2010-06-11

Population Invariance of Linking Functions of Curriculum-Based Measures of Math Problem Solving

Jia Huang
University of Miami, jiahuang.cn@gmail.com

UNIVERSITY OF MIAMI

POPULATION INVARIANCE OF LINKING FUNCTIONS ACROSS ALTERNATE FORMS OF CURRICULUM-BASED MEASURES OF MATH PROBLEM SOLVING

By

Jia Huang

A DISSERTATION

Submitted to the Faculty
of the University of Miami
in partial fulfillment of the requirements for
the degree of Doctor of Philosophy

Coral Gables, Florida

June 2010

UNIVERSITY OF MIAMI

A dissertation submitted in partial fulfillment of
the requirements for the degree of
Doctor of Philosophy

POPULATION INVARIANCE OF LINKING FUNCTIONS ACROSS ALTERNATE FORMS OF CURRICULUM-BASED MEASURES OF MATH PROBLEM SOLVING

Jia Huang

Approved:

_____ Marjorie Montague, Ph.D. Professor of Teaching and Learning	_____ Terri A. Scandura, Ph.D. Dean of the Graduate School
_____ Batya Elbaum, Ph.D. Associate Professor of Teaching and Learning	_____ Wendy M. Cavendish, Ph.D. Assistant Professor of Teaching and Learning
_____ Randall Penfield, Ph.D. Professor of Educational and Psychological Studies	

HUANG, JIA (Ph.D., Teaching and Learning)
Population Invariance of Alternate Forms of Curriculum-Based Measures Of Math Problem Solving (June 2010)

Abstract of a dissertation at the University of Miami.

Dissertation supervised by Professor Marjorie Montague.
No. of pages in text. (83)

The purpose of this study was to investigate population invariance of the true-score linking functions with respect to the ability subgroups (i.e., average-achieving students, low-achieving students, and students with learning disabilities). The mean/mean linking functions for five alternate forms of a curriculum-based math problem solving measure were based on the Rasch model. Most studies of curriculum-based measurement have reported only the reliability and validity of alternate forms of measures. This is necessary but insufficient for establishing alternate forms of curriculum-based measures. It is also necessary to establish equivalency of the forms. The present study was based on data from a previous study that developed equivalent forms of curriculum-based measures using Item Response Theory. The participants in the present study were 1,861 seventh- and eighth-grade students. Equatability indices were used to evaluate population invariance of the Rasch mean/mean linking functions over the ability subgroups. Results indicated that the Rasch mean/mean linking functions were population invariant for the ability subgroups across the five alternate forms. The differences between the linking functions computed on the ability subgroups and the linking function on the whole group were negligible for the five forms. Several implications and recommendations for future

studies on population invariance of the linking functions with alternate forms of curriculum-based measures were discussed.

DEDICATION

I would like to dedicate this Doctoral dissertation

to my dear parents, who have helped so much with baby-sitting and have given me their unwavering support and unselfish love;

to my considerate husband, who has provided me the greatest encouragement and fullest love;

to our precious son, who is the joy of our lives.

ACKNOWLEDGEMENT

I would have never been able to complete my dissertation without the guidance of my committee members, help from my colleagues and friends, and support from my family members.

I would like to express my deepest appreciation to my dissertation advisor, Dr. Marjorie Montague, for her helpful criticism, careful reading of countless drafts, and valuable suggestions. Without her continuous guidance and persistent help, this dissertation would not have been possible. I would like to extend a special thank you to Dr. Randy Penfield for his expert statistical knowledge and good patience to help me with my methodology. Many thanks must go to Dr. Batya Elbaum and Dr. Wendy Morrison-Cavendish for their thoughtful suggestions, unflinching encouragement and support along the way.

I would like to thank my colleagues who helped collect data and manage the database for this study. I thank all of my friends who read and reread my dissertation and provided useful feedback.

Last, but certainly not least, I am most grateful to my parents for their unconditional love and endless support. Their faith in me has been essential to my success in this doctoral program and in all of my other endeavors. I am also extremely grateful to my husband Zhigang, who has inspired and encouraged me throughout this entire process and to my son Lucas, who was born at the beginning of this doctoral program and spent much time at daycare to allow me to focus. I am deeply sorry for the time we spent apart.

TABLE OF CONTENTS

LIST OF FIGURES

Page

LIST OF TABLES

Page

Chapter One

Introduction

Curriculum-based measures have been used widely in educational decision-making for the purposes of screening, identifying, and referring students at risk for academic failure; gauging students' responsiveness to intervention; evaluating the efficacy of instructions; making instructional decisions; and predicting students' achievement on high-stakes assessments (Deno & Fuchs, 1987; Fuchs & Deno, 1992; Deno, 2003a; Deno, 2003b; Fuchs, 2004; Madelaine & Wheldall, 1999; Meherns & Clarizio, 1993). The majority of curriculum-based measurement (CBM) research studies have evaluated only the technical adequacy of alternate forms of curriculum-based measures (i.e., validity and reliability) rather than their equivalency. Alternate forms of curriculum-based measures must be equated if scores of students are to be compared over time. Researchers in CBM have endeavored to construct equivalent forms of tests, but the forms, in practice, generally rely only on raw scores and thus have inevitable differences in difficulty. Raw scores do not have intrinsic normative meaning. Using only raw scores will result in misinterpretations of an examinee's underlying ability if one examinee takes a more difficult form of a test than that taken by another (Montague, Penfield, Enders, & Huang, 2010).

The purpose of this study was to investigate whether the Rasch mean/mean linking functions were population invariant for the subgroups that differed in ability. Data from five alternate forms of a math problem solving measure were used to determine population invariance of the linking functions with respect to the average-achieving (AA) students, low-achieving (LA) students, and students with learning disabilities (LD). Math

