INVESTIGATION INTO THE VALIDITY OF THE QUANTITY DISCRIMINATION CURRICULUM-BASED MEASURE OF EARLY NUMERACY

By

MICHAEL HOFFMAN

Bachelor of Science in Psychology
Oklahoma State University
Stillwater, Oklahoma
2006

Master of Science in Educational Psychology
Oklahoma State University
Stillwater, Oklahoma
2008

Submitted to the Faculty of the
Graduate College of the
Oklahoma State University
in partial fulfillment of
the requirements for
the Degree of
DOCTOR OF PHILOSOPHY
May, 2014

INVESTIGATION INTO THE VALDITIY OF THE QUANTITY

DISCRIMINANTION CURRICULUM-BASED MEASURE OF EARLY NUMERACY

Dissertation Approved:

Gary Duhon, PhD

Dissertation Adviser

Brian Poncy, PhD

Terry Stinnett, PhD

Dale Fuqua, PhD

Name: MICHAEL HOFFMAN, M.S.

Date of Degree: MAY, 2014

Title of Study: INVESTIGATION INTO THE VALIDITY OF THE QUANTITY DISCRIMINATION CURRICULUM-BASED MEASURE OF EARLY NUMERACY

Major Field: EDUCATIONAL PSYCHOLOGY

Abstract: Children come to school with a number of informal math skills (Gersten et. al., 2005; National Mathematics Advisory Panel, 2008). These informal skills include counting, ability to identify numbers, ability to formulate mental number lines, and ability to discriminate between quantities (Clements, Sarama, & DiBiase, 2004). Tests of Early Numeracy (TENs), Quantity Discrimination (QD) being one, are robust indicators that are based on numeracy skills and can be used to help educators identify students potentially at-risk for later failure in formal mathematical computation (Clarke & Shinn, 2002). There is no current single TEN that accounts for enough of the variance in informal math that can be used as an independent screening measure for early numeracy. This study investigated changing probe construction for the QD measure from finding the larger of two numbers to finding the middle number of a three item set to improve the concurrent validity with the Number Knowledge Test (NKT), so it possibly could be used as an independent measure for early numeracy. If a broader range of informal math skills could be investigated within one measure, as was suggested in this study by changing probe construction, then it is possible that the validity of that measure could be increased. Changing the probe construction from finding the larger of two numbers to finding the middle number of a three number set did not improve the validity of the original QD measure in this study. Whether administration time and scoring method could impact the validity was questioned as well. Neither changing the administration time from one minute to two or three minutes, nor changing the scoring of the QD measure from a purely fluency based measure to an outcome measure that takes into account fluency while weighting the score for accuracy improved the validity of the original QD measure. Theoretical, research, and applied implications regarding the findings of this study, as well as how the study findings could possibly affect future directions for early numeracy CBM screening and CBM screening in general are discussed.

TABLE OF CONTENTS

Chapter Page

LIST OF TABLES

Table Page

CHAPTER I

INTRODUCTION

In order to provide additional instruction in math for students who need it, a valid process to identify these students needs to be identified and followed. One way to approach student identification is by the use of universal screening (Clarke & Shinn, 2002). A number of curriculum-based measures have been experimentally designed and researched for this purpose (Chard et al., 2005; Clarke & Shinn, 2004; VanDerHeyden, Witt, Naquin, & Noell, 2001). Early universal assessment is an important component when identifying students who need additional instruction in mathematics. The 'Matthew Effect Phenomenon' states that students with early math skills continue to prosper over the course of their education, while children who struggle at kindergarten entry tend to experience great degrees of problems in math (Methe & Riley-Tillman, 2008). Differences in children who do acquire a clear knowledge base of core competencies in early math and children who do not include early informal experiences with number, strategies used to store and access the knowledge, and fluency with numeric combinations (Gersten, Jordan, & Flojo, 2005; National Mathematics Advisory Panel,

2008). Early mathematical assessment is warranted as early as kindergarten, so early mathematical intervention can be implemented for those students who need it.

Tests of Early Numeracy (TENs)

There appears to be a low degree of consensus about the best approach to take when developing Curriculum-Based Measures (CBM) for mathematics (Foegen, Jiban, and Deno, 2007). Foegen et al. (2007) summarized two different approaches used for developing mathematical CBMs. One approach for developing mathematical CBMs is termed as the curriculum sampling approach (Mathematical Curriculum-Based Measurement; M-CBM), which involves sampling curriculum used in the classroom for that grade level and including those problems on the CBM probes. The second approach for developing mathematical CBMs is termed as the robust indicators approach, which involves identifying measures that represent broadly defined proficiencies in mathematics.

Educators use Mathematical Curriculum-Based Measurement (M-CBM) to monitor student progress in mathematical computation after they are doing formal mathematics; however, mathematical computation is generally not taught until the middle to end of first grade (Clarke & Shinn, 2002). Since M-CBM assessment cannot be used as an assessment tool before mathematical computation is taught, educators need tools to identify at-risk students before problems occur. At-risk students are students who could be failing to acquire the early mathematical knowledge (numeracy skills) needed to acquire mathematical computation understanding (Gersten et al., 2005; National Mathematics Advisory Panel, 2008). Robust indicators can be used to identify these at-risk students (Foegen et al., 2007). Tests of Early Numeracy (TENs) are robust indicators that are based on numeracy skills and

can be used to help educators identify students potentially at-risk for later failure in formal mathematical computation (Clarke & Shinn, 2002).

What are TENs and what do they measure? Children come to school with a number of informal math skills (Gersten et. al., 2005; National Mathematics Advisory Panel, 2008). These informal skills include counting, the ability to identify numbers, the ability to formulate mental number lines, and the ability to discriminate between quantities (Clements, Sarama, & DiBiase, 2004). Measures of early mathematics (TENs) can be based upon these early informal math skills (Clarke & Shinn, 2002). Measurement examples include Oral Counting (OC), Number Identification (NI) the ability to identify numbers in print, Missing Number (MN) the ability to formulate a mental number line and identify a missing number in a sequence of three numbers (i.e. 4__5), and Quantity Discrimination (QD) the ability to discriminate between two quantities in which the larger of two numbers is generally identified (Clarke & Shinn, 2002; 2004).

The Development of TENs

Clarke and Shinn (2004) investigated the sensitivity, reliability, and validity of four early mathematics measures (OC, NI, MN, and QD) that were experimental and designed to be used in early identification and formative evaluation. Participants consisted of 52 first grade students in the Pacific Northwest. Concurrent and predictive validity was examined using three criterion measures.; M-CBM (required students to answer math problems that were drawn from the students' first-grade mathematics curriculum and were two minutes in duration), WJ-R applied problems subtest (requires the participants to analyze and solve practical problems in mathematics by deciding on the appropriate mathematical operations to

use; Woodcock & Johnson, 1989), and the *Number Knowledge Test* (developmental test designed to measure the intuitive knowledge of number that the average child has available at the age-levels of 4, 6, 8, and 10 years; Okamoto & Case, 1996). Each of the four experimental measures (OC, NI, MN, and QD) had sufficient evidence for sensitivity, reliability, and validity. No measure had poor reliability, and the QD measure (numbers 0-20) was the most reliable measure. Support was strongest for the use of the QD measure (numbers 0-20) as a single indicator of early mathematics.

The OC, MN, QD, and NI measures investigated by Clarke and Shinn (2004) are a part of the AIMSweb Tests of Early Numeracy (TEN; Clarke & Shinn, 2002) mathematical probe set for kindergarten and first grade. AIMSweb is a curriculum-based measurement tool that allows educators and school psychologists to screen and progress monitor students who are demonstrated to have difficulties in reading, writing, and mathematics when compared to national and local norms. Can the QD measure be improved to a degree where it can be used as a single measure of early numeracy? In order to answer this question, the technical adequacy of the current QD measure needs to be considered.

Technical Adequacy of the QD Measure

The concurrent validity for the QD measure has been evaluated for use with first grade and kindergarten students (Chard et al., 2005; Clarke & Shinn, 2004; Lembke, Foegen, Whittaker, & Hampton, 2008). Chard et al. (2005) evaluated kindergarten (numbers 1-10) and first grade (numbers 1-20) samples on three of the four mathematical probes (including the QD measure) that were developed by Clarke and Shinn (2004). The correlation between the *Number Knowledge Test* (Okamoto & Case, 1996) and QD for kindergarten students was

.55 in the fall and .50 in the spring, and for first grade students it was .45 in the fall and .53 in the spring (Chard et al., 2005). Lembke et al. (2008) evaluated the technical adequacy of the NI, MN, and QD (numbers 0-10 for kindergarten & numbers 0-20 for first grade) measures for monitoring progress in early numeracy among kindergarten and first grade students. The *Stanford Early School Achievement Test* (SESAT; Psychological Corporation, 1996) was used to assess criterion validity and was administered to the first grade sample (N=29) in the fall prior to the Lembke et al. (2008) study. Concurrent criterion validity between the SESAT and the QD measure was statistically significant ($r = .50$; $p \leq .01$; Lembke et al., 2008).

The Quantity Discrimination (QD) Measure

Currently, the QD measure requires students to name the larger of two visually presented numbers (numbers 0-10 for kindergarten and numbers 0-20 for first grade), and the measure is scored as the number of larger numbers correctly identified in one minute (Chard et al., 2005; Clarke & Shinn, 2004; Lembke et al., 2008). The numeracy skill assessed with the QD measure is comparing and ordering. With comparing and ordering quantities and sets are compared using nonverbal identification, progressing to verbal labels such as more, equal, and less (Methe & Tillman, 2008; National Council of Teachers of Mathematics, 2006). The ability to recognize and produce ordinal numbers begins with 1-5 and progresses through 30 by age 6 (Methe & Tillman, 2008). Comparing and ordering is a prerequisite for establishing the idea of movement and change, and requires the envisioning of a mental number line (Methe & Tillman, 2008).

The QD measure was initially developed as a measure of whether or not a student could discriminate the larger of two numbers (Clarke & Shinn, 2002; 2004). Being able to identify the larger of two numbers is a skill students need to identify which number is placed first (the larger number) in a computational problem. Whether identifying the larger of two numbers is the best procedure for conducting a quantity discrimination problem has not currently been investigated. Could requiring a student to identify the middle number of a three number set improve the validity of the QD measure? If so, this would decrease the chance of obtaining a correct answer by a function of guessing, while increasing assessment integrity.

QD Probe Durations

The current QD measure uses a probe duration time of one-minute (Chard et al., 2005; Clarke & Shinn, 2004; Lembke et al. 2008). Is the standard one-minute administration time sufficient for the QD measure, or are longer administration times significantly more reliable? There is no known research study that has investigated whether the QD measure's reliability can be significantly increased by increasing the administration time.

Although no known research has investigated whether increasing administration time is necessary to increase the reliability of the QD measure, evaluating the effect of measuring time as an outcome variable for the universal screening of reading has been investigated (Williams et al., 2011). The most common Reading Curriculum-Based Measure (R-CBM) monitors Words Read Correct per Minute (WRCPM; reading aloud; reading fluency). Williams et al. (2011) recorded the time it took for a student to read a 400 word passage, compared that outcome to the WRCPM outcome, and determined WRCPM is an indirect

measure of reading speed. Reading speed was found to account for the greatest variance in most reading measures (Williams et al., 2011). Like reading, it is possible that the speed of completing the QD measure could account for the most variance in discriminating quantity. Evaluating whether the standard one-minute administration time for the QD measure is sufficient (or longer administration times are more reliable) will help ensure that the most variance possible is being accounted for regarding administration time. Increased administration duration could potentially improve the validity of the measure by reducing floor and ceiling effects.

Researchers have investigated probe administration time regarding other M-CBM measures. Jiban and Deno (2007) investigated the influence of aggregating administration scores (thus increasing and investigating administration time) in a study investigating if cloze math facts were as equally reliable as basic math facts when the measures were used to predict performance on a criterion. The scores on the cloze math measure were evaluated individually and aggregated. Aggregating the scores of two administrations on the one-minute cloze math probes, which increased administration time, appeared to be a potential method to improve reliability of the scores (from .65 to .86 for a fifth grade sample). This aggregating procedure will not be used in this current study; however, administration time will be evaluated at one, two, and three minutes (of a three minute administration period) to investigate the effect of administration time of the QD measure on concurrent validity, which is not known to have been investigated to this date.

QD Outcome Scoring

Would changing the scoring of the QD measure from a purely fluency based measure (Digits Correct per Minute; DCM) to an outcome measure that weights the fluency score (DCM) for accuracy improve the validity and sensitivity of the QD measure? If so, this would help rank order students better (purpose of screening), and it would add evaluating accuracy along with fluency into one score (which could help better evaluate a student's true performance). To date, no known research has investigated the effectiveness of using this scoring procedure to potentially improve the validity of rank ordering students on QD.

