss et al., 2005). The AUC represents the probability of a certain predictor measure distinguishing between students who are likely to meet the state mathematics standards versus students who are not likely to meet the standards.

Diagnostic accuracy, as determined by the AUC: A predictor measure's ability to accurately predict student membership into one of two groups: students who are likely to meet the state mathematics standards versus students who are not likely to meet the standards.

True Positive Fraction (TPF; also known as sensitivity): Among the students who did not meet the mathematics standards, the proportion who scored below the predictor measure cut score. TPF signifies the accuracy of the predictor measure among the students who did not meet the standards. TPF is not dependent on base rate.

TPF = TP / (TP + FN)

True Negative Fraction (TNF; also known as the specificity): Among the students who met the mathematics standards, the proportion who scored above the predictor measure cut score. TNF signifies the accuracy of the predictor measure among the students who did meet the standards. TNF is not dependent on base rate.
TNF = TN / (TN + FP)

False positive fraction (FPF): Rate of students who meet the standards but are incorrectly identified by a positive result on the predictor measure. A more colloquial term for FPF is the "false alarm rate."
FPF = 1 – TNF

False negative fraction (FNF): Rate of students who do not meet the standards but are incorrectly identified by a negative result on the predictor measure.
FNF = 1 – TPF

Positive Predictive Power (PPV): The percentage of students who fall below a certain predictor cut score who do not meet the standards. The PPV is influenced by the base rate/the prior probability and thus does not easily transfer to other samples. The PPV signifies the posttest probability of not meeting the standards. That is, the PPV offers the following interpretation for a student who scores below the cut score: "The student has a __% chance of not meeting the mathematics standards at the end of the year."
PPV = ρTPF / (ρTPF + (1 − ρ)FPF)

Negative Predictive Power (NPV): The percentage of students who fall above a certain NAEP cut score who truly do meet the standards. The NPV is influenced by the base rate/the prior probability and thus does not easily transfer to other samples. The NPV offers the following interpretation for a student who scores above the cut score: "The student has a __% chance of meeting the mathematics standards at the end of the year."
NPV = (1 − ρ)(TNF) / ((1 − ρ)(TNF) + ρ(1 − TPF))

Diagnostic likelihood ratio for a positive predictor measure result (DLR+): The odds that a predictor measure score less than a cut score will correctly identify a student who does not meet the standards. The DLR+ is independent of a sample's base rate and therefore can generalize to other samples of students (Pepe, 2003).
DLR+ = TPF / FPF

Diagnostic likelihood ratio for a negative predictor measure result (DLR-): The odds that a predictor measure score greater than a cut score will correctly identify a student who meets the standards. The DLR- is independent of a sample's base rate and therefore can generalize to other samples of students (Pepe, 2003).
DLR- = FNF / TNF

Appendix D

IRB APPROVAL

RESEARCH OFFICE

210 Hullihen Hall
University of Delaware
Newark, Delaware 19716-1551
PH: 302/831-2136
Fax: 302/831-2828

DATE: August 10, 2010

TO: Nancy Jordan
FROM: University of Delaware IRB

STUDY TITLE: [183637-1] Improving Students' Understanding of Fractions

SUBMISSION TYPE: New Project

ACTION: APPROVED
APPROVAL DATE: August 10, 2010
EXPIRATION DATE: August 10, 2011
REVIEW TYPE: Expedited Review

REVIEW CATEGORY: Expedited review category #7

Thank you for your submission of New Project materials for this research study. The University of Delaware IRB has APPROVED your submission. This approval is based on an appropriate risk/benefit ratio and a study design wherein the risks have been minimized. All research must be conducted in accordance with this approved submission.

This submission has received Expedited Review based on the applicable federal regulation.

Please remember that informed consent is a process beginning with a description of the study and insurance of participant understanding followed by a signed consent form. Informed consent must continue throughout the study via a dialogue between the researcher and research participant. Federal regulations require each participant receive a copy of the signed consent document.

Please note that any revision to previously approved materials must be approved by this office prior to initiation. Please use the appropriate revision forms for this procedure.

All SERIOUS and UNEXPECTED adverse events must be reported to this office. Please use the appropriate adverse event forms for this procedure. All sponsor reporting requirements should also be followed.

Please report all NON-COMPLIANCE issues or COMPLAINTS regarding this study to this office.

Please note that all research records must be retained for a minimum of three years.

Based on the risks, this project requires Continuing Review by this office on an annual basis. Please use the appropriate renewal forms for this procedure.