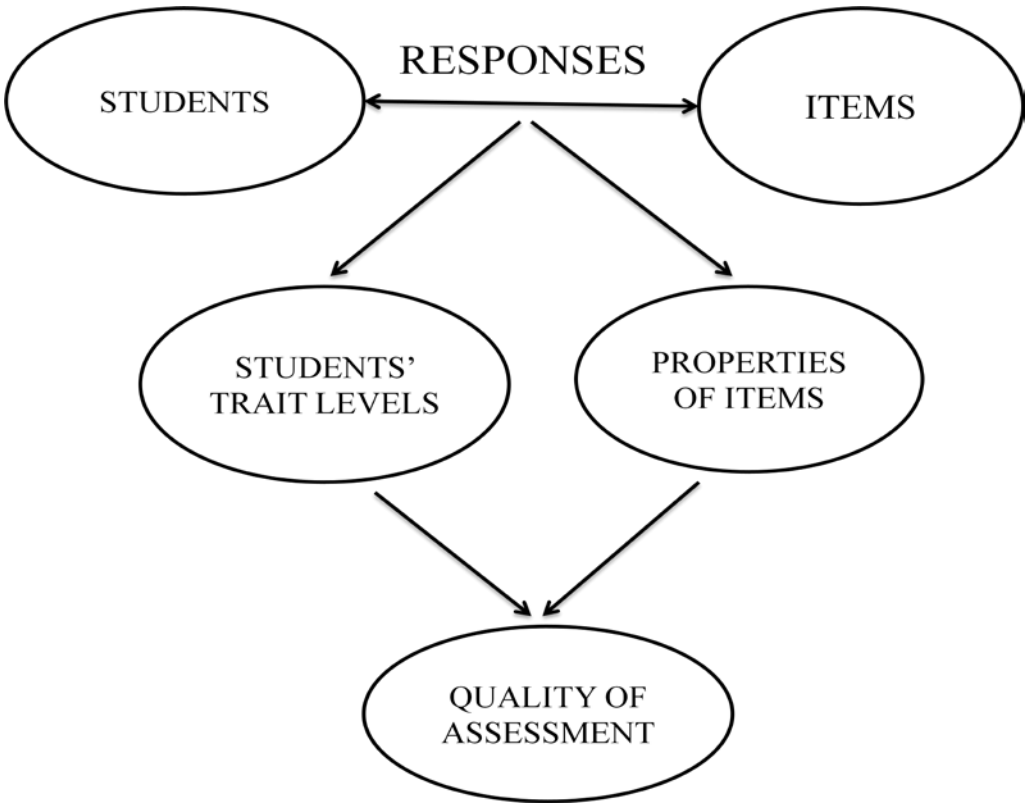
problem solving was selected as the content focus because this academic area is critically important to success in school and, consequently, there is a need to document accurately student progress in mathematics over time. This chapter provides an overview of educational assessment and test theories and also discusses the importance of the study in the context of today's educational environment.

Educational Assessment

The *Standards for Educational and Psychological Testing* (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1999) defines assessment as a systematic process used to obtain information about the characteristics of people or objects and then to make inferences. It is an extensive and complex procedure and helps to understand these characteristics (Reynolds, Livingston, & Willson, 2009). Assessment plays a substantial role in educational settings. One purpose of educational assessment is to observe and measure student learning. These measurements provide teachers with a better understanding of what students are learning in order to help students engage more deeply in the process of learning in a particular domain. Three metaphors are used to understand and describe the concept of educational assessments: "sitting beside, judging, and coaching" (Musial, Nieminen, Thomas, & Burke, 2009, p. 5). The metaphor of "sitting beside" indicates that assessments are information-collecting procedures in which teachers can have knowledge of what students need by interacting with them. In the "judging" metaphor, teachers can gain an understanding of what students have mastered. The "coaching" metaphor implies that when students have some difficulty, teachers can help them overcome it and reach a specific goal (Musial et al., 2009). There are two main

types of educational assessment, summative assessment and formative assessment. Summative assessment is used to determine the value of an outcome. In contrast, formative assessment is employed to provide specific feedback to students so that they can tell what they have learned. The specific feedback can also help teachers adjust their instruction to meet student needs (Musial et al., 2009; Reynolds et al., 2009). Educational assessment can be regarded as a two-way street of measurement (see Figure 1). Students' responses to the items of an instrument or test not only tell their ability levels but also indicate properties of the items and the instrument as a whole. For example, on the one hand, a student who does not correctly answer many items of a test may have a low ability level; on the other hand, an item for which only a few students correctly answer may be a very difficult item. In doing educational assessment, it is critical to consider both the ability levels of students being measured as well as the properties of items used in the measurement process.

Figure 1. Educational Assessment Model.



Validity. The *Standards* (AERA et al., 1999) defines validity as "the degree to which evidence and theory support the interpretations of test scores entailed by proposed uses of test" (p. 9). This definition indicates that validity is about making decisions by using a test's scores. A valid score is a score for which test developers have evidence indicating that it allows the test to do what it is supposed to do (Osterlind, 2006). Osterlind (2006) argued that validity is a property of scores rather than an instrument or test because the scores of an instrument may be valid for one population but not another. For example, a valid score on the Florida Comprehensive Assessment Test (FCAT) is a score for which evidence suggests that it allows teachers to have information about student mastery of skills. The *Standards* (AERA et al., 1999) also addresses validity as "the most fundamental consideration in developing and evaluating tests" (p. 9). Various sources of evidence can interpret different aspects of validity rather than different types of validity because "validity is a unitary concept" (p. 11). These sources of evidence can be used in evaluating the intended interpretation of test scores for specific purposes.

First of all, one of the simplest forms of validity evidence is evidence based on test content that concerns how well the content of the items of an instrument or test matches the content domain intended to be measured by the items. It is also referred to as item content-relevance validity because it is related to the content of items. It has implications for the properties (i.e., content, relevance, and meaning) of the items themselves. Second, validity evidence based on response processes can be obtained from an analysis of the fit between the responses actually engaged in by examinees and the construct a test is intended to measure. When examinees take a test, they go through mental processes so as to give their responses to the items. The examiner assumes that

their processes reflect what the test is intended to measure. For example, if a test is to measure mathematical problem solving ability, then examinees' mental processes should reflect problem solving strategies rather than remembering formula. Third, validity evidence based on internal structure refers to the degree to which the relationships among test items and test components match with the construct underlying the proposed test interpretations. That is to say, one group of test items may rise and fall together in one pattern, and another group of test items may rise and fall in a different pattern, or all of the items rise and fall together. Validity evidence based on internal structure can be obtained through mathematical techniques such as structural equation modeling, factor analysis, and differential item functioning. These techniques all help discover how many things a test measures. Last, validity evidence based on relations to other variables involves test scores' relation to other external criterion variables. Relationships between test scores and other tests of similar constructs contribute to convergent evidence, relationships between test scores and other tests of different constructs provide discriminant evidence, and relationships between test scores and a relevant criterion reflect criterion evidence. For example, when graduate schools recruit students into graduate programs, they usually consider the students' scores on the Graduate Record Examination (GRE) as one of the requirements. The criterion of interest is students' academic performance while obtaining a Master's degree, which is measured by grade point average (GPA). If the scores on the GRE are unrelated to GPA, then the scores will not provide any criterion-related evidence of validity.