Study Rational and Research Questions

The purpose of the current study is to investigate and improve the validity of the QD curriculum-based measure (determining the most effective administration and scoring procedures). This will assist educators in making the most valid assessment conclusions, while accomplishing this quickly with as few resources as possible. Three research questions are investigated in the current study. First, will changing the administration time from one minute to two or three minutes improve the validity of the QD measure? This could potentially improve floor and ceiling effects. Second, will changing the probe construction from finding the larger of two numbers to finding the middle number of a three number set improve the validity of the QD measure? If so, this would decrease the chance of obtaining a correct answer by a function of guessing and increase assessment integrity. Third, will changing the scoring of the QD measure from a purely fluency based measure (Digits Correct per Minute; DCM) to an outcome measure that takes into account fluency (DCM) while weighting the score for accuracy (A) improve the validity of the QD measure (DCM*A)? If

so, this could help rank order students better, and it would add evaluating accuracy along with fluency into one score. If proven effective, the proposed alterations (to administration time, probe construction, and scoring) to the QD measure could offer a procedure to increase the validity of the QD measure when it is used to make screening decisions.

CHAPTER II

REVIEW OF LITERATURE

The administration of curriculum-based measures is becoming common place in education. The purpose of this review of the literature is to cover the domain of curriculum-based measurement (CBM) in early mathematics. Recent research is included that has been conducted for the purpose of developing CBM early mathematics measures to screen for mathematical difficulties, progress monitor mathematic performance, and predict mathematic performance on a criterion. Curriculum and assessment recommendations that have been made by the National Council of Teachers of Mathematics (1989 & 2006) and the National Mathematics Advisory Panel (U.S. Department of Education, 2008) are listed. Recommendations for future research needed in regards to CBM early mathematics assessment are discussed.

Mathematical Curriculum Recommendations

Rivera (1997) reported that in the 1990s many professionals in mathematics set forth recommendations in regards to restructuring curricular and instructional emphasis in mathematics programs with a goal of improving mathematical instruction for all students,

including those individuals with a specific learning disorder in mathematics. In 1989 The National Council of Teachers of Mathematics (NCTM) published the *Curriculum and Evaluation Standards for School Mathematics*. The NCTM has been a leader in advocating for a change in the way that mathematics is taught, but discussion and controversy have been generated in the professional community by the *Curriculum and Evaluation Standards for School Mathematics* (1989). Some special educators challenged these standards because the standards did not explicitly mention students with learning disorders and there were limited replicable and validated instructional practices. In regards to mathematics assessment and instruction for all students, there is a need present for empirical validation of instrumentation that will successfully identify students (including students with learning disorders) who are having difficulty in mathematics and for interventions to remediate those difficulties (Rivera, 1997).

The latest publication by the NCTM is the *Curriculum Focal Points for Prekindergarten through Grade 8 Mathematics* (2006) and has the same limitations as the previous publication by the NCTM in 1989. However, the latest publication (2006) has recommended updated curriculum focal points. The NCTM (2006) stated it is important to identify curriculum focal points for each grade in order to provide a way to answer the question of how to organize curriculum standards and build mathematical facts and strategies upon one another from one grade to the next. Curriculum focal points are areas of emphasis and not grade level mastery objectives. The purpose of identifying curriculum focal points for each grade level is to help students build mathematical skills in the context of a cohesive and focused curriculum that includes reasoning, critical

thinking, and problem solving. The NCTM (2006) does not provide valid instructional procedures; however, it does provide specific curriculum goals for each grade level. These curriculum focal points are designed for a general classroom, and do not include separate guidelines for students with a mathematical learning disorder. Although the NCTM (2006) gives detailed focal points for prekindergarten through the eighth grade, only a few curriculum focal points for kindergarten and first grade will be reviewed in this report. One focal point recommended in kindergarten is numbers and operations which are defined as comparing, representing, and ordering whole numbers, along with joining and separating number sets. Numbers and operations is also a curriculum focal point in the first grade; however, it is defined as developing an understanding of whole number relationships including groupings in ones and tens. Another curriculum focal point identified in first grade is number and operations and algebra, which is defined as the development of an understanding of addition, subtraction, and strategies for basic addition facts and related subtraction facts. The NCTM (2006) suggests first grade is where a child begins to understand the connections between counting and the operations of addition and subtraction. The NCTM (2006) postulates that this connection between counting and the understanding of the related operations of adding and subtraction is the foundation needed for the acquisition of algebra skills later. In kindergarten focus is put on the relationship of whole numbers, while in first grade the relationship of whole numbers is linked with the related operations of addition and subtraction.

How do educators evaluate if students are learning necessary early mathematical facts and strategies, and how can students who are having difficulty be identified early so

intervention can be initiated before they fail at laying the foundation needed to build new facts and strategies upon in later grades? Curriculum-based measures (CBM) have been designed and researched that assess early mathematical skills which need learned by students.

A Two-Factor Model of Mathematics

Can curriculum-based measurement (CBM) be used successfully to identify students who are having difficulty in mathematics? To answer this question we must first identify what mathematical curriculum-based measurement (M-CBM) actually assesses. Thurber, Shinn and Smolkowski (2002) performed a study with the purpose of examining the relationship of the constructs of general mathematics achievement, computation, and application to M-CBM by using confirmatory factor analysis. Computation is also referred to as operations, while application is also known as problem solving. The main idea of the study was to determine what construct M-CBM actually measures. The participants in this study were 207 fourth graders that were from four general education classrooms in a midsized Northwestern public school district.

The measures used in the Thurber et al. (2002) study were the M-CBM, Basic Math Facts Probe, Stanford Diagnostic Mathematics Test (SDMT; Psychological Corporation, 1996), California Achievement Tests (CAT; CTB/McGraw-Hill, 1992), National Assessment of Educational Progress (NAEP; 1992), and a Reading Maze Test. The M-CBM included three mixed-operation probes that were sampled from the annual curriculum of typical mathematics textbooks. The participants were given five minutes to complete as many problems as possible on each probe. Scoring was accomplished by

counting the number of correct digits that were obtained while calculating the correct answer, but this scoring procedure possibly compromised the interscorer reliability. The Basic Math Facts Probes were administered by two fact worksheets. Descriptions of the types of problems included in the probes are documented in the Thurber et al. (2002) article. Participants were instructed to complete as many problems as possible in two minutes, and scoring of the probes was accomplished by counting the number of correct problems. The computation subtest and applications subtest of the SDMT were administered. The math computation subtest and the math concepts and applications subtest of the CAT were administered. The applications items of the NAEP were administered. Three reading maze tests were administered.

The results of the Thurber et al. (2002) study showed the best fit for the data was a two-factor model where computations and applications are distinct but related constructs, reading skill was highly correlated with both computation and applications, and M-CBM was a measure of computations. A secondary finding was that reading skills play an important role in general mathematics assessment.

One potential application for CBM mathematics measures is for use in screening and identifying children who are having difficulty in mathematics. Several researches are developing mathematical measures that can be applied for this use (VanDerHeyden, Witt, Naquin, & Noel, 2001; VanDerHeyden, Broussard, Fabre, Stanley, & Creppell, 2004; VanDerHeyden, Broussard, & Cooley, 2006; Clarke & Shinn, 2004; and Chard et al. 2005). A second potential application for CBM mathematics measures is for use in progress monitoring during mathematical interventions. Researchers that have studied

the use of math measures for progress monitoring are Foegen, Jiban, and Deno (2007) and Shapiro, Edwards, and Zigmond (2005). A third potential use for the application of CBM mathematics measures is for the prediction of future performance on a criterion such as state standardized tests (Jiban & Deno, 2007) and early prediction of a mathematical learning disorder (Fuchs et al. 2007).

The purpose for the remainder of this literature review will be to give a description of the work done by researchers and investigate two questions. The first question is what content do CBM mathematic measures need to include, and how long do CBMs need to be to be reliable and valid (Jiban et al., 2007; Fuchs, Fuchs, Hamlett, & Walz, 1993)? The second question asks whether a one skill math measure can be identified as a screener for early numeracy (similar to the ORF in reading) which will help educators and school psychologists make valid criterion and norm-referenced decisions about individual students (Hintze, Christ, & Keller, 2002). The future of mathematics assessment (U.S. Department of Education, 2008) and mathematics assessment research will also be discussed.

CBM -Mathematics Probe Development for Screening

In order to provide additional instruction in math for students who need it, a valid process to identify these students needs to be identified and followed. One way to approach student identification is by the use of universal screening. A number of curriculum-based measures have been experimentally designed and researched for this purpose. Several researchers (Gersten & Chard, 1999; Clarke & Shinn, 2004; Chard et al., 2005) have conducted research and developed curriculum-based measures based on a concept known as number sense. Number sense is defined by Berch (1998) as, "An emerging construct

that refers to a child's fluidity and flexibility with numbers, the sense of what the numbers mean, and an ability to perform mental mathematics and to look at the world and make comparisons." Case (1998) stated that students who have a good number sense are, figuratively speaking, able to move seamlessly between the real world of quantity and the mathematical world of numerical expression and numbers. Gersten and Chard (1999) postulated, in an article defining and explaining number sense, that number sense is as an important concept for mathematical learning as phonemic awareness is for the development of reading ability, and that this concept can be helpful in enhancing mathematical instruction for students with learning disabilities. Students with good number sense will be able to navigate through a number set by the use of the concept of magnitude and will be able to apply this automatic use of mathematic information to solving basic arithmetic computations. Researchers who have developed and/or studied curriculum-based measures used as screeners based on number sense are Clarke and Shinn (2004) and Chard et al. (2005).

Clarke and Shinn (2004) investigated the sensitivity, reliability, and validity of four early mathematics measures that were experimental and designed to be used in early identification and formative evaluation. Reliability assessment included test-retest, alternate forms, and inter-scorer reliability. Concurrent and predictive validity was examined using three criterion measures. Measures were designed for the purpose of assessing precursors of mathematics understanding that are present before children are able to perform formal mathematics. Clarke and Shinn (2004) stressed that identifying and intervening with students who may be at-risk the most for failure later are keys to

preventing mathematics difficulties. Participants consisted of 52 first grade students from the Fall 2000 to Spring 2001 academic year. Participants were recruited from two schools in a school district of 2,500 students located in the Pacific Northwest.

All four of the experimental measures designed in the Clarke and Shinn (2004) study were individually administered and were one minute in duration. The measures consisted of the following: Oral Counting (OC); Number Identification (NI); Quantity Discrimination (QD); and Missing Number (MN). The OC measure required students to count orally, and the measure was scored as number of numbers correctly counted in one minute. The NI measure required students to orally identify numbers from 0 to 20 when presented as a set of printed number symbols, and the measure was scored as numbers identified correctly in one minute. The QD measure required students to name the larger of two visually presented numbers, and the measure was scored as the number of larger numbers correctly identified in one minute. The MN measure required students to name the missing number from a string of numbers from 0 to 20, and the measure was scored as the number of missing numbers identified correctly in one minute. The criterion measures used were the M-CBM grade one computation probes (which required students to answer math problems that were drawn from the students' first-grade mathematics curriculum and were two minutes in duration), the WJ-R applied problems subtest (Woodcock & Johnson, 1989), and the *Number Knowledge Test* (Okamoto & Case, 1996).

The results of the Clarke and Shinn (2004) study found that each of the four experimental measures had sufficient evidence for sensitivity, reliability, and validity.

No measure had poor reliability, the QD measure was the most reliable measure, and the NI was the second most reliable measure. All four experimental math curriculum-based measures (EM-CBM measures) were found to be reliable when making screening decisions about an individual student. In general, moderate to strong evidence was found for the concurrent validity of all four EM-CBM measures, with the QD measure being the strongest measure of early mathematics due to the significantly stronger relationship with the criterions when compared to the other three measures. Support was strongest for the use of the QD measure as a single indicator of early mathematics. Limitations of the study and future research recommendations are not discussed here, but are listed in the article (Clarke & Shinn, 2004). The OC, MN, QD, and NI measures investigated by Clarke and Shinn (2004) are a part of the AIMSweb Tests of Early Numeracy (TEN; Clarke & Shinn, 2002) mathematical probe set for kindergarten and first grade. AIMSweb is a curriculum-based measurement tool that allows educators and school psychologists to screen and progress monitor students who are demonstrated to have difficulties in reading, writing, and mathematics when compared to national and local norms.

Three of the four mathematical probes that were developed by Clarke and Shinn (2004) were also evaluated in a kindergarten and first grade sample by Chard et al. (2005). The three measures evaluated were Number Identification (1-20), Quantity Discrimination (1-20), and Missing Number (1-20). This study expanded research with these measures in three ways; first, an increase of the first grade sample to 483 students; second, the probes were administered to kindergarten students; third, the number range of

all three measures were changed from 1-20 to 1-10 in order to evaluate the kindergarten sample.

The overall purpose of the Chard et al. (2005) study was to investigate ten curriculum-based measurements of early mathematics, based on the concept of number sense, and designed to screen students in kindergarten and first grade to identify those at risk for potential mathematic difficulties. The mathematics measures investigated were Number Identification, Quantity Discrimination, and Missing Number developed by Clarke and Shinn (2004), as well as a number writing measure and six counting measures. The kindergarten and first grade samples were taken from seven schools in a school district of 5,550 students from the Pacific Northwest. The mathematics measures were administered in the fall, winter, and spring of the 2002-2003 academic year. All the measures were administered in the fall, and a selected set of measures (selection based on a correlation of $\geq .50$ with the criterion in the fall) were administered in the winter and spring. The *Number Knowledge Test* (Okamoto & Case, 1996) was used as the criterion for this study. Assessment scores were only reported for those participants who were present for all three administration periods (kindergarten n = 168; first grade n = 207).

A number writing measure and six counting measures were developed for the Chard et al. study (2005). These measures were administered only in the fall of 2002. No further data was collected on a measure in the winter or spring if the correlation with the criterion was $< .50$ (assumed the measures would not be valid for screening purposes due to the low correlation), there was a floor effect, or there was difficulty with administration/scoring the probe. The number writing measure instructed the child to

write the number symbol for a number administered orally. Students were randomly given all numbers from 1 to 20 and five random numbers from 20 to 100. The correlation in the fall between this probe and the *Number Knowledge Test* (Okamoto & Case, 1996) was .63 for kindergarten and .46 for first grade; however, the probe administration was not repeated in the winter or spring due to administration and scoring difficulty. The six counting measures were arranged in three different categories. The three categories were count to 20 (orally count from 1 to 20), count from 3 and 6 (orally count the next five numbers starting from 3 and 6), and count by 2, 5, and 10 (the student is asked to count the next 10 numbers in the count by sequence). For kindergarten, the correlation range in the fall between the measures and the *Number Knowledge Test* was .41-.50 (count by 10s was the only measure that was .50). For first grade, the correlation range in the fall between the measures and the Knowledge Number Test was .07-.48. These measures were not proven to have statistical or practical utility as screeners in the fall, and therefore were not administered in the winter and spring.

The Number Identification, Quantity Discrimination, and Missing Number probes were administered in the fall, winter, and spring of the 2002-2003 academic year during the Chard et al. (2005) study. The correlation between the *Number Knowledge Test* (Okamoto & Case, 1996) and Number Identification for kindergarten students was .65 in the fall and .58 in the spring, and for first grade students it was .56 in the fall and .58 in the spring. The correlation between the *Number Knowledge Test* and Quantity Discrimination for kindergarten students was .55 in the fall and .50 in the spring, and for first grade students it was .45 in the fall and .53 in the spring. The correlation between the

Number Knowledge Test and Missing Number for kindergarten students was .69 in the fall and .64 in the spring, and for first grade students was .61 in the fall and .61 in the spring. The Chard et al. study provided validation for the Number Identification, Quantity Discrimination, and Missing Number measures that were studied in the Clarke and Shinn (2004) study and are included in the AIMSweb Tests of Early Numeracy (TEN; Clarke & Shinn, 2002) probe set.

Other attempts have been made at developing early numeracy measures for screening. Three early numeracy measures were developed and investigated in a study by VanDerHeyden, Witt, Naquin, and Noell (2001). These measures were developed for the purpose of assisting in the identification of kindergarten students who exhibit deficient readiness skills in math, reading, and writing by proactively administering the measures to classes. The 107 participants in the study were from six classrooms at two suburban public schools in south Louisiana.

The three mathematics measures developed in the VanDerHeyden (2001) study were Circle Number, Write Number, and Draw Circles. The measures were designed to be group administered. The Circle Number measure gives the student one minute to count sets of circles (ranging from one to ten) on one side of a page and circle the correct number of circles from four choice lists on the other side of the page. The Write Number measure gives the student one minute to count sets of objects (ranging from one to ten) and write the number of objects counted in corresponding answer boxes. The Draw Circles measure gives the students one minute to view numbers (ranging from one to ten) and draw circles that represent those numbers on the opposite side of the page.

Phase one of the VanDerHeyden et al. (2001) study was to establish the reliability of each of the measures. Three measures of reliability were evaluated. The first was the coefficient alpha that was calculated to estimate the internal consistency of the alternate forms. The Circle Number measure and the Write Number measure had coefficient alphas of ≥.9, while the Draw Circle measure had a coefficient alpha of >.8. The second reliability measure was alternate forms. The Circle Number measure correlation was .84, the Write Number measure was .81, and the Draw Circle measure was .70. The third reliability measure was interscorer agreement, and the average percent agreement was calculated as the number of items for which the scorers agreed divided by the number of items the scorers disagreed plus the number of items the scores agreed. The average percent agreement for the Circle Numbers measure was 99.3 (range of 90-100), 95.56 (range of 80-100) for the Write Number measure, and 100 (range of 100-100) for the Draw Circle measure.

Phase two of the VanDerHeyden et al. (2001) study was the validity phase of the study. The criterion used for the mathematics component of this study was the *Comprehensive Inventory of Basic Skills, Revised* (CIBS-R; Brigance, 1999). Three measures of validity were obtained and these were concurrent validity (correlation with the CIBS-R), predictive validity (discriminant functional analysis), and social validity (social validity Likert-scale given to eight teachers). The three math measures were part of a set of curriculum-based measures (included reading and writing as well as mathematics) that were used to identify students exhibiting deficient readiness skills for reading, writing, and mathematics in kindergarten. The predictive validity was evaluated

for the set of curriculum-based measures as a whole; therefore, this data is not reported, because this literature review deals specifically with the subject of mathematics. Social validity is not discussed in this literature review due to the limited number of teachers sampled (eight). The concurrent validity correlations between the CIBS-R and the mathematics probes were found to be significant only for the Circle Number measure. The Circle Number measure was significant (at the .003 level) with the CIBS-R: understands quantitative concepts subtest (r = .55), CIBS-R: counts objects subtest (r = .53), and the CIBS-R: math composite (r = .61). The results of this study demonstrate that curriculum-based measures can have technical adequacy for screening students in the primary grades, and identifying students in kindergarten who exhibit deficient skills.

In a second study, VanDerHeyden, Broussard, Fabre, Stanley, & Creppell (2004) developed six curriculum-based measurement probes that were designed for use in measuring math competence (important early math skills) and growth of preschool children enrolled in public school programs. Reliability of the probes was addressed (to include source, item, and alternate forms) and concurrent validity (with two measures) was assessed. The study was performed in southern Louisiana at two rural public preschool programs that served primarily children at risk. The randomly selected sample from the two programs included 102 children (mean age of 59 months).

In the VanDerHeyden et al. (2004) study the six mathematics probes were correlated with two criterion measures, *Brigance Screens* (1999) and the *Test of Early Mathematics-2* (Ginsburg, & Baroody1990), that were chosen to indicate math-specific performance. The six measures developed were Choose Number, Number Naming,

Count Objects, Free Count, Discrimination, and Choose Shape. All measures were scored as number correct divided by total time to respond, with the exception of the Discrimination measure which was scored as number correct per minute. The Choose Numbers, Number Naming, Count Objects, and Discrimination measures were found to be valid curriculum-based measurement (moderately correlated with the criterion measures) for evaluating performance in mathematics among preschool students. The Free Count and Choose Shape measures were not found to be valid (weak reliability and concurrent correlation estimates were found, and it was suggested more research and development on these two probes would be needed for the probes to be acceptable).

In a study by VanDerHeyden, Broussard, and Cooley (2006) four of the six measures (Choose Number, Count Objects, Free Count, and Discrimination) used in the VanDerHeyden et al. (2004) study were re-administered to the original study participants; however, the participants were in kindergarten when they were re-evaluated. The reason for re-evaluating the Free Count measure (not proven valid in the previous study) and not evaluating the Number Naming measure (proven valid in the previous study) was not documented within the article. One purpose of this study was to re-evaluate the validity and reliability of four of the six measures that were evaluated during the VanDerHeyden et al. (2004) study. The probe scores were correlated with one criterion measure (Brigance Screens, 1999), and a moderate correlation was found. The Choose Number and Count Objects scoring procedures were modified from number correct divided by time to respond to number of items correct per minute. The new scoring procedure was evaluated to determine if using shorter mathematics probe duration (one minute) was as

reliable as the original procedure. The number of measurement items for the Choose Number and Count Objects measures were increased from 10 to 20 when using the measures with a kindergarten population. Both new procedures (scoring procedure and increased measurement items) were found to be reliable.

A second purpose of this study was to evaluate whether these four math measures would be effective for monitoring performance in mathematics as students transition from preschool to kindergarten (VanDerHeyden et al., 2006). Findings supported the conclusion that the measures are sensitive to performance differences at the different grade levels. Children in kindergarten would be expected to perform better on the measures than they performed in pre-school. However, the measures for the kindergarten population had a broader scope of measurable items (increased from numbers 10 to 20) than for the preschool population; therefore, evaluation is needed to determine if the probes' durations need to be greater than one minute for use in a kindergarten population.

This study also developed and evaluated interventions (seven sessions) designed to improve scores on the four mathematics measures (VanDerHeyden et al., 2006). The goal was to examine the sensitivity of the four CBMs at detecting growth resulting from interventions.

One intervention was designed to be used with the Count Objects measure (VanDerHeyden et al., 2006). The intervention had three steps, which consisted of experimenter modeling, completion of the task with the experimenter, and independent task completion. Plastic bears were used for the intervention, and the intervention

involved counting the number of bears (from 1 to 20) using a random scale without repeating a number.

For the Discrimination intervention an array of four objects were presented to the child (VanDerHeyden et al., 2006). First, the child was asked to, "Find the one that is different." Second, if the child chose wrong the correct response was modeled by the researcher, while saying, "This one is different." Third, the researcher scrambled the items and he/she would state, "Which one is different?" Praise was given for correct responses. Fourth, if the child responded incorrectly a second time, then while guiding the child's hand physically to touch the correct item the researcher would state, "This one is different."

The last intervention designed matched the Choose Number measure (VanDerHeyden et al., 2006). The child was asked to choose a number from 1 to 20 that was presented on a printed number line. Ten intervention trials were used per session, and the child was asked to pick a number randomly without choosing a number twice during each session. The child was then asked to place an object on the number chosen. If the object was placed incorrectly the researcher would point to the correct number while saying, "This is the number five; put the object on the number five." Praise was given when the responses were correct. If a child had responded incorrectly after the first prompt, then the researcher would have physically guided the child's hand to the correct response while saying, "No, this is the number five; put the object on the number five."

Even though the four math measures designed by VanDerHeyden et al. (2006) appeared to be sensitive for monitoring performance in mathematics as students transition

from preschool to kindergarten, it was questioned whether these measures reflect growth that occur as a function of interventions. The children who participated in the interventions did score higher on average on all measures but one (Choose Number) after intervention; however, these differences were not shown to be statistically significant. The authors recommended the need for further research to evaluate whether these measures could evaluate growth that occurs as a result of successful interventions.

CBM-Mathematics Probe Development for Progress Monitoring

Foegen, Jiban, and Deno (2007) compiled a review of the literature designed to examine a full array of mathematics CBMs ranging from preschool to secondary schools. In general, the measures reviewed have acceptable levels of reliability, while the criterion validity of CBMs for mathematics appears to be lower than for CBMs for reading. There appears to be a low degree of consensus about the best approach to take when developing CBMs for mathematics.

Foegen, Jiban, and Deno (2007) summarized two topics clearly and concisely in their article. The first topic summarized the two different approaches used for developing mathematics CBMs. One approach for developing mathematics CBMs is termed as the curriculum sampling approach, which involves sampling curriculum used in the classroom for that grade level and including those problems on the CBM probes. The second approach for developing mathematics CBMs is termed as the robust indicators approach, which involves identifying measures that represent broadly defined proficiencies in mathematics. The second topic summarized in the article was an explanation of the three stage continuum that research studies developing CBMs in

mathematics fall on. Stage one includes studies that explore the technical adequacy of CBMs used as static indicators. Stage two includes studies that examine the technical adequacy of slopes generated as a result of repeated measurement with CBMs used for progress monitoring to detect growth. Stage three includes studies that examine the CBMs for instructional utility. Stage three studies are applied studies that investigate if the CBMs used by teachers to inform instructional decisions actually result in improved student achievement. The research studies included in the Foegen et al. (2007) literature review are categorized using this three stage continuum and by identifying which CBM development approach was used.

Chard et al. (2005) evaluated the effect that the Number Identification, Missing Number, and Quantity Discrimination measures have when used to progress monitor. According to Foegen et al. (2007) the Chard et al. (2005) study would be classified as a stage one and stage two robust indicator study. It was demonstrated that these measures were effective when used as screeners for identifying students who have mathematical difficulties; however, research that evaluated whether or not these could be effectively used for progress monitoring was limited. The Chard et al. study explored the technical adequacy of the Number Identification, Missing Number, and Quantity Discrimination measures' use as static indicators for use in progress monitoring. Data was collected in the fall, winter, and spring of the 2002-2003 academic year, which provided three data points for evaluation. The magnitudes of the change differences across the three measures were similar in kindergarten and first grade. The only measure that demonstrated substantial growth from fall to spring was the Number Identification probe. Missing

Number and Quantity Discrimination demonstrated less change from fall to spring, and it appeared that these probes may not be sensitive enough to actual learning that occurs over time. Since only three data points were collected across an entire year, validation of this study's findings needs replication with more data points collected to demonstrate more precisely the sensitivity of the three measures.

Lembke, Foegen, Whittaker, and Hampton (2008) evaluated the technical adequacy of using the Number Identification, Missing Number, and Quantity Discrimination measures for monitoring progress in early numeracy, and provide validation/replication needed to investigate the measures' sensitivity to change in mathematical learning. The specific purpose of the study was to examine the use of these three early numeracy measures to monitor student progress. The Lembke et al. (2008) study explored the technical characteristics of slopes generated as a result of repeated measurement with the Number Identification, Missing Number, and Quantity Discrimination measures for progress monitoring.