Reliability. The *Standards* (AERA et al., 1999) defines reliability as "the consistency of such measurements when the testing procedure is repeated on a population

of individuals or groups" (p. 25). The definition indicates that individuals or groups should demonstrate some degree of stability in their behaviors when they take a measurement multiple times. However, the same person cannot exhibit identical behaviors on different occasions, so there are variations in the person's scores. These variations reflect measurement errors that are usually random and unforeseeable. Understanding the role of measurement errors is crucial to proper data interpretation. Reliability can be estimated through three categories of coefficients: test-retest, alternate-form, and internal consistency coefficients. Test-retest coefficients are concerned with the agreement of measuring instruments over time. They are estimated by administering an instrument or test with the same sample at different times. Results are compared and correlated with the initial instrument to give a measure of stability. Alternate-form coefficients are associated with the use of two or more parallel forms of a test. They are computed by correlating the scores obtained on two or more parallel forms of a test when they are administered to the same sample. Internal consistency coefficients are based on the correlations between different items on the same test or the same subscale on a larger test. They measure whether the items that are intended to measure the same general construct produce similar scores.

Curriculum-based measurement. Curriculum-based measurement (CBM) was originated by Deno and Mirkin (1977) as a type of formative assessment to monitor students' academic progress. CBM was originally established for special educators as a measurement system to obtain valid and reliable repeated measurement data of the academic performance of students with disabilities and to test the effectiveness of a special education model termed *data-based program modification* (DBPM; Deno &

Mirkin, 1977). An essential difference between CBM and traditional psychoeducational measurement is that CBM provides student responses to the local classroom curriculum and these responses can enable teachers to evaluate student progress over time ((Deno & Fuchs, 1987; Deno, Fuchs, Marston, & Shin, 2001; Fuchs, Fuchs, & Courey, 2005; Fuchs, Fuchs, McMaster, & Otaiba, 2003). Over the past 30 years, researchers and teachers have extensively applied CBM in academic domains for a variety of special education purposes such as screening, identifying, and referring students at risk for academic failure.

In accord with the purposes of CBM, Deno and Fuchs (1987) conceptualized a 3 × 3 matrix, which specified three fundamental questions for developing a measurement system: "what to measure, how to measure, and how to use data" (p. 4) and three criteria to answer these questions: "technical adequacy, instructional effectiveness, and logistical feasibility" (p. 5). The selection of what to be measured is the foremost concern in developing CBM because the targeted performance should be responsive to the effectiveness of instruction and attainable for repeated measurement. Therefore, performance needs to be specific and its indicators for growth should be determined before developing CBM. The task of how to measure performance requires that measurement activities produce reliable and valid data and efficient procedures for administering curriculum-based measures.

Technical adequacy plays a fundamental role in developing the measures because CBM is used to make instructional decisions to solve problems for students with academic difficulties. The relevant issues of technical adequacy are the validity and reliability of measures. Three related types of validity are applicable to CBM: content,

construct, and criterion-related validity (Deno & Fuchs, 1987). Content validity requires CBM to reflect a specific intended domain of content when researchers develop it. Construct validity asks CBM to provide correlation between a theoretical concept and a specific measurement procedure. To have criterion validity, CBM should demonstrate the accuracy of a measure through comparing it with another valid measure. There are also four types of reliability chiefly applied to CBM: test-retest, alternate form, internal consistency, and interscorer reliability (Deno & Fuchs, 1987). To obtain the reliability, each repeated measure of CBM must be designed to assess identical concepts at an equivalent level of difficulty so that students can respond to different but equivalent measures that are drawn from the same task. Thus, students' growth in performance can be precisely determined. If a sample of students takes an instrument at different times, their scores on different administrations of the instrument should be highly correlated. In addition to technical adequacy, instructional effectiveness is another criterion used for developing CBM because CBM should reflect the effects the measures have on teacher decision-making and student progress (Deno & Fuchs, 1987). Fuchs and Fuchs' (1986) meta-analysis indicated CBM had effects on student achievement and also that different procedures in CBM produced differential student achievement. Logistical feasibility is also included as a criterion for developing CBM because the time used for administering CBM should be logically flexible.

Compared to norm-referenced achievement tests, CBM has the following characteristics: simple and efficient, sensitive, and inexpensive (Deno, 1985, 2003a). Teachers cannot administer norm-referenced tests to students in order to determine student academic progress. They usually have to wait months or even a year before they

gain knowledge of whether a student is benefiting from an academic program. In contrast, CBM probes can be given repeatedly in a short span of time and are quick to administer. For example, in order to measure a student's reading fluency, a teacher can ask the student to read a passage aloud for three minutes. The resulting information can then be graphed to demonstrate student progress. In addition, CBM probes are made up of materials taken from the local curriculum and are used to measure achievement in small environments of time (i.e., weekly or monthly), whereas norm-referenced tests are designed to measure accumulated achievement over a longer span of time (i.e., yearly). Norm-referenced tests may miss evidence of small but important improvements in a student's academic functioning. Therefore, CBM is more sensitive to students' short-term academic progress than norm-referenced tests. Last but not least, norm-referenced tests are more expensive than CBM.

Test Theories

There are primarily two test theories and their corresponding models used in measurement, classical test theory (CTT) and item response theory (IRT). CTT assumes that an observed score obtained by an individual is made up of a true score and an error score (Hambleton & Jones, 1993). A true score is an unbiased estimate of a person's latent trait, and an observed score is a test score. The discrepancy between a true score and an observed score is an error score which is called measurement error (Osterlind, 2006). The classical true score model (CTSM) within CTT is based on observed score continuum. The assumptions in the CTSM are (1) there is no correlation between true scores and error scores, (2) the expected error for a particular person is zero, and (3) error scores on parallel tests are not correlated (Hambleton & Jones, 1993). The CTSM has

several major limitations. First, standard error of measure is assumed constant for each individual in a particular population. That is, the standard error of measurement is distributed equally for all score levels. Second, CTSM estimates a person's trait level based on the summated scores so it is difficult to obtain interchangeable test forms and deal with missing data. Third, item difficulty and discrimination properties are dependent on the samples of examinees. Fourth, an observed score is computed without consideration of difficulty of items (Embretson & Hershberger, 1999).