Probe administration for the measures in the Lembke et al. (2008) study was one minute for each individual probe administration. The content of the three measures were modified from the content in the Clarke and Shinn (2004) study and the Chard et al. (2005) study. The probe content was identical for kindergarten and first grade. The Quantity Discrimination measure instructed the student to identify the larger of two numbers. Two numbers were used for each probe item and these numbers were randomly selected from either of two number sets (a number set of 0-10 and a number set of 0-20). The Number Identification measure instructed the student to identify a number (0-100)

that is visually presented to the student. Fifty percent of the sample items ranged from 0-20, 30% ranged from 0-50, and 20% ranged from 0-100. The Missing Number measure instructed the student to identify one number that is missing in a series of four numbers. Eighty percent of the items were presented in intervals of 1s (range 1-20), and 20% were presented in intervals of 5s and 10s (range 5-50 for 5s; range 10-100 for 10s).

Criterion validity was assessed during the Lembke et al. (2008) study using two measures. One criterion measure was a Likert-scale (7-point) given to teachers in order to rate their students' overall general mathematics proficiency relative to other students in the class. The total number of teachers completing the measure was apparently not documented in this study; therefore, the results of the concurrent criterion validity between this measure and the probes is not discussed in this literature review, because there is not a sufficient amount of data available in order to validate the reliability of the measure itself.

The second criterion measure (Stanford Early School Achievement Test; SESAT; Psychological Corporation, 1996) used to assess criterion validity was administered to the first grade sample (N=29) in the fall prior to the study. Concurrent criterion validity between the SESAT and the Quantity Discrimination measure was statistically significant ($r = .50$; $p \leq .01$), meaning that 25% of the variability in mathematic performance was shared between the SESAT and Quantity Discrimination measure. Concurrent criterion validity between the SESAT and the Number Identification measure was statistically significant ($r = .41$; $p \leq .01$), meaning that 16.8% of the variability in mathematic performance was shared between the SESAT and Number Identification measure.

Concurrent criterion validity between the SESAT and the Missing Number measure was not statistically significant (r = .21), meaning that only 4.4% of the variability in mathematic performance was shared between the SESAT and Missing Number measure.

The results of the Lembke et al. (2008) study revealed that there is a significant increase in linear growth across time on the Number Identification measure in both kindergarten and first grade. There were significant differences in Number Identification final scores and growth rate of Number Identification scores between kindergarten and first grade. In kindergarten, the growth rate for the Number Identification measure was 2.04 between each measurement time, and the weekly growth rate was .34. In first grade, the growth rate for the Number Identification probe was 1.42 between each measurement time, and the weekly growth rate was .24. The slopes were statistically significant ($p < .05$) for the Quantity Discrimination and Missing Number measures; however, growth rates were found to be nonlinear. Estimates of growth rates were given in the research article; however, it was cautioned in the article that these estimates were rough and were not likely to be constant over time, and these results are not listed in this literature review due to this caution.

How do the results in the Chard et al. (2005) and Lembke et al. (2008) studies affect the instructional utility of the Number Identification, Quantity Discrimination, and Missing Number measures in the classroom? The Chard et al. study demonstrates that these three measures are technically adequate for use as universal screeners. The Lembke et al. study demonstrates that the Number Identification measure is technically adequate for monitoring progress. More investigation is needed into the use of these three

measures for monitoring progress. Lembke et al. stated that their study findings were limited due to the single geographic location and the lack of diversity in the student sample. A recommendation was given for further research to be conducted that examines these three measures on a weekly basis while teachers administer and graph data in order to assess effectiveness of instruction. According to Foegen et al. (2007), this would be a stage three robust indicator study. A stage two study that would replicate the findings in the Lembke et al. study would also help improve the reliability of the research that has already been conducted. Long-term (several months or more) data that would evaluate progress monitoring slopes needs to be collected using these three measures on a weekly basis, at a number of geographic locations, with a diverse sample of students.

Shapiro, Edward, and Zigmond (2005) conducted a stage two curriculum sampling study that investigated weekly progress monitoring slopes of learning for mathematics that was implemented with special education students. Both factors of the two-factor mathematics model (Thurber et al., 2002) were investigated in the Shapiro et al. study, which included computation (operations) and applications (problem solving). Students in the first through sixth grade were used as participants in this study.

Fifteen school districts in Pennsylvania volunteered to participate in the Shapiro et al. (2005) study. Five school districts were from the western regions of the state, five were from the eastern regions of the state, and five were from the central regions of the state. Five special education teachers and a site coordinator were chosen by each school district. Each teacher identified a minimum of two students for whom progress monitoring by a General Outcome Measure (GOM) would be used in math computation,

math concepts/applications, and reading. Only mathematics data was reported in the Shapiro et al. article. One hundred and twenty students were progress monitored in math computation, while 109 students were progressed monitored in math concepts/applications. Two to 15 students in each district were assessed; the mean number assessed per school district was nine for math computation and eight for math concepts/applications. Of the 120 students, 16 were in middle school (two school districts) while the remainder of the study sample was in an elementary school building. Of the 120 students 5.8% were students who were in programs for emotional and behavioral support. Out of the seven students who were receiving emotional support, four students received part-time support and three students received itinerant support. One hundred and thirteen participants were students with learning disabilities in need of learning support. Thirty-one students received full-time support (out of the general education classroom for > 50% of the school day), 69 students received part-time or resource room support (receiving special education services between 25-50% of the school day), and 13 students received itinerant support (fully included students receiving special education services for < 25% of the school day).

Shapiro et al. (2005) chose to use the blackline masters from *Monitoring Basic Skills (MBSP) – Math Computation (2nd edition)* (Fuchs, Hamlett Fuchs, 1998) to assess progress in math computation, and the blackline masters from *Monitoring Basic Skills (MBSP) – Math Concepts and Applications (2nd edition)* (Fuchs, Hamlett Fuchs, 1999) to assess progress in math concepts/applications. The MBSP Computation and the *Math Concepts and Applications (2nd edition)* are group administered by computer; however,

the blackline masters used in the Shapiro et al. study were paper copies of the computerized probes, which allowed for computer skill to be controlled as a potential extraneous variable in test administration. Every student was assessed once per week at his/her instructional level. Computation and concepts/applications assessments were alternated from week to week; therefore, students received assessment in computation and concepts/applications once every two weeks.

In the Shapiro et al. (2005) study each math computation probe consisted of 25 mixed-operation problems presented on a single sheet. The first and second grade probes consisted of addition and subtraction (without and with grouping). In third grade simple multiplication and division facts were introduced. In fourth grade fractions, multi-digit multiplication, and simple division were added to the probes. In fifth and sixth grade decimals, complex fractions, and multi-digit division with remainders were introduced. The dependant measure used to reflect student performance on math computations was the total number of digits correct. Students in the first and second grades were given two minutes to complete the probes, third and fourth grades were given three minutes to complete the probes, and fifth grade was given five minutes to complete the probes. Probe administration time was not listed for sixth grade.

In the Shapiro et al. (2005) study each math concepts/applications probe consisted of 18 problems for second grade and 24 problems for third through sixth grades. The purpose of the math concepts/applications probes was to assess mastery of math concepts and application skills. The concepts/applications that the measures covered were number concepts, names of numbers, counting, measurement, money, charts and graphs,

fractions, word problems, and applied computations. Problems increased with grade level, required between one and three responses, and varied between multiple choice and fill in the blank. The dependant measure used to reflect student performance on math concepts/applications was the number of parts (one to three responses for each problem) answered correctly, and a point was given for each part answered correctly. Students in the second grade were given eight minutes to complete the probes, third and fourth grades were given six minutes to complete the probes, and fifth grade was given seven minutes to complete the probes. Probe administration time was not listed for sixth grade.

The analysis goal in the Shapiro et al. (2005) study was to report mean slopes for students with disabilities. The goal set for math computations and math concepts/applications performance was 0.5 digits per week (one digit on each probe type assessed every two weeks). The mean cross-grade (N=120) weekly growth rate was 0.38 digits per week on measurement of math computation. The mean cross-grade (N=109) weekly growth rate was 0.38 points per week on measurement of math concepts/applications. The within-grade weekly growth rate (digits/week) on measurement of math computation was 0.32 for first grade (N=15), 0.28 for second grade (N=42), 0.43 for third grade (N=50), 0.72 for fourth grade (N=10), no growth for fifth grade (N=3), and no reported data for sixth grade. The within-grade weekly growth rate on measurement of math concepts/applications was not evaluated for first grade, .36 for second grade (N=47), 0.37 for third grade (N=49), 0.44 for fourth grade (N=10), 0.52 for fifth grade (N=3), and no reported data for sixth grade.

The focus of the Shapiro et al. (2005) study was frequent data analysis (to perform slope analysis) and instructional adjustment in response to these analyses (using CBM procedures to adjust individual instruction); however, the use of these CBM procedures for individual instructional adjustment was not investigated empirically in this study. Further investigation comparing the CBM procedures used in this study to other educational procedures used to improve student achievement in math computation and concepts/applications is needed. This further investigation is needed to assess the utility of using the Shapiro et al. (2005) CBM procedures for adjusting individual instruction with special education and general education students.

CBM-Mathematics Probe Development for Predicting Performance

Fuchs et al. (2007) conducted a study that provided another important step in the development of reliable and accurate CBM and RTI procedures. Fuchs et al. (2007) evaluated four screening measures for use as predictors of mathematics disorder by the end of second grade and as math progress-monitoring tools. The four measures were Number Identification/Counting, Fact Retrieval, Curriculum-Based Measurement Computation, and Curriculum-Based Measurement Concepts/Applications. The sample for the study was 225 students who participated from the entrance of first grade to the completion of second grade. For this study, mathematics disorder was operationally defined as below the tenth percentile on word problems and calculations by the end of the second grade. Classification accuracy for identifying mathematics disorder was primarily driven by CBM Computation and CBM Concepts/Applications (best predictor), while CBM Computation appeared valid for use with progress monitoring.

The Number Identification, Missing Number, and Quantity Discrimination measures that were evaluated for use with screening and progress monitoring in the Chard et al. (2005) study were also investigated for statistical predictive utility when they were used as variables in a regression analysis. The research question for this segment of the study was how well the fall administration of these measures predicted scores on the *Number Knowledge Test* (Okamoto & Case, 1996) administered in the spring? In the kindergarten sample, a model including two predictor measures was used (Missing Number and Quantity Discrimination), and this model was found to be statistically significant [$F=63.0$, (2, 165), $R=.66$, $p<.01$] accounting for 43.6% of the variance on the Knowledge Number Test administered in the spring. In the first grade sample, a model including three predictor measures was used (Number Identification, Missing Number, and Quantity Discrimination), and this model was found to be statistically significant [$F=59.2$, (3, 203), $R=.683$, $p<.01$]. All three measures were reported to account for unique and significant variance on the *Number Knowledge Test*. Overall, 46.6% of the variance was accounted for by a fall administration of the Number Identification, Missing Number, and Quantity Discrimination measures.

Jiban and Deno (2007) investigated concurrent validity between three simple one-minute curriculum-based measurements and a state (Minnesota) standards test of mathematics among 35 third and 49 fifth graders. Two numeric CBMs, a new "cloze" math facts measure and a traditional basic math facts measure, were investigated as predictors. A silent reading CBM was evaluated as a third predictor. An additional question investigated was whether a one-minute administration of the cloze math

measure was technically adequate or if scores needed to be aggregated from two separate administrations to be reliable.

The basic math facts measures for the Jiban and Deno (2007) study were from the Basic Academic Skills Sample (Espin, Deno, Maruyama, & Cohen, 1989) and were adjusted to a horizontal versus a vertical format. Measures consisted of problems from all four basic operations (addition, subtraction, multiplication, and division) using two single-digit numbers (4+2=?, 5-4=?, etc.). One minute was given to each student to complete as many problems as possible. Two forms were administered on consecutive days, one on each day. Scoring was assessed for problems correct and problems correct minus problems incorrect. Both scoring procedures were found to be equivalent.

The cloze math facts measures for the Jiban and Deno (2007) study were selected from the Basic Academic Skills Sample (Espin et al., 1989) and were adjusted to a horizontal versus a vertical format. The cloze math facts were similar to the basic math facts, with the exception that the missing number could be in any position (?+3=5, 6-?=2, 3x2=?, etc.). The basic math facts measures all had the missing number after the equals sign. One minute was given to each student to complete as many problems as possible. Two forms were administered on consecutive days, one on each day. Scoring was assessed for problems correct and problems correct minus problems incorrect. Both scoring procedures were found to be equivalent.

The maze reading measures for the Jiban and Deno (2007) study were selected from the Basic Academic Skills Sample (Espin et al., 1989). Reading passages were presented to each student with every seventh word missing. Each student was given three

word choices to choose from (two incorrect and one correct). Each measure was scored by taking the mazes correct and subtracting the mazes incorrect. Scoring was stopped if three consecutive mazes were missed and the score stopped at the last correct maze response. One minute was given to each student to complete as many mazes as possible. Two forms were administered on consecutive days, one on each day.

The criterion variable for the Jiban and Deno (2007) study was the *Minnesota Comprehensive Assessment in Mathematics* (2007). This is a primarily multiple–choice test, and no problems require direct computation of basic math facts in isolation. The math portion of the assessment served as the criterion measure for each grade level, and areas included number sense, problem solving, procedures and concepts, as well as other mathematical concepts. All problems in both grades include text that requires reading in order to understand and solve the problems.

Results for third grade participants were inconclusive (Jiban & Deno, 2007) due to inadequacies in the research design (low reliability of scores). The best regression model that predicted performance for fifth grade students on the *Minnesota Comprehensive Assessment in Mathematics* (2007) included the cloze math and maze reading measures. Predictions for fifth graders on the criterion were improved by 52% when using the combination of cloze math facts and maze reading as predictors. Jiban and Deno (2007) recommended that reading performance be included as a predictor for performance on mathematics achievement tests, but it was stressed that this study was a first look into the use of cloze math facts and maze reading to predict mathematics performance on a criterion and further examination is needed. Jiban and Deno (2007) also

discovered that cloze math facts were as equally reliable as basic math facts when the measures were used to predict performance on a criterion. Third grade and firth grade scores on the cloze math measure were evaluated individually and aggregated. Aggregating the scores of two administrations on the one-minute cloze math probes appeared to be a potential method to improve reliability of the scores (from .65 to .86 for fifth grade), but it was reported there is a need for further investigation.

Developing reliable and accurate CBMs is an important step in developing CBM and RTI procedures that can be used to predict and identify mathematics difficulties early in a student's academic experience. Once CBMs are in place that can reliably and accurately identify mathematics difficulties, then interventions can be developed that can be used in an RTI paradigm and the CBMs can be used as measures for monitoring progress. Once students do not respond to intervention, then they can be formally evaluated and diagnosed for a specific learning disorder in mathematics.

CBM-Mathematics Probe Duration Variance by Breadth of Concepts

Should CBM mathematics probe durations differ for different grade levels (increasing probe duration as grade level increases), due to the differences in the breadth of measurable concepts? There is literature that suggests probe durations in mathematics may need to vary among different grade levels (Fuchs, Fuchs, Hamlett, & Walz, 1993). The Fuchs et al. (1993) study was an empirical, longitudinal study where participants were followed across two years in the first through sixth grades in five school districts in the upper Midwest. The study's purpose was to produce standards for intraindividual norms for slopes of achievement (weekly rates of growth) using CBM procedures in

mathematics, reading, and spelling. For the math procedures the probe durations used were 45 seconds in first grade, one minute in second grade, 1.5 minutes in third grade, three minutes in fourth grade, five minutes in fifth grade, and six minutes in sixth grade. This increase in probe duration at increasingly higher grade levels was suggested to be a reliable procedure, due to the difference in the breadth of measurable concepts at different grade levels.

Single-skill Versus Multiple-Skill CBM-Mathematics Probes

Hintze, Christ, and Keller (2002) designed a study to explore the extent to which performance within two types (single-skill and multiple-skill) of survey-level CBM mathematics assessments vary. The hypotheses of the study was that because single-skill CBM math assessments are controlled for difficulty, little practical difference would be observed across three probe administrations during CBM survey-level assessments, and given that multiple-skill probes are not controlled for difficulty, performance on these probes would vary during CBM survey-level assessments.

A total of 67 students were participants in the Hinzte et al. (2002) study and they were from the first through fifth grade. The participants were from 21 classrooms at one elementary school in the Northeast United States. The school served 335 students from first through fifth. A systematic sampling procedure was used for this study, where every fifth student was chosen from alphabetized class lists until 20% of every grade was selected. The sample was comprised of 14 first graders, 15 second graders, 12 third graders, 15 fourth graders, and 13 fifth graders.

Hintze et al. (2002) developed three probes for each measurement series (single-skill computation and multiple-skill computation). Problems on each probe were randomly selected and no one problem was repeated. The number of problems on each probe ranged from 20 to 35, depending upon the difficulty of the problem set for each probe (an easier problem set had more problems than a harder problem set). The dependant measure was the number of correct digits computed per minute for each individual probe. The multiple skill probes contained randomly selected problems that were a mix of all mathematical skills taught at that grade and all previous grades (multiple operations and multiple levels of difficulty). The single skill probe problems were selected by a different method. Each single skill probe assessed one consistent skill at one consistent level of difficulty (i.e., two-digit by two-digit addition without regrouping). Each teacher at each grade identified a particular skill or algorithm, from a hierarchically arranged set of mathematics computational skills (Shapiro, 1996), that the majority of the students were being taught. At each grade level the median skill was selected as the computational operation for all three probes at that grade level. All the students were administered all six probes (three single skill probes and three multiple skill probes) for that grade level, and were given two minutes to complete the problems on each probe. All six probes were administered during one group administration setting, and the administration sequence was counterbalanced for within and across probe type.

Hintze et al. (2002) conducted a full model generalizability analysis to evaluate the extent to which performance observed on the multiple-skill math probes could be generalized to the single-skill math probes and vice versa. Separate reduced model

generalizability analyses were also conducted on the single-skill and the multiple-skill math probes to evaluate how much variability could be attributed to the various facets of measurement within the singe-skill and multiple-skill math probes. The results of these analyses suggest that one administration of a single-skill math probe provides information of sufficient quality to make norm-referenced and criterion-referenced decisions. For the multiple-skill probes, similar results were observed; however, a repeated measures ANOVA was conducted and variations in performance across the three probe administrations was found in the first, second, and fifth grade. Therefore, Hinzte et al. recommended that three multiple-skill math probes still be used as a part of survey-level CBM math assessments. Further practical recommendations were given by Hintze et al. The type of evaluation needed should drive the choice between using one single-skill math probe or three multiple-skill math probes. If the evaluation concern is for assessing isolated skills, then obtaining one single-skill math probe administered for two minutes is a reliable and valid course for practitioners to take. If the evaluation concern is for assessing a wide variety of skills, then the most reliable course of action is to obtain three multiple-skill math probes while reporting the median score.

Future of Mathematics Assessment

As mathematics instruction changes so should mathematics assessment; therefore, it is necessary to consider and evaluate current mathematics curriculum recommendations when developing assessments. The National Mathematics Advisory Panel (U.S. Department of Education, 2008) published *the Final Report of the National Mathematics Advisory Panel* in order to recommend the actions needed to strengthen the students of

the United States in mathematical learning. One emphasis on curricular content was to ensure that students learn the critical foundations of algebra necessary for learning algebra in high school. Algebraic skills were a big focus in the National Mathematics Advisory Panel final report. For elementary school students, fluency with whole numbers and fluency with fractions was specifically listed. In regards to fluency with whole numbers, one benchmark for the critical foundation of algebra needs to be obtained by the end of third-grade. Students need to be proficient in addition and subtraction of whole numbers. Second, students need to be proficient with multiplication and division of whole numbers by the end of fifth grade. In regards to fluency with fractions, one benchmark for the critical foundation of algebra needs to be attained by the end of fourth grade. Students need to be able to identify and represent fractions and decimals, and able to compare them on a number line as well as with other common representations of fractions and decimals. Second, students need to be proficient with comparing fractions and decimals as well as percents, and able to add and subtract fractions and decimals by the end of fifth grade.

In addition to recommending curriculum changes in mathematics, The National Mathematics Advisory Panel (U.S. Department of Education, 2008) made recommendations on how to change and improve mathematics assessment for State tests and the NAEP (National Assessment of Educational Progress). These recommendations are important to consider, because CBM math probes that are developed by the curriculum sampling approach sample the curriculum. The curriculum should be relevant to the material that is contained in the State tests and the NAEP; therefore, curriculum

sampled CBM math probes should be relevant to the State test and NAEP as well. Educators can then use the CBM math probes to screen students who need intervention regarding math skills, to monitor progress as intervention is initiated, and to predict how individual students will perform on State tests and the NAEP. The National Mathematics Advisory Panel states the most critical skills that are foundational to algebra are whole numbers, operations with whole numbers, and facility with fractions. The National Mathematics Advisory Panel final report recommended a reorganization of the NAEP content strands. The content strands recommended for the fourth grade were whole numbers, fractions and decimals, geometry and measurement, algebra, and data display. Due to these recommendations CBM early math measures should continue to assess skill with whole numbers, relationships between whole numbers, and operations with whole numbers. The development of CBM math measures that would assess a student's skill at fractions and decimals is needed.

Future of CBM-Mathematics Research

At least three domains need additional investigation in the area of early mathematics CBM development, and are as follows. First, research that will investigate CBM probe administration time and/or individual versus group administration is needed. Regarding the administration time of mathematics probes, probes need to be short enough to be efficient, but long enough to be reliable. Investigation into whether or not probes need to be administered individually or group administered is needed. Individual administration would make assessment longer but could make it more reliable. Group administration would make assessment shorter but potentially could make it less reliable

due to more variability. All CBM early mathematics measures that are developed in the future should be investigated to determine the optimal probe administration time, and whether individual administration or group administration is more reliable and efficient.

Second, improving the construction and/or scoring of existing CBM early mathematics measures could help increase the validity of these measures. Examples would be expanding the content of current CBM early mathematics probes and/or altering scoring methods in a manner that would decrease the chance of obtaining a correct answer by a function of guessing and change scoring in a manner that could help rank order students better by increasing floor and ceiling effects.

The third and final domain that requires further investigation is the development of CBMs in early mathematics that can be used to assess early algebra skills and skills with fractions and decimals. Procedures for the cloze math measure (Jiban & Deno, 2007) could be used to assess early algebra skills. Procedures for early numeracy measures such as Quantity Discrimination, Number Identification, and Missing Number (Chard et al. 2005; Clarke & Shinn, 2002 & 2004; Lembke et al. 2008) could be used to assess skills with fractions and decimals as is done with whole number relationships. Researchers who can make novel use of early mathematics CBMs already developed, and/or who can develop new measures are desperately needed within the field of early mathematics CBM assessment.

Addressing the first two of the three domains mentioned above could help researchers find a one skill screener for numeracy, which would be helpful to educators because it would allow for the reduction of the total number of various measures

currently needed to accomplish a thorough screening of a student's early numeracy skills. In regards to reading, the oral reading fluency (ORF) measure is used as a reliable screener for literacy. If a CBM early mathematics measure could be demonstrated as a reliable and valid one-skill screener for numeracy, then this would be valuable to CBM mathematics assessment.

Purpose of the Current Study

Selecting the QD measure as the TEN to investigate for the current study was chosen because research (Clarke & Shinn, 2004) has demonstrated that the QD measure is the best possible TEN for use as a one skill screener for numeracy. Regarding the validity of the QD measure, the current study asks three research questions that address the first two of the three domains identified earlier in the Future of CBM-Mathematics Research section of this chapter. First, will changing the administration time from one minute to two or three minutes improve the validity of the original QD measure? Second, will changing the probe construction from finding the larger of two numbers to finding the middle number of a three number set improve the validity of the original QD measure? Third, will changing the scoring of the QD measure from a purely fluency based measure (Digits Correct per Minute; DCM) to an outcome measure that takes into account fluency (QDC) while weighting the score for accuracy (DCM*A) improve the validity of the original QD measure?

The three changes (administration duration, probe construction, and scoring procedure) to the original QD measure identified in the current study could, by main effect(s) or interaction effect(s), improve the validity of the QD measure to a point where

the QD measure could explain enough of the variance in early numeracy to be used as a one-skill screener for numeracy. Seven hypotheses have been identified for investigation during the current study. Hypothesis one, changing the administration time from one minute to two or three minutes will improve the validity of the original QD measure. Hypothesis two, changing the probe construction from finding the larger of two numbers to finding the middle number of a three number set will improve the validity of the original QD measure. Hypothesis three, changing the scoring of the QD measure from a purely fluency based measure (Digits Correct per Minute; DCM) to an outcome measure that takes into account fluency (DCM) while weighting the score for accuracy (DCM*A) will improve the validity of the original QD measure. Hypothesis four, the combination of changing administration time and changing the scoring method together will interact to improve the validity of the original QD measure. Hypothesis five, the combination of changing administration time and changing probe construction together will interact to improve the validity of the original QD measure. Hypothesis six, the combination of changing probe construction and changing the scoring method together will interact to improve the validity of the original QD measure. Hypothesis seven, the combination of changing the administration time, changing probe construction, and changing the scoring method together will interact to improve the validity of the original QD measure.

CHAPTER III

METHODOLOGY

Participants

The participants in the current study included 94 first grade students from two elementary school(s) in one Midwestern United States school district. Participants ranged from 6 to 8 years of age, and 38 were male and 56 were female. Twenty-six percent of the population of the school district received free or reduced lunches. Convenience sampling was used: every student who returned a parent consent form allowing participation and assented to participation had his or her data used in the study. Two groups were needed (one group was administered a two-item QD probe, while a second group was administered a three-item QD probe). Participants were listed alphabetically (per classroom, and seven classrooms were used to recruit participants), each participant was assigned a participant number starting at the top of each classroom list and continuing to the bottom of each classroom list, then participants with an odd participant number were assigned to the two-item QD probe group and participants with an even participant number were assigned to the three-item QD probe group. Forty-six

participants were in the two-item QD probe group (originally 48 were recruited but two participants did not assent), and 48 participants were in the three-item QD probe group.