In contrast, IRT is a theory about specifying information about a person's latent trait and properties of items (Osterlind, 2006). Within the general IRT framework, there are many models in which the relationship between item responses and latent traits can be determined, such as the Rasch model or the one-, two-, or three-parameter logistic model. The IRT models express the probability of observing each response option as a function of the respondent's latent trait level. It estimates a person's latent trait depending on responses to items and properties of items. The key assumptions in the IRT models are (1) a unidimensional trait, (2) local independence of items, and (3) relations between latent traits and observed responses have a specific form (Embretson & Reise, 2000). The line relating a trait and response is called an item characteristic curve (ICC). Furthermore, the IRT models have the following advantages. First, the IRT models calculate the standard error of measurement separately for each particular trait level and the same trait level has the same standard error of measurement. Second, the IRT models estimate a person's trait by using a maximum likelihood which is not influenced by different test forms or missing data. Third, shorter tests are more reliable than longer tests in IRT because the standard error of measurement is calculated for each trait level so that greater

reliability for a group as a whole will be achieved. Fourth, item difficulty and discrimination properties are independent of respondents (Embretson & Hershberger, 1999).

Test scores are usually employed as one of the conditions in making important decisions. For instance, a graduate school needs to decide what test score is required to admit a student into a Master's program or K-12 achievement test scores provide information about student mastery of skills so teachers can revise instruction based on these scores. Whatever decisions are to be made, the most accurate information is expected. However, students take different test forms on different test dates. If these different forms of a test are different in their statistical properties (i.e., content, reliability, and difficulty), the test scores might produce incorrect results and information so that false decisions might be made. Therefore, making decisions in different contexts requires that alternate forms of a test be equal in content and difficulty. As mentioned, raw test scores have a nonlinear relationship with students' underlying ability and no inherent meaning, so the lack of a common metric for raw scores will provide inaccurate information about students' underlying abilities if students take alternate forms of a test (Montague et al., 2010). Obtaining a common metric for raw scores is necessary.

The process of equating is applied in situations where a measure has alternate forms and scores obtained on alternate forms need to be compared with each other. The goal of equating is to obtain a common metric for raw scores and produce scores that can be used interchangeably and are comparable. Equating removes the effects of some unintended differences in the statistical properties of alternate forms on students' scores. It is necessary that an equating method be fair to students, so that the scaled scores used

for reporting students' performance have the same meaning regardless of the form administered (Dorans, Pommerich, & Holland, 2007; Kolen & Brennan, 2004). There are five requirements that have been extensively considered as important and necessary for successful testing equating (Dorans &Holland, 2000; Dorans et al., 2007; Kolen & Brennan, 2004; Lord, 1980):

a. The equal construct requirement: The two tests should measure the same constructs.
b. The equal reliability requirement: The two tests should have the same reliability.
c. The symmetry requirement: The equating function for equating the scores of Y to those of X should be the inverse of the equating function for equating the scores of X to those of Y.
d. The equity requirement: It should be a matter of indifference to an examinee to be tested by either one of two tests that have been equated.
e. The population invariance requirement: The equating functions should be invariant across subpopulations from the same population.

Since population invariance can be empirically evaluated, the present study focused on this requirement. The equating or linking functions used to convert raw scores to scaled scores should be invariant for different populations or subpopulations of individuals. Such populations or subpopulations can differ in gender, race, ethnicity, or ability. If linking functions computed on populations or subpopulations deviate from each other, alternate forms of a test will not be equal (Kolen, 2004). Nevertheless, "population invariance never really holds exactly, but it might hold approximately" (Dorans &

Holland, 2000, p. 286). When alternate forms of a test are similar in content, difficulty, and reliability, population invariance will hold approximately. Lack of population invariance can be regarded as evidence that a linking is not an equating.

The standardized root mean square difference (RMSD; Dorans & Holland, 2000), the root expected mean square difference (REMSD; Dorans & Holland, 2000), and the root expected square difference ($RESD_{j;}$ Yang, 2004) are used to investigate the population invariance of the linking functions with the whole group and ability subgroups. The two general measures RMSD and REMSD are used to evaluate population invariance by comparing linking functions obtained on subpopulations with those obtained on the whole population. The RMSD compares linking functions computed on different subpopulations with linking function computed on the whole population at a given score level and the REMSD focuses on overall differences between the whole population and subpopulation linking functions across score levels (Dorans & Holland, 2000). Yang (2004) pointed out that the subpopulation-specific $RESD_j$ measure also should be computed to supplement the overall REMSD measure of equitability because the inconsistency in some subpopulations might be ignored.

Significance of the Study

The No Child Left Behind Act (NCLB, 2001) requires that all students meet high standards in academics, including mathematics. As an important component of mathematics education, problem solving serves as a vehicle for students not only to learn and understand new mathematical concepts and skills but also to reinforce the knowledge and skills they already acquired. The National Council of Teachers of Mathematics (NCTM) posited the notion that problem solving allows students to experience the

"power and beauty of mathematics" because it is associated with various emotions during a solution process (NCTM, 1989). NCTM (2000) also advocated for developing students' conceptual understanding and problem solving rather than simply focus on computation and procedural knowledge. Therefore, developing the ability to solve math problems can help students meet the standards and significantly contribute to the outcomes of mathematics education. There is a need for measures that adequately monitor students' progress as they acquire the ability to solve math problems.

Stecker, Lembke, and Foegen (2008) addressed the importance of progress monitoring for states, schools, teachers and students. Teachers can oversee student performance in a particular area through progress monitoring and revise their instructions in time, based on an individual student's need. Schools can examine the efficacy of instructional programs and make instructional decisions for students who need more instruction than their peers by monitoring student progress over time. Therefore, states can evaluate the effectiveness of their educational systems by using progress-monitoring data (Stecker et al., 2008). As one of the formative assessment tools, CBM can effectively and efficiently monitor student progress across domains that have been validated by researchers (e.g. Deno, 1985; Deno & Fuchs, 1987; Deno & Mirkin, 1977; Fuchs & Deno, 1992; Marston, 1989).

With the reauthorization of the Individuals with Disabilities Education Act (IDEA) in 2004, *Response to Intervention* (RtI) was included as an alternative method to identify students with LD. Within a three-tier RtI model, students who are at risk for LD are given more than one research-validated interventions and their academic progress is frequently monitored. If they fail to respond to intervention, they may be identified

eventually as having LD. That is to say, frequent progress monitoring is necessary to index student performance over time. CBM uses brief, frequent, targeted formative assessments to monitor a student's response to intervention. By administering these formative

assessments, a student's progress is monitored closely and modifications are made to the intervention. Therefore, CBM can be widely used within RtI models (Busch & Reschly, 2007; Fuchs & Fuchs, 2006; Speece, Case, & Molloy, 2003; Wallace, Espin, McMaster, Deno, & Foegen, 2007).

To adequately measure progress over time, it is necessary that alternate forms of a test be equivalent in difficulty level and also that a statistical relationship between raw scores be established when student progress is monitored over time (Dorans, et al., 2007). To reiterate, the purpose of this study was to investigate whether the Rasch mean/mean linking functions computed on the whole group and subgroups are invariant. To accomplish this, population invariance of the linking function for each subgroup (i.e., AA students, LA students and students with LD) were examined at each score level and across score levels.

Chapter Two

Literature Review

Two decades of research have proven the technical adequacy of CBM in providing reliable and valid indicators of student performance and progress (Marston, 1989). CBM has had an influence on measuring student progress in all academic domains (Stecker, Fuchs, & Fuchs, 2005). The majority of CBM studies have been conducted at the elementary school level, and reading has received more attention than mathematics, written expression, and spelling. Researchers primarily have attended to the technical adequacy, instructional effectiveness, and logistical feasibility of these measures. More recently, CBM has been considered as a mechanism for monitoring students' progress in general education classrooms and identifying and classifying students with learning difficulties within RtI models. This chapter will review specific research applications of CBM across academic domains with a particular focus on mathematics and research studies on population invariance.

CBM Research in Reading

The majority of research on the development of CBM for assessing reading performance has concentrated on oral reading fluency (ORF), which has been determined to be a valid and reliable way of measuring students' general reading ability and comprehension (Carver, 1992; Deno, Mirkin, & Chiang, 1982; Fuchs, Fuchs, & Maxwell, 1988; Madelaince & Wheldall, 1999). The research suggests that there are three main measures used to examine students' oral reading: read aloud, maze selection, and word identification (Wayman, Wallace, Wiley, Ticha, & Espin, 2007). Deno et al. (1982) conducted a series of studies with elementary school students to identify possible

measures of their reading progress and investigated the correlations between these measures and different standardized reading tests. The results revealed that read aloud measures had high correlations with standardized reading tests with a range from 0.73 to 0.91. Other researchers came to a similar conclusion that read-aloud measures could be a valid and reliable indicator of students' reading proficiency (e.g., Fuchs et al., 1988; Jenkins & Jewell, 1993; Marston, 1989).

In the Deno et al. study (1982), students were asked to read aloud a passage sampled from curricular materials for one minute and the number of words that were correctly read by students was counted. Hesitation of more than three seconds, substitution, mispronunciation, and omission were counted as incorrect, but self-correction within three seconds was marked as correct. They found that read-aloud measures had stronger correlations with standardized reading achievement tests than the correlations between cloze comprehension measures and standardized comprehension tests. Fuchs et al. (1988) also examined the relationship between the performance of students with mild disabilities from grades four to eight on read-aloud measures and performance on standardized reading comprehension tests. The results showed that read aloud scores strongly correlated with standardized reading comprehension scores. They also compared read-aloud measures to other measures for assessing students' reading comprehension, such as written cloze, story retellings, and question answering. Among these measures, read-aloud measures were found to have a stronger correlation with standardized reading comprehension tests than the other three measures. Jenkins and Jewell (1993) explored the validity of two curriculum-based measures, a read-aloud and a maze task. They examined the relationship between students' performance on these two

measures and their performance on two standardized reading tests, the Gates-MacGinitie Reading Test (MacGinitie, Kamons, Kowaski, MacGinitie, & Mckay, 1978) and the Metropolitan Achievement Test (MAT; Prescott, Balow, Hogan, & Farr, 1984). The results indicated that the two curriculum-based measures were significantly correlated with the two standardized achievement tests. However, the correlation coefficient between ORF and two achievement tests decreased when the grade level increased. Confirmatory factor analysis (CFA) was used by Shinn, Good, Knuston, Tilly, and Collins (1992) to investigate third and fifth grade students' reading decoding, fluency, and comprehension skills through read aloud measures. A one-factor model, emerging with reading competence for the third graders, indicated that all the reading skills significantly contributed to students' reading competence.

Read-aloud measures have good technical adequacy, but they need to be administered individually and whether they are appropriate for older students needs to be clarified (Wayman et al., 2007). Due to the limitations of read-aloud measures, the maze task is considered by researchers as another valid and reliable measure to monitor students' reading progress (e.g., Fuchs & Fuchs, 1992; Fuchs, Fuchs, Hamlett, & Ferguson, 1992; Jenkins & Jewell, 1993). Fuchs and Fuchs (1992) established the rules for creating maze passages in which typically every seventh word was eliminated and replaced by a three-word multiple choice with only one correct answer. The criterion validity of four alternative measures: question answering, story recall, cloze, and maze tasks, were compared in their study. Levels of teacher satisfaction were also collated for these four measures. For the maze task, special education teachers administered the measure to the upper elementary students with mild to moderate disabilities twice weekly

over eighteen weeks via computer. The results suggested that the maze task could be a promising monitoring measure because there was a small ratio between slope and standard error of estimate, which indicated that the maze task could easily detect students' progress in performance. Jenkins and Jewell (1993) also found there was a significant correlation between the maze tasks and standardized reading achievement tests.

Daly, Wright, Kelly, and Martens (1997) investigated the technical adequacy of eleven curriculum-based measures of early academic competencies including reading, math, color, and shape with first-grade students. Reading measures involved letter reading, word list reading, letter copying, letter-sound production, and letter-sound selection. The results suggested that word list reading and letter reading measures were the best technically among these eleven measures. The validity of word identification fluency and nonsense word fluency measures were also compared in Fuchs, Fuchs, and Compton's study (2004). Two curriculum-based measures were administered to the first grade at-risk students once a week for seven weeks and twice a week for another thirteen weeks. The results demonstrated that the word identification measure had greater validity than the nonsense word fluency measure. Compton, Fuchs, Fuchs, and Bryant (2006) administered two assessments, a prediction battery and outcome assessments, to first-grade students to assess their word identification fluency (WIF). The prediction battery included measures of phonemic awareness, rapid naming, oral vocabulary, and WIF. It was administered in the fall term of the first grade. The outcome assessments contained standardized reading tests and were administered at the end of the second grade. Moreover, the researchers monitored the students' progress for five continuous weeks by

administering alternate forms of WIF once a week. They found that the WIF measures produced more classification accuracy for identifying at-risk students than the other three measures.