Materials

Two QD probe types were used to assess quantity discrimination performance (each probe type was generated into three alternate forms). One probe contained 120 QD number sets, was administered to the first group of two, and each participant (in group one) was instructed to verbally choose the larger of a two number set presented visually (refer to Appendix B). A second probe contained 120 QD number sets, was administered to the second group of two, and each student (in group two) was instructed to verbally choose the middle digit of a three number set presented visually (refer to Appendix B).

The *Number Knowledge Test* (NKT; Okamoto & Case, 1996) was administered and used as a criterion to calculate concurrent validity. The NKT is designed to measure the intuitive knowledge of number that the average child has available at the age-levels of 4, 6, 8, and 10 years. It is a developmental test, meaning that knowledge at level 0 is a prerequisite for progressing to the next level (up to level 3). The NKT has been used as a criterion measure to calculate concurrent validity for the QD measure in previous research and during development of the original two-item QD measure (Chard et al., 2005; Clarke & Shinn, 2004). This is the rationale for using the NKT as the criterion measure to calculate concurrent validity for the current study.

Procedures

Administration occurred during the second semester of the academic year. Each participant was administered one QD probe (either the two-item or three-item QD probe)

and the criterion (NKT; Okamoto & Case, 1996), and both were administered on the same day. The QD and NKT administrations were counterbalanced (half of the participants of both groups were administered the QD probe first and the NKT second, while the other half of the participants were administered the NKT first and the QD probe second). Once the participant lists were divided into two groups (a two-item QD group and a three-item QD group), for every other participant in each group the administration order was changed (i.e. the first participant in the two-item QD group was administered the QD probe first, the second participant in the two-item QD group was administered the QD probe second, the third participant in the two-item QD group was administered the QD probe first, etc.).

Each student was individually administered one of the two QD probe types (two-item or three-item). This created two groups. Each participant in both groups was administered one three-minute QD probe that corresponded to the type of probe (two-item or three-item) assigned to his/her group. The two-item probe type included 120 QD number sets and the participant was instructed to verbally choose the larger of a two number set presented visually. The administration directions for the two-item probe were identical to the directions designed by Clarke & Shinn (2002) for the two-item QD probe, with exception that the administration time was altered from a one-minute to a three-minute administration. A participant copy of the two-item QD probe was placed in front of the student. A research assistant's copy was placed on the examiner's clipboard and positioned so the participant could not see what the research assistant recorded. Each was administered two examples to ensure he/she understood the procedure for completing the

probe. These specific directions were given to the participant (regarding the examples), "Look at the piece of paper in front of you.....the box in front of you has two numbers in it (demonstrate by pointing)…..I want you to tell me the number that is bigger." If the correct response was given by the participant (for example one) the following feedback and directions were given, "Good….the bigger number is 7…..now look at this box (demonstrate by pointing)…..it has two numbers in it…..tell me the number that is bigger." If an incorrect response was given by the participant (for example one) the following feedback and directions were given, "The bigger number is 7…..you should have said 7 because 7 is bigger than 4…..now look at this box (demonstrate by pointing)…..it has two numbers in it…..tell me the number that is bigger." If the correct response was given by the participant (for example two) the following feedback was given, "Good, the bigger number is 4." (Turn the page). If an incorrect response was given by the participant (for example two) the following feedback was given, "The bigger number is 4…..you should have said 4 because 4 is bigger than 2." Once the two examples were given, the directions for completing the two-item QD probe were given. "The paper in front of you has boxes on it. In the boxes are two numbers. When I say start, I want you to tell me the number in the box that is bigger. Start here and go across the page (demonstrate by pointing). If you come to a box and you don't know which number is bigger, I'll tell you what to do. Are there any questions? Put your finger on the first one. Ready, start."

Each research assistant administered and scored the two-item QD probe using the following administration and scoring guidelines. The three minute administration time

was started. If a participant failed to answer the first problem after 3 seconds, the participant was told to "try the next one." If the participant did not get any correct within the first 5 items, the administration was discontinued and a score of zero was recorded. All incorrect responses were recorded on the research assistant's copy of the QD probe. The maximum response time allowed for each item was 3 seconds; if a participant did not provide an answer within 3 seconds, the participant was told to "try the next one." After one, two, three minutes a bracket ( ]) was placed on the research assistant's copy to record how many items were completed at one, two, and three minutes. The probes were scored immediately, and participants received 1 point for every item correctly completed. If a participant stated the bigger number the item was scored as correct (all other responses, or if the participant was prompted to "try the next one," were scored as incorrect). If a participant skipped an item or a row of items, then the items skipped were omitted from scoring (in order to accurately calculate accuracy proportions for all items actually attempted).

The three-item probe type included 120 QD number sets and the participant was instructed to verbally choose the larger of a two number set presented visually. The administration directions for the three-item probe were similar to the directions designed by Clarke & Shinn (2002) for the two-item QD probe, with the following three exceptions; the administration time was altered from a one-minute to a three-minute administration, the participant was instructed to pick the middle number of a three number set, and the participant was given 15 seconds to respond to each item set. A participant copy of the three-item QD probe was placed in front of the student. A research

assistant's copy was placed on the examiner's clipboard and positioned so the participant could not see what the research assistant recorded. Each was administered two examples to ensure he/she understood the procedure for completing the probe. These specific directions were given to the participant (regarding the examples), "Look at the piece of paper in front of you…..the box in front of you has three numbers in it (demonstrate by pointing)…..I want you to tell me the number that goes in the middle of the other two numbers." If the correct response was given by the participant (for example one) the following feedback and directions were given, "Good…..the middle number is 7…..now look at this box (demonstrate by pointing)…..it has three numbers in it…..tell me the number that goes in the middle of the other two numbers." If an incorrect response was given by the participant (for example one) the following feedback and directions were given, "The middle number is 7…..you should have said 7 because 7 is bigger than 4 and smaller than 10…..now look at this box (demonstrate by pointing)…..it has three numbers in it…..tell me the number that goes in the middle of the other two numbers." If the correct response was given by the participant (for example two) the following feedback was given, "Good, the middle number is 4." If an incorrect response was given by the participant (for example two) the following feedback was given, "The middle number is 4…..you should have said 4 because 4 is bigger than 2 and smaller than 6." Once the two examples were given, the directions for completing the two-item QD probe were given. "The paper in front of you has boxes on it. In the boxes are three numbers. When I say start, I want you to tell me the number in the box that goes in the middle of the other two numbers. Start here and go across the page (demonstrate by pointing). If

you come to a box and you don't know which number goes in the middle, I'll tell you what to do. Are there any questions? Put your finger on the first one. Ready, start."

Each research assistant administered and scored the three-item QD probe using the following administration and scoring guidelines. The three minute administration time was started. The maximum response time allowed for each item was 15 seconds; if a participant did not provide an answer within 15 seconds, the participant was told to "try the next one." If the participant did not get any correct within the first 5 items, the administration was discontinued and a score of zero was recorded. All incorrect responses were recorded on the research assistant's copy of the QD probe. After one, two, three minutes a bracket ( ]) was placed on the research assistant's copy to record how many items were completed at one, two, and three minutes. The probes were scored immediately, and participants received 1 point for every item correctly completed. If a participant stated the middle number the item was scored as correct (all other responses, or if the participant was prompted to "try the next one," were scored as incorrect). If a participant skipped an item or a row of items, then the items skipped were omitted from scoring (in order to accurately calculate accuracy promotions for all items actually attempted).

The QD probes for both groups (the two-item and three-item) were scored and investigated using six different procedures (a combination of three different administration times and two scoring methods). One scoring procedure counted the number of correctly completed QD number sets in one minute (for the first minute of the QD probe administration) to arrive at the digits correct per 1 minute (DC1M) score

(measuring fluency). The second scoring procedure counted the number of correctly completed QD number sets in one minute (for the first minute of the QD probe administration), and that number was multiplied by the proportion of number sets attempted that are answered correctly (measuring fluency while weighting the score for accuracy; DC1M*A). The third scoring procedure counted the number of correctly completed QD number sets in two minutes (for the first two minutes of the QD probe administration) to arrive at the digits correct per 2 minutes (DC2M) score (measuring fluency). The fourth scoring procedure counted the number of correctly completed QD number sets in two minutes (for the first two minutes of the QD probe administration), and that number was multiplied by the proportion of number sets attempted that are answered correctly (measuring fluency while weighting the score for accuracy; DC2M*A). The fifth scoring procedure counted the number of correctly completed QD number sets in three minutes to arrive at the digits correct per 3 minutes (DC3M) score (measuring fluency). The sixth scoring procedure counted the number of correctly completed QD number sets in three minutes, and that number was multiplied by the proportion of number sets attempted that are answered correctly (measuring fluency while weighting the score for accuracy; DC3M*A).

The NKT was administered using the standardized instructions designed by Okamoto & Case (1996). Participant's answers were recorded by research assistants on provided answer sheets, and were scored by circling the problem number of all that were answered correctly (for problem numbers that had two parts, both parts had to be correct in order for the problem number to be counted correct). Correct answers were provided

on the answer sheets for the research assistants to accurately score answers. All participants were started at Level 0. Administration continued to the next level until the student had not answered sufficient items at any level to progress to the next level (the criterion for each level was in bold type next to the corresponding level that was being administered). The standardized instructions were read to each participant. "I will be asking you some questions about numbers. The questions will be easy at first, but they will get harder and harder. I do not expect you to know all of the answers. Some of them are even challenging for older children to get right!" The overall score for each participant consisted of the number of items that were answered correctly for all of the levels administered.

CHAPTER IV

FINDINGS

Three independent variables were investigated. One independent variable was a with-in group variable, which was administration time (1 minute, 2 minutes, and 3 minutes). A second independent variable was a between groups variable, which was probe construction (choosing the larger of a two numbers, and choosing the middle number of a three item set). The third independent variable was a with-in group variable, which was scoring method (two methods used). The first scoring method was fluency, digits correct per (x) minutes (DCxM). The second scoring method was fluency while weighting for accuracy, digits correct per (x) minutes while weighted for accuracy (DCxM*A). The dependent variable was the correlation between QD performance at the different levels and performance on the criterion which was the Number Knowledge Test (NKT).

Before calculating the statistics, three participant scores were removed due to withdrawal from the study or indication of an invalid administration. Originally 46 participants were assigned to the two item set probe construction group, and 48 participants were assigned to the three item set probe construction group. One

participant in the two item set probe construction group and one participant in the three item set probe construction group withdrew from the study. One participant's total number of QD sets administered in the three item set probe construction group was indicative that he or she was allowed more than 15 seconds to complete each set; therefore, this participant's data was removed. The final two item set QD probe construction group n = 45 and the three item set QD probe construction group n = 46. This was within the recommended participant range of 30 to 500 participants for a criterion-related validation study to have real world meaning (Roscoe, 1975).

Statistical analyses chosen were used to investigate concurrent validity and compare concurrent validity coefficients between the two QD probe construction groups. The Pearson-r correlation coefficient was used to calculate concurrent validity between QD probe administration and NKT administration (refer to Table 1 in Appendix A), and Microsoft Excel (2007) was used to calculate the correlations coefficients and means and standard deviations for each probe construction and procedure cell (refer to Table 2 in Appendix A). Fisher r-z transformation was calculated to evaluate whether the Pearson-r correlation coefficient (concurrent validity) of the original QD measure (two item set QD probe construction, with a one minute administration time, and scoring for fluency) was significantly improved statistically from the Pearson-r correlation coefficients (concurrent validity) of the three-item set probe construction group at the different levels of administration time and scoring method levels (refer to Table 1 in Appendix A). The Fisher r-z transformation was a one-tailed test, the alpha level for statistical significance was set at p = .05., and Lowry's VasserStats: Website for Statistical Computation (2001-

2014) was used to calculate the Fisher r-z transformation. Statistical results listed below are organized by hypothesis.

Administration Time on the Two Item Set Probe When Scoring for Fluency

First, did changing the administration time from one minute to two or three minutes improve the validity of the original QD measure? No, the Pearson-r correlation coefficients (concurrent validity) were r = 0.66 at one minute, r = 0.62 at two minutes, and r = 0.66 at three minutes. The concurrent validity for the two minute administration was relatively the same as the one minute administration time on the two item set QD probe construction when scoring for fluency (DCM). The concurrent validity for the three minute administration was not different than the one minute administration time on the two item set QD probe construction when scoring for fluency (DCM).

Probe Construction at One Minute when Scoring for Fluency

Second, did changing the probe construction from finding the larger of two numbers to finding the middle number of a three number set improve the validity of the original QD measure? No, the Pearson-r correlation coefficients (concurrent validity) were r = 0.66 for the two item set probe construction, and r = 0.31 for the three item probe construction. The Fisher r-z calculation was z = - 2.18, p = .99. The concurrent validity for the three item set probe construction at one minute administration time scored for fluency (QD) was less than the two item set QD probe construction at the one minute administration time scored for fluency (DCM).

Scoring at One Minute on the Two Item Set Probe

Third, did changing the scoring of the QD measure from a purely fluency based measure (Digits Correct per Minute; DCM) to an outcome measure that takes into account fluency (DCM) while weighting the score for accuracy (DCM*A) improve the validity of the original QD measure? No, the Pearson-r correlation coefficients (concurrent validity) were r = 0.66 when scoring for fluency (DCM), and r = 0.66 when scoring for fluency weighted for accuracy (DCM*A). The concurrent validity when scoring for fluency (DCM) was not different than when scoring for fluency weighted for accuracy (DCM*A) at the one minute administration time on the two item set QD probe construction.

Administration Time and Scoring on the Two Item Set Probe

Fourth, did the combination of changing administration time and changing the scoring method improve the validity of the original QD measure? No, the Pearson-r correlation coefficients (concurrent validity) were r = 0.66 at one minute administration time on the two item set probe construction when scoring for fluency (DCM), r = 0.63 at two minutes and scoring for fluency while weighting for accuracy (DCM*A), and r = 0.67 at three minutes and scoring for fluency while weighting for accuracy (DCM*A). The concurrent validity at the two and three minute administration times when scoring for fluency while weighting for accuracy (DCM*A) was relatively the same as the one minute administration time on the two item set QD probe construction when scoring for fluency (DCM).

Administration Time and the Three Item Set Probe

Fifth, did the combination of changing administration time and changing probe construction improve the validity of the original QD measure? No, the Pearson-r correlation coefficients (concurrent validity) were r = 0.66 at one minute administration time on the two item set probe construction when scoring for fluency (DCM), r = 0.31 at a two minute administration time on the three item set probe, and r = 0.33 at a three minute administration time on the three item set probe. The Fisher r-z calculations were z = - 2.18, p = .99 between a two minute administration time on the three item set probe and the original QD measure, and z = -2.07, p = .98 between a three minute administration time on the three item set probe and the original QD measure. The concurrent validity for the two and three minute administrations on the three item set probe when scoring for fluency (DCM) was less than the one minute administration time on the two item set QD probe construction when scoring for fluency (DCM).

Three Item Set Probe and Weighting Scoring for Accuracy

Sixth, did the combination of changing probe construction and changing the scoring method improve the validity of the original QD measure? No, the Pearson-r correlation coefficients (concurrent validity) were r = 0.66 at one minute administration time on the two item set probe construction when scoring for fluency (DCM), and r = 0.34 on the three item probe set when scoring for fluency while weighting for accuracy (DCM*A). The Fisher r-z calculation was z = - 2.02, p = .98. The concurrent validity at the one minute administration time on the three item probe set when scoring for fluency

while weighting for accuracy (DCM*A) was less than the one minute administration time on the two item set QD probe construction when scoring for fluency (DCM).

Administration Time, Three Item Set Probe, and Weighting Scoring for Accuracy

Seventh, did the combination of changing the administration time, changing probe construction, and changing the scoring method improve the validity of the original QD measure? No, the Pearson-r correlation coefficients (concurrent validity) were r = 0.66 at one minute administration time on the two item set probe construction when scoring for fluency (DCM), r = 0.40 at two minutes administration time on the three item set probe construction when scoring for fluency while weighting for accuracy (DCM*A), and r = 0.43 at three minutes administration time on the three item set probe construction when scoring for fluency while weighting for accuracy (DCM*A). The Fisher r-z calculations were z = - 1.70, p = .96 between the two minutes administration time on the three item set probe construction when scoring for fluency while weighting for accuracy (DCM*A) and the original QD measure, and z = -1.53, p = .94 between a three minute administration time on the three item set probe construction when scoring for fluency while weighting for accuracy (DCM*A) and the original QD measure. The concurrent validity at the two and three minutes administration times on the three item set probe construction when scoring for fluency while weighting for accuracy (DCM*A) was less than the original one minute administration time on the two item set QD probe construction when scoring for fluency (DCM).

Post-Analysis Data

The mean percentage correct (accuracy rate) on the two item set QD probe construction and three item set QD probe construction at one, two, and three minutes of administration time were calculated, since accuracy rates appeared less on the three item set QD probe construction. On the two item set QD probe construction the mean percentage correct ranged from 97% to 97.4% for the 1, 2, and 3 minute administration times. On the three item set QD probe construction the mean percentage correct ranged from 51% to 53% for the 1, 2, and 3 minute administration times. Mean accuracy rates for both the two item set QD probe construction and three item set QD probe construction were better than chance, but accuracy rate was less on the three item set QD probe construction than on the two item set QD probe construction. A floor effect was also noted on the three item set QD probe construction (refer to Table 2 in Appendix A for means and standard deviations on the three item set QD probe construction).

On the three item set QD probe construction where the mean percentage correct was less, the scoring method that scored for fluency while weighting for accuracy (DCM*A) demonstrated larger Pearson-r correlation coefficients than the fluency scoring method (DCM). This did not improve the Pearson-r correlation coefficients to the point that overall concurrent validity was improved over the original QD measure, but did demonstrate that weighting fluency scores for accuracy can increase Pearson-r correlation coefficients when the mean percent correct is not relatively 100%. This was most prevalent at the two and three minute administration times. At two minutes the fluency scoring method (DCM) $r = 0.31$, and the fluency while weighting for accuracy scoring

method (DCM*A) r = 0.40. At three minutes the fluency scoring method (DCM) r = 0.33, and the fluency while weighting for accuracy scoring method (DCM*A) r = 0.43. Little difference was noted at the one minute administration time where the fluency scoring method (DCM) r = 0.31, and the fluency while weighting for accuracy scoring method (DCM*A) r = 0.34.

It was noted also that simple overall accuracy percentage also demonstrated that on the three item set QD probe construction where the mean percentage correct was less, Pearson-r correlation coefficients were larger than the fluency scoring method (DCM). This also did not improve the Pearson-r correlation coefficients to the point that overall concurrent validity was improved over the original QD measure, but did demonstrate that scoring simply for overall accuracy percentage instead of fluency scores can increase Pearson-r correlation coefficients when the mean percent correct is not relatively 100%. This was noted at the one, two, and three minute administration times. At one minute the fluency scoring method (DCM) r = 0.31, and the simple overall accuracy scoring r = 0.40. At two minutes the fluency scoring method (DCM) r = 0.31, and the simple overall accuracy scoring r = 0.43. At three minutes the fluency scoring method (DCM) r = 0.33, and the simple overall accuracy scoring r = 0.43.

CHAPTER V

CONCLUSION

Children come to school with a number of informal math skills (Gersten et. al., 2005; National Mathematics Advisory Panel, 2008). These informal skills include counting, the ability to identify numbers, the ability to formulate mental number lines, and the ability to discriminate between quantities (Clements, Sarama, & DiBiase, 2004). Tests of Early Numeracy (TENs), QD being one, are robust indicators that are based on numeracy skills and can be used to help educators identify students potentially at-risk for later failure in formal mathematical computation (Clarke & Shinn, 2002). Changing QD probe construction from finding the larger of two numbers to finding the middle number of a three item set was suggested for this study, because it requires students to identify numbers, formulate mental number lines, and discriminate between quantities which include three of the four informal math skills that children come to school with. If a broader range of informal math skills could be investigated within one measure, then it is possible that the validity of that measure could be increased. Whether administration time

and scoring method could impact the validity was questioned as well in this study. The QD probes were scored for fluency and fluency weighted for accuracy. Based on the original three research questions seven hypotheses were developed and a summary of the seven findings are discussed within this chapter.

Summary and Discussion of Findings

First finding, changing the administration time from one minute to two or three minutes did not improve the validity of the original QD measure. The literature did not reflect whether lengthening administration time could improve the concurrent validity of the QD measure. Jiban and Deno (2007), however, investigated the influence of aggregating administration scores on predictive validity in a study investigating if cloze math facts were as equally reliable as basic math facts. The scores on the one minute cloze math measure were evaluated individually and aggregated (two administrations equaling two minutes), and the increased time increased correlation coefficients from .65 to .86 among a fifth grade sample (criterion was *Minnesota Comprehensive Assessment in Mathematics*, 2007*).* On reading measures Williams et al. (2011) found reading speed accounted for the greatest variance in most reading measures, and it was questioned whether this could be a factor regarding early numeracy measures as well and required further investigation. If this was true, it was questioned whether speed of identification could be better captured with an increased administration time. Administration time had no effect on the validity of the original QD measure. Concurrent validity was not improved with increased administration time on the original two item set QD probe

construction when scored for fluency, and the original one minute administration time was demonstrated as sufficient.

Second finding, changing the probe construction from finding the larger of two numbers to finding the middle number of a three number set did not improve the validity of the original QD measure. The literature did not reflect whether adding numbers to the QD set could improve the measure. Adding numbers (changing from two to three for this study) to each QD probe set does lower the theoretical probability of choosing the correct number by a function of guessing, but the accuracy rate on the three item set QD probe construction was much less than the accuracy rate on the two item set QD probe construction. The concurrent validity on the three item set QD probe construction was also less than the original two item set QD probe construction, and the three item set QD probe construction was more difficult for students to identify the target number correctly. Concurrent validity was not improved by changing the two item set probe construction to a three item set QD probe construction at the original one minute administration time when scored for fluency.

Third finding, changing the scoring of the QD measure from a purely fluency based measure (Digits Correct per Minute; DCM) to an outcome measure that takes into account fluency while weighting the score for accuracy (DCM*A) did not improve the validity of the original QD measure. The literature did not reflect whether weighting a fluency score for accuracy would improve the measure. It was questioned whether or not weighting fluency scores for accuracy might improve the measure and help rank order students in a more valid manner. A student with a fluency score of 20 DCM who is 100%

accurate should be ranked differently than a student with a fluency score of 20 DCM who is 50% accurate, which is the reasoning why weighting for accuracy might have improved the validity and rank order students more effectively. However, concurrent validity was not improved on the original two item set QD probe construction with a one minute administration time by changing the scoring method from scoring for fluency (DCM) to fluency weighted for accuracy (DCM*A).

Fourth finding, the combination of changing administration time and changing the scoring method did not improve the validity of the original QD measure. There was no relatively no change in validity due to an interaction of these changes. Concurrent validity was not improved on the original two item set QD probe construction by increasing the administration time and changing the scoring method from scoring for fluency (DCM) to fluency weighted for accuracy (DCM*A).

Fifth finding, the combination of changing administration time and changing probe construction did not improve the validity of the original QD measure. The concurrent validity on the three item set QD probe construction at two and three minutes administration time scored for fluency was rather less than the original two item set QD probe construction at one minute administration scored for fluency. Concurrent validity was not improved by increasing the administration time and changing the original two item set probe construction to a three item set QD probe construction.

Sixth finding, the combination of changing probe construction and changing the scoring method did not improve the validity of the original QD measure. The concurrent validity on the three item set QD probe construction scored for fluency while weighting

for accuracy (DCM*A) was rather less than the original two item set QD probe construction at one minute administration scored for fluency. Concurrent validity was not improved, at the original one minute administration time, by changing the original two item set QD to a three item set QD probe construction while also changing the scoring method from scoring for fluency (DCM) to fluency weighted for accuracy (DCM*A).

Seventh finding, the combination of changing the administration time, changing probe construction, and changing the scoring method did not improve the validity of the original QD measure. The concurrent validity on the three item set QD probe construction at two and three minutes administration time scored for fluency while weighting for accuracy (DCM*A) was rather less than the original two item set QD probe construction at one minute administration scored for fluency. Concurrent validity was not improved, compared to the original QD construction and procedures, on the three item set QD probe construction when the administration time was increased and the QD probe was scored for fluency weighted for accuracy (DCM*A).

General Implications of Findings

What does this mean to researchers and educators? This section of the chapter will review theoretical, research, and applied implications of the findings of this study on early numeracy CBM screening and CBM screening in general.

Theoretical Implications

Fluency could be more important than the overall QD measure itself, or in other words processing speed may be a robust indicator of math performance. Williams et al. (2011) demonstrated that reading speed accounted for the most variance in reading

measures. In the current study when accuracy rates were high, as they were on the two item set QD probe construction, weighting fluency scoring for accuracy did not affect the overall concurrent validity. When accuracy rates are high, processing speed (which affects fluency) could be a robust indicator of math performance. This requires further investigation. In contrast when accuracy rates were low, as they were on the three item set QD probe construction in the current study, weighting the fluency score for accuracy did increase the concurrent validity. When accuracy rates are lower, speed of responding may be less of a predictor. If so, then the accuracy rates of responses could affect whether or not speed of responding could be used as a robust indicator of math performance.

Research Implications

On measures where mean percentage correct is lower (no matter subject domain) could weighting fluency scores for accuracy increase concurrent validity? In the current study when the mean percentage correct was lower, weighting fluency scores did increase concurrent validity. Future investigation whether or not validity could be significantly improved on CBM measures (no matter subject domain) that have lower mean percentage rates (such as Mathematical-CBM) is recommended.

On measures where mean percentage correct is lower (no matter subject domain), is overall accuracy percentage more important than fluency and a better robust indicator? In the current study on the three item set QD probe construction where accuracy rates were lower, the simple accuracy rate increased the concurrent validity as much as fluency scoring weighted for accuracy. More investigation into the overall role simple accuracy rate plays in measures where overall accuracy rates are lower could be beneficial.