Thus, numerous studies have indicated the technical adequacy of curriculum-based measures for monitoring students' progress in reading. However, the adequacy of the alternate forms of the measures with respect to difficulty level was rarely established. Two recent studies have addressed this issue (Betts, Pickat, & Heistad, 2009; Christ & Ardoin, 2009). In the first, Christ and Ardoin (2009) looked at four procedures for developing oral reading passage sets from 50 third grade CBM reading passages with second and third grade students. The four procedures included random selection, Spache readability statistics, mean level of performance evaluation, and Euclidean Distance (ED). The results indicated that the passage selection procedures had an influence on the measurement outcomes; however, none of the procedures could generate optimal passage sets. In the second, Betts et al. (2009) determined the degree to which readability statistics could provide an appropriate measure of difficulty levels of curriculum-based reading passages with students in first grade through eighth grade. The results suggested that readability statistics could distinguish passage difficulties between grades but not within grades. They also highlighted the importance of obtaining equivalent forms of curriculum-based reading passages by using equating or linking methodology and suggested that IRT could be used to evaluate word-by-word responses and place reading passages on the same scale so that the scores would be comparable.

CBM Research in Written Expression

Like the research in CBM of reading, research on CBM in written expression has examined the validity and reliability of measures (e.g., Deno, Mirkin, & Marston, 1980; Espin, De La Paz, Scierka, Roelofs, 2005; Espin et al., 2000;Videen, Deno, & Marston, 1982). Also, like the reading research, no study addressed equivalency of alternate forms of CBM of written expression. Deno et al. (1980) conducted three concurrent validity studies on the relationship between performance on behavioral measures of written expression and performance on standardized written expression achievement tests with elementary school students. The behavioral measures of written expression involved T-units length, mature words, total words written (WW), large words, and words spelled correctly (WSC). The standardized written expression achievement tests included the Test of Written Language (TOWL; Hammill and Larson, 1978), Stanford Achievement Test, Intermediate I, Word Usage Subtest (SAT; Madden, Gardner, Rudman, Karlsen, & Merwin, 1978), Developmental Sentence Scoring System (DSSS, Lee & Canter, 1971), and Program Placement. The researchers also provided three prompts including picture stimulus, story prompt, and topic sentences with the students. Across these studies, all behavioral measures except T-unit had high correlation coefficients ranging from 0.69 to 0.88. Based on Deno et al. studies, Videen et al. (1982) included another written expression measure in their study which is correct word sequences (CWS). CWS referred to "two adjacent, correctly spelled words that are acceptable within the context of the phrase to a native speaker of the English language" (p. 7). The results revealed that there was a strong relationship between the CWS measure and standardized achievement tests,

and the CWS measure could be considered as a valid and reliable measure of written expression.

In addition to the research on elementary school students, a number of studies for secondary school students also focused on the validity of CBM of written expression (e.g., Tindal & Parker, 1989a; Tindal & Parker, 1989b). Tindal and Parker (1989a) analyzed six-minute writing samples with middle school students who were either in special education or remedial programs by utilizing eight indices. Factor analysis indicated that the eight indices of written expression were intercorrelated, and a two-factor model emerged with a production factor and production-free factor. The two-factor model accounted for 82.8% of variance in students' writing performance. Along with validity, four types of reliability of CBM of written expression are commonly examined: test-retest, alternate-form, internal consistency, and interscorer reliability (e.g., Deno, Mirkin, & Marston, 1980; Espin et al., 2000; Fuchs, Deno, & Marston, 1982; Marston & Deno, 1981; Weissenburger & Espin, 2005). Deno et al. (1982) investigated the reliability of written expression measures developed by Deno et al. (1980) on elementary school students in grade one to grade six across three states. The students took the written expression measures twice during the fall and spring semesters. Each time they were provided with two story starters and asked to finish each one in three minutes. The results revealed that the WW, WSC, and Correct Letter Sequence (CLS) measures had moderate to strong reliability across the grades (*r*s = .70 - .86). Marston and Deno (1981) conducted a series of studies to look at the reliability of Deno et al.'s (1980) curriculum-based measures of written expression. In the first study, they evaluated test-retest reliability of these measures. The WW, WSC, and CLS measures were found to have

strong correlations over one-day interval because coefficients were .91, .81, and .92, respectively and moderate coefficients ranging from .62 to .70 over a three-week interval. In the second study, they looked at alternate-form reliability of these measures. The results demonstrated that the WW (r = .95), WSC (r = .95), and CLS (r = .96) measures had relatively strong correlations with the two five-minute story tasks. Moreover, they investigated the internal consistency of these measures in their third study. Likewise, the reliability coefficients were high for the WW, WSC, and CLS measures.

Additionally, there are some studies in CBM of written expression investigating the reliability of written expression measures at the middle school level (e.g., Espin et al., 2005; Espin et al., 2000; Parker, Tindal, & Hasbrouck, 1991a, 1991b; Weissenburger & Espin, 2005). Parker et al. (1991 a) explored the technical adequacy of writing samples with the middle school students with mild learning disabilities for six months. Four sets of the students' writing samples were obtained over a school year and seven indexes were applied to these samples. All scoring procedures had strong interscorer reliabilities (rs = .83-.98). The alternate-form reliability was investigated in both Espin et al.'s studies (2000, 2005). Espin et al. (2000) evaluated the alternate-form reliability of various indicators of writing performance and the effects of writing type and sample duration on these indicators. They asked 7th and 8th grade students to produce two narrative and two expository samples in five minutes each. They also located the students' writing positions at three minutes. The results indicated that the writing types and duration of samples had no significant effects on the indicators. Moreover, the WW, WSC, CWS, and CIWS measures had moderate to strong alternate-form reliability. CBM research in mathematics mirrors the research in reading and written expression.

CBM Research in Mathematics

The majority of studies in CBM of mathematics also investigated the validity and reliability of measures. The research on CBM of elementary school mathematics has received more attention than CBM of early mathematics or secondary school mathematics. Research on CBM in early mathematics highlights number identification, number naming, visual discrimination, objects counting, and missing number (Chard et al., 2005; Clarke & Shinn, 2004; VanDerHeyden et al., 2004; VanDerHeyden, Witt, Naquin, & Noell, 2001). VanDerHeyden et al. (2001) developed a series of curriculum-based measures on reading, mathematics, and writing and investigated the technical adequacy of these measures through two phases of the study. With respect to the measures of math, they included circling number (circle one particular number out of four numbers), writing number (count the number of objects), and drawing circles (draw the specified number of circles). Two types of reliability of these measures were assessed, alternate-form and interscorer reliability, with a sample of kindergarten children. The results demonstrated that the three measures had moderate to strong alternate-form reliability with the coefficients ranging from .70 to .84. The interscorer reliability coefficients across the three measures all were greater than .95. To estimate the concurrent, predictive, and social validity of the measures, the researchers randomly selected 40 students from the participants and administered nine subtests of the Comprehensive Inventory of Basic Skills, Revised (CIBS-R; Brigance, 1999) and Dynamic Indicators of Basic Early Literacy Skills (DIBELS; Good & Kaminski, 1996; Kaminski & Good, 1996). The findings suggested that all three measures had low to moderate correlations with the subtests of CIBS-R. Similar results were found in

VanDerHeyden et al.'s study (2004). They extended that study and specifically examined the validity and reliability of six mathematics probes with preschool children. The six probes included number identification, number naming, objects count, free count (the height of counting), visual discrimination, and shape identification. The participants finished each probe in three minutes. In the first phase of the study, the researchers randomly selected three out of six probes and their alternate forms and administered them to half of the participating children. The remaining three probes and their alternate forms were administered to another half of the children. They noticed that the probes except shape identification and free count had strong alternate-form reliability with the coefficients ranging from .83 to .88. The interscorer reliability was also reported as strong. In the second phase of the study, they evaluated the criterion validity of these measures. The criterion variables involved Brigance Screens (Brigance, 1985) and Test of Early Mathematics Ability (TEMA-2; Ginsburg & Baroody, 1990). The results demonstrated that both number identification and visual discrimination measures had moderate correlations with both criterion measures in a range from .50 to .57. The free count measure had moderate correlation with Brigance Screens (r = .56), but weak correlation with TEMA-2 (r = .19). There was no correlation between the shape identification measure and Brigance Screens.

In a similar study, Clarke and Shinn (2004) gauged the psychometric properties of four curriculum-based measures on mathematics: oral counting, number identification, quantity discrimination, and missing number on the first grade students. Another three measures, Math CBM Grade One Computation Probes (M-CBM), WJ-R Applied Problems subtest (Woodcock & Johnson, 1989), and Number Knowledge Test (Okamoto

& Case, 1996) were administered to determine the concurrent and predictive validity of these four measures. The results showed that the reliability coefficients across the four measures were all close to or greater than .80. Most validity coefficients were moderate to strong. Similar results were also found in Chard et al.'s study (2005).

The most commonly used CBM in elementary school mathematics is the Monitoring Basic Skills Progress measures (MBSP; Fuchs, Fuchs, Hamlett, &Allinder, 1989; Fuchs, Hamlett, & Fuchs, 1998; 1999). They include two types of measures, computation (CBM-COMP) and concepts/applications (CBM-APP). There are 30 parallel forms per grade for grades one through six in computation and 30 parallel forms per grade for grades two through six in concepts/applications. The MBSP measures have been widely utilized and validated through a series of studies for over 20 years (e.g., Fuchs, Fuchs, Hamlett, Stecker, 1990; Fuchs, Fuchs, Hamlett, 1989; Fuchs et al., 1989; Fuchs, Fuchs, Hamlett, Stecker, 1991; Fuchs et al., 1994; Fuchs, Fuchs, Hamlett, Walz, & Germann, 1993; Shapiro, Edwards, & Zigmond, 2005). The research, however, has again explored only the technical adequacy of these measures.

Fuchs et al. (1994) adopted the CBM-COMP (Fuchs et al., 1989) and developed a set of CBM in mathematics concepts and application (CBM-APP) for second grade through fourth grade students. The students took one of the measures once a week over 20 weeks. They also graphed students' performance over time. The Comprehensive Test of Basic Skills (CTBS; Macmillan/McGraw-Hill, 1989), fourth edition, which consisted of three math subtests, Mathematics Computation, Mathematics Concepts and Applications, and Mathematics Total Battery, were included as criterion measures. The

internal consistency coefficients were .98 for second graders, .94 for third graders, and .97 for sixth graders. The correlation between the CBM-APP and the CTBS ranged from .74 to .81 across three grades. The students in grade two had higher slopes in the COM-APP (.40) than those in the COM-COMP (.25). However, the fourth grade students had similar slopes for both measures (.69 and .70). The slopes for the CBM-COMP and CBM-APP at grade two were both lower than those at grade four. Similar results were found in other studies. For example, Shapiro et al. (2005) used the CBM-COMP and CBM-APP to assess student progress in math. The slopes for students' performance on both measures were calculated and graphed with the students in grade one through grade six. The results revealed that mean of change across participants in both the CBM-COMP and CBM-APP were the same (.38) and the students showed improvement of 13.7 digits over a period of nine months. Fuchs et al. (1993) reported that CBM math slopes were in a range of .20 (second grade) to .77 (fourth grade) during the first year of their study and .28 to .74 during the second year.

There are only two studies that investigated CBM in math word problems (Jitendra, Sczesniak, & Deatline-Buchman, 2005; Leh, Jiterdra, Caskie, & Griffin, 2007). Jitendra et al. (2005) assessed reliability and validity of eight forms of curriculum-based measures in math word problem solving fluency (WPS-Fluency) with third-grade students with and without LD. The students took the measure once every two weeks over sixteen weeks. They also took the MBSP once a week to examine its predictive power. Criterion measures included Math Problem Solving and Procedures subtests of the Stanford Achievement Test-9 (Standford 9; Harcourt Brace & Comparny, 1996) and Mathematics Computation and Concepts and Applications subtests of the TerraNova

(CTB/McGraw-Hill, 2001). Cronbach's alpha coefficients demonstrated that each probe had moderate reliability in a range of .60 to .75. These coefficients increased to the range of .76 to .83 when odd and even probes were aggregated. The WPS-Fluency measure was identified to have stronger correlations with Stanford 9 Problem Solving test (r = .71 and .54) than MBSP (r = .49 and .50). In contrast, the WPS-Fluency measure (spring and winter) had lower correlations (r = .58 and .38) with Stanford 9 Procedures test than the MBSP (spring and winter) (r = .64 and .66). Additionally, the WPS-Fluency measure had stronger predictive validity (r = .77) than the MBSP (r = .43). Leh et al. (2007) replicated Jitendra et al.'s study (2005) with third grade students who were categorized as low- (LO) and averaging-performing (AV) students and then evaluated the degree to which the WPS measure was sensitive to students' growth across time. The results indicated that LO and AV students had the same rate of change on both the WPS and MBSP measures and WPS was again considered as a valid and reliable measure to monitor students' progress in math word problem solving. The slope for the MBSP measure was similar to that reported by Fuchs et al. (1993).