Applied Implications

There is no current single measure for early numeracy screening, including QD, which accounts for enough of the variance on the NKT to use that one measure as an independent screening measure for early numeracy. Educators need to continue to use a combination of several CBM measures to adequately screen for early numeracy. The concurrent validity Pearson-r correlation coefficient in the current study ($r = 0.66$) for the original QD measure was moderate in strength, according to Gay, Mills, & Airasian (2006), which did not explain enough of the variance (coefficient of determination = 0.44) on the NKT to be used as a single screener for early numeracy. These findings were within the range of previous findings. Clarke and Shinn (2004) concluded the concurrent validity with the NKT on the original QD developed and investigated in that study was $r = .80$ for first grade students, which was increased in comparison to the results of the current study. Chard et al. (2005) found that the correlation between the NKT and QD for first grade students was 0.45 in the fall and 0.53 in the spring, which was decreased in comparison to the results of the current study. Concurrent validity coefficients on the original QD measure are not strong enough between the QD measure and NKT in the current study or previous studies to use QD as a single screening measure for early numeracy, and the QD concurrent validity with the NKT could not be improved with any of the hypothesized modifications to probe construction or administration and scoring procedures made in the current study.

General Limitation of Study

One design and internal validity factor of note was the overall difficulty of the three item set QD probe for numbers 1-20. It was desired that the overall mean accuracy rates would be lower than what they are on the two item set QD probe; however, the mean accuracy rates were not anticipated to be lower than 75%. The mean accuracy rates were not expected to be 51 to 53% as they were during the current study, and a floor effect was noted as well. The mean accuracy rates were better than chance (33%), but improving mean accuracy rates by decreasing the difficulty of the three item set QD probe construction, for example by using numbers 1-10 instead of 1-20, may have improved the overall mean accuracy rates and improved concurrent validity with the NKT for the three item set probe construction.

Future Directions

So where do we go from here with the QD measure if we cannot improve the concurrent validity of the measure and are forced to administer three to four TENs at a time to screen adequately? Two options are suggested that may decrease the administration time, making the time to screen shorter and more manageable for educators. One, if concurrent validity is acceptable at one minute of administration, would it still be acceptable at 45 seconds? In other words, if increased administration time does not improve concurrent validity, then how short can we make administration time on the original QD measure without decreasing concurrent validity? This would save some administration time for educators if an administration time of less than one minute was still effective. Two, an option that would save even more administration time

would be if the QD measure could be group administered instead of individually administered. First graders can take group administered Mathematical-CBM probes for simple addition and subtraction, so why not attempt to group administer the QD measure and evaluate if concurrent validity remains unchanged? Whatever researchers and educators decide to do, the goal is to find the most reliable, valid, and efficient measures as possible to screen for early numeracy skills.

REFERENCES

Beatty, L.S., Gardner, E.G., Madden R., & Karlsen, B. (1985). *The Stanford Diagnostic Mathematics Test* (3rd ed.). San Antonio, TX: The Psychological Corporation.

Berch, D.B. (Ed.). (1998). *Mathematical cognition: From numerical thinking to mathematics education*. Conference presented by the National Institute of Child Health and Human Development, Bethesda, MD.

Brigance, A. (1999). *Comprehensive Inventory of Basic Skills* (rev. ed.). North Billerica, MA: Curriculum Associates Inc.

Case, R. (1998). *A psychological model of numbers sense and its development*. Paper presented at the annual meeting of the American Educational Research Association, San Diego.

Chard, D., Clarke, B., Baker, S., Otterstedt, J., Braun, D., & Katz R. (2005). Using measures of number sense to screen for difficulties in mathematics. *Assessment of Effective Intervention, 30*(2), 3-14.

Clarke, B. & Shinn, M.R. (2004). A preliminary investigation into the identification and development of early mathematics curriculum-based measurement. *School Psychology Review, 33* (2), 234-247.

Clarke, B., & Shinn, M.R. (2002). *Tests of Early Numeracy (TEN): Administration and scoring of AIMSweb early numeracy measures for use with AIMSweb*. Eden Prairie, MN: Edformation Inc.

Clements, D. H., Sarama, J., & DiBiase, M. (Eds.). (2004). *Engaging young children in mathematics: Standards for early childhood mathematics education*. Mahwah, NJ: Erlbaum.

CTB/McGraw-Hill. (1992). *California Achievement Tests* (5th ed.). Monterey, CA: CTB Macmillam/Mc-Graw-Hill.

Espin, C.A., Deno, S.L., Maruyama, G., & Cohen C. (1989). *The Basic Academic Skills Sample (BASS): An instrument for the screening and identification of children at risk for failure in regular education classrooms*. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA.

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

Fuchs, L.S., Fuchs, D., Compton, D.L., Bryant, J.D., Hamlett, C.L., & Seethaler, P.M. (2007). Mathematics screening and progress monitoring at first grade: Implications for responsiveness to intervention. *Exceptional Children,73(*3), 311-330.

Fuchs, L., Fuchs, D., Hamlett, C., & Walz, L. (1993). Formative evaluation of academic progress: How much growth can we expect? *School Psychology Review, 22*(1), 27-48.

Fuchs, L.S., Hamlett, C., L., & Fuchs, D. (1998). *Monitoring basic skills progress: Math computation*. Austin: TX: PRO-ED.

Fuchs, L.S., Hamlett, C., L., & Fuchs, D. (1999). *Monitoring basic skills progress: Concepts and applications*. Austin: TX: PRO-ED.

Gay, L.R., Mills, G.E., & Airasian, P. (2006). *Educational research: Competencies for analysis and applications* (8th ed.). Upper Saddle River, New Jersey: Pearson Education, Inc.

Gersten, R., & Chard, D. (1999). Number sense: Rethinking arithmetic instruction for students with mathematical disabilities. *The Journal of Special Education, 33*(1), 18-28.

Ginsberg, H.P. (1989). *Children's arithmetic: How they learn it and how you teach it* (2nd ed.). Austin, TX: PRO-ED.

Ginsburg, H.P, & Baroody, A.J. (1990). *Test of Early Mathematics Ability* (2nd ed.). Austin, TX: PRO-ED.

Hintze, J., Christ, T., & Keller, L. (2002). The generalizability of CBM survey-level mathematics assessments: Just how many samples do we need? *School Psychology Review, 31*(4), 514-528.

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.

Lembke, E.S., Foegen, A., Whittaker, T.A., & Hampton, D. (2008). Establishing technically adequate measures of progress in early numeracy. *Assessment for Effective Intervention, 33*(4), 206-214.

Lowry, R., (2001-2014). VasserStats: Website for Statistical Computation [Computer Software]. Vassar College, Poughkeepsie, NY. http://www.vassarstats.net/

Methe, S.A., & Riley-Tillman, T.C. (2008). An informed approach to selecting and designing early mathematics interventions. *School Psychology Forum: Research in Practice*. 2(3), 29-41.

Minnesota Department of Education (2007). *Minnesota Comprehensive Assessment in Mathematics.*

National Assessment of Educational Progress. (1992). *NAEP 1992 mathematics report card for the nation and the states* (report no. 23-ST02). Washington, DC: National Center for Educational Statistics.

National Council of Teachers of Mathematics. (1989). *Curriculum and evaluation standards for school mathematics*. Reston, VA: Author.

National Council of Teachers of Mathematics. (2006). *Curriculum focal points for prekindergarten through grade 8 mathematics: A quest for coherence.* Reston, VA: Author.

National Mathematics Advisory Panel. (2008). *Final report of the national mathematics advisory panel*. Retrieved June 2, 2008, from http://www.ed.gov/about/bdscomm/list/mathpanel/reports.html.

Okamoto, Y., & Case, R. (1996). Exploring the microstructure of children's central conceptual structures in the domain of number. In R. Case & Y. Okamoto (Eds.), *The Role of Central Conceptual Structures in the Development of Children's Thought: Monographs of the Society for Research in Child Development* (vol. 1-2). pp. 27-58. Malden, MA: Blackwell Publishers.

Psychological Corporation. (1996). *Stanford Early School Achievement Test* (9th ed.). San Antonio, TX: Harcourt Assessment.

Rivera, D.P. (1997) Mathematics education and students with learning disabilities: Introduction to the special series. *Journal of Learning Disabilities 30*(1), 2-19, 68.

Roscoe, J.T. (1975). *Fundamental research statistics for the behavioral sciences* (2nd edition). New York: Holt Reinhardt and Winston.

Shapiro, E.S. (1996). *Academic skills problems workbook*. New York: Guilford Press.

Shapiro, E.S., Edwards, L., & Zigmond, N. (2005). Progress monitoring of mathematics among students with learning disabilities. *Assessment for Effective Intervention, 30*(2), 15-3.

Thurber, R., Shinn, M., & Smolkowski, K. (2002). What is measured in mathematics tests? Construct validity of curriculum-based mathematics measures. *School Psychology Review, 31*(4), 498-513.

U.S. Department of Education. (2008). *The final report of the national mathematics advisory panel*.

VanDerHeyden, A. M., Broussard, C., & Cooley, A. (2006). Further development of measures of early math performance for preschoolers. *Journal of School Psychology, 44*, 533-553.

VanDerHeyden, A. M., Broussard, C., Fabre, M., Stanley, J. L., & Creppell, R. (2004). Development and validation of curriculum-based measures of math performance for preschool children. *Journal of Early Intervention, 27*(1), 27-41.

VanDerHeyden, A.M., Witt, J. C., Naquin, G., & Noell, G. (2001). The reliability and validity of curriculum-based measurement readiness probes for kindergarten students. *School Psychology Review, (30)*3, 363-382.

Williams, J.L., Skinner, C.H., Floyd, R.G., Hale, A.D., Neddenriep, C., Kirk, E. (2011). The variance in standardized reading scores accounted for by reading speed. *Psychology in the Schools 48*(2), 87-101.

Woodcock, R.M., & Johnson, M.B. (1989). *Woodcock-Johnson Psycho-Educational Battery* (revised). Allen, TX: DLM Teaching Resources.

APPENDICES

Appendix A

Table 1

Concurrent Validity between QD Measure and NKT

Administration time & scoring method	Two item set QD probe n = 45	Three item set QD probe n = 46	Fisher r-z comparing three item set QD probe r to original QD measure r = 0.66*
1 minute & scored for fluency	r = 0.66*	r = 0.31	z = -2.18 p = .99
1 minute & fluency score weighted for accuracy	r = 0.66	r = 0.34	z = -2.02 p = .98
2 minute & scored for fluency	r = 0.62	r = 0.31	z = -2.18 p = .99
2 minute & fluency score weighted for accuracy	r = 0.63	r = 0.40	z = -1.70 p = .96
3 minute & scored for fluency	r = 0.66	r = 0.33	z = -2.07 p = .98
3 minute & fluency score weighted for accuracy	r = 0.67	r = 0.43	z = -1.53 p = .94

*Original QD measure

Table 2

Means and Standardized Deviations of QD Probe Construction and Procedure

Administration time & scoring method	Two item set QD probe n = 45	Three item set QD probe n = 46
1 minute & scored for fluency	Mean = 33 SD = 10	Mean = 7 SD = 4
1 minute & fluency score weighted for accuracy	Mean = 32 SD = 11	Mean = 4 SD = 4
2 minute & scored for fluency	Mean = 64 SD = 19	Mean = 11 SD = 7
2 minute & fluency score weighted for accuracy	Mean = 63 SD = 20	Mean = 7 SD = 7
3 minute & scored for fluency	Mean = 93 SD = 24	Mean = 16 SD = 10
3 minute & fluency score weighted for accuracy	Mean = 91 SD = 25	Mean = 10 SD = 10

Appendix B

Original two item set QD probe construction example

7	10
6	9
18	10

3	2
6	19
13	14

17	8
15	9
3	15

Three item set QD probe construction example

17	9	5
16	4	20
16	10	15

17	9	12
7	12	8
4	19	16

3	6	8
17	19	13
10	1	14

VITA

Michael Hoffman

Candidate for the Degree of

Doctor of Philosophy

Thesis: INVESTIGATION INTO THE VALIDITY OF THE QUANTITY DISCRIMINATION CURRICULUM-BASED MEASURE OF EARLY NUMERACY

Major Field: Educational Psychology (School Psychology)

Biographical:

Education:

Completed the requirements for the Doctor of Philosophy in Educational Psychology at Oklahoma State University, Stillwater, Oklahoma in May, 2014.

Completed the requirements for the Master of Science in Educational Psychology at Oklahoma State University, Stillwater, Oklahoma in December, 2008.

Completed the requirements for the Bachelor of Science in Psychology at Oklahoma State University, Stillwater, Oklahoma in December, 2006.

Experience:

July 2013 to June 2014 – Texas A&M University Health Science Center at Scott & White Healthcare, Temple, Tx.; APICC (program) member - Clinical Child/Adolescent Psychology Resident (pre-doctoral internship)

August 2010 to May 2011 – Behavioral Solutions, Stillwater, Ok.; RtI Site Visitor/Consultant

November 2009 to Present – University of Phoenix; Online Main Campus; College of Social Sciences – Associate Faculty

Professional Memberships:

American Psychological Association

National Association of School Psychologists