A few studies looked at CBM of secondary math (e.g., Foegen, 2000; Foegen & Deno, 2001; Helwig, Anderson, & Tindal, 2002; Helwig & Tindal, 2002). Foegen and Deno (2001) probed the technical adequacy of four researcher-developed curriculum-based measures on math with the students in grade six through eight. These measures incorporated Basic math Operations Task (BMOT), Basic Estimation Task (BET), and Modified Estimation Tasks (METs) with form A and B. The criterion measures consisted of California Achievement Test (CAT), students' math and overall GPA, and the teacher's rating with respect to students' competence in mathematics. Their findings

demonstrated that the BET and BMOT measures had stronger internal consistency coefficients (.93 and .92, respectively) than the METs measure (.77-.81). All the measures had strong test-retest reliability with the coefficients ranging from .80 to .88. The BMOT and METs measures also had moderate to strong alternate-form reliability in a range of .79 to .86. The criterion coefficients suggested that the BMOT measure had a strong relationship with students' overall GPA and the mathematics computation subtest of CAT. The MET-B measure was found to have a closer relationship with the reading subtest of CAT than the other measures. Additionally, regression analyses demonstrated that the BMOT measure was identified as the strongest predictor of the computation subtest of CAT. Likewise, Foegen (2000) explored the validity and reliability of two general outcome measures, the *fact probe* and *estimation probe* with a sample of sixth-grade students. The sensitivity of the measures to changes in student performance was attended to at the same time. There were 80 single-digit facts in the one-minute fact probe and each of the four operations contained 20 facts. The three-minute estimation probe included 20 computation problems and 20 word/story problems. The criterion measures included the Iowa Tests of Basic Skills (Hoover, Hieronymus, Fribie, & Dunbar, 1993), students' math grade, and students' semester GPA. The math teacher also rated and ranked the overall math capability for all participants. Both general outcome measures were administered to the students once a week for consecutive ten weeks. The results indicated that the estimate probe had moderate to strong reliability in a range of .75 to .87 and weak to moderate correlations with the criterion measures. The moderate mean effect sizes for the fact probe and estimation probe were reported as .70 and .51, respectively. An ordinary least-squares regression was run to produce slope values between scores and

calendar days for each student. The results suggested that students' scores on both measures were significantly higher at the end than those at the beginning. The mean slope for the fact probe (.55) was twice the size of the mean slope for the estimation probe (.25).

Two studies examined middle school students' math conceptual understanding and application rather than computation skills (Helwig, Anderson, & Tindal, 2002; Helwig & Tindal, 2002). Helwig and Tindal (2002) developed a fifteen-item general outcome measure of mathematics and administered four alternate forms of this measure to the eighth grade students across an academic year. The measure contained one computation problem, nine real-life word problems, and five mathematics conceptual understanding problems. The results showed that the four curriculum-based measures had high reliability coefficients in a range of .81 to .86. Correlations between each CBM and the statewide achievement test also ranged from .81 to .88. There was a significant increase among each form except between form two and three which demonstrated a decrease of .07. Similar results were found in Helwig et al.'s study (2002).

Again, like the research in reading and written expression, these studies of CBM of mathematics focused solely on technical adequacy of measures. The measures may be technically adequate, but still may not be a good indicator of progress over time because the alternate forms are not equivalent in level of difficulty. There is only one study that addressed equivalent forms of curriculum-based measures in math problem solving (Montague et al., 2009). This study used IRT to equate alternate forms. The researchers developed seven equivalent forms of curriculum-based measures to monitor progress in math problem solving by applying the dichotomous Rasch model (Rash, 1980). The item

difficulty parameters were placed on a common metric and the obtained scores were then comparable and interchangeable. These estimates of student ability are more accurate than raw scores (i.e., number of correct responses). The equating methodology was found to have a significant influence on the longitudinal assessment of student progress. That is, the raw score analysis did not capture the considerable growth of student underlying math ability because raw scores were on a different metric and did not have a nonlinear relationship with the underlying ability metric. Therefore, this study underscored the need to equate scores of alternate forms used in CBM.

Population Invariance

A number of research studies have examined population invariance by using the data from standardized testing programs like the College-Level Entrance Placement (CLEP), Law School Admission Test (LSAT), Scholastic Aptitude Test (SAT), etc. No study has investigated population invariance of linking functions for non-standardized tests, such as CBM.

Yi, Harris, and Gao (2008) examined population invariance of the IRT true- and observed-scores equating as well as equipercentile equating methods for a science achievement test with subgroups that differed in ability. Four forms of the science achievement test (e.g., Form A, B, C, and X) were spiraled and administered to randomly equivalent groups of examinees on one test date. They divided the participating examinees into two subgroups (low-ability and high-ability groups) under the following three situations: examinees' self-reported GPA, the average of four test scores of the science achievement test, or whether examinees had taken a physics course. The

examinees who had the sum of self-reported GPA for four science courses more than 8.0 were in the high-ability group; otherwise, they were in the low-ability group. The examinees who had taken a physics course were assigned to the high-ability group and the students who had not would be assigned to the low-ability group. If the examinees within one test center had a lower average test center composite score than the mean of the average test center composite score, they were assigned to the low-ability group; otherwise, they were assigned to the high-ability group. They found evidence of population invariance of the IRT true-score equating functions for subgroups defined by whether examinees had taken a physics course on Form B not Form C. There was population dependence of the IRT true-score equating functions on Form C across the subgroups defined by examinees' self-reported GPA. The results indicated that the linking functions were more population sensitive if subpopulations' abilities related to performance on the test.

In addition, subgroups can be defined by gender, ethnicity, or race. Von Davier and Wilson (2008) examined population sensitivity of the IRT equating functions for gender subgroups on the Advanced Placement Program Calculus AB exam which includes both multiple-choice and free-response sections. The IRT equating functions involved the IRT true-score equating, chained equipercentile equating, and Tucker linear equating. They applied an internal anchor test data collection design to equate the multiple-choice test and the test including both multiple choice and free-response items. They found that the IRT true-score linking differences between each subgroup and the whole group were population invariant for the multiple-choice test only and the test including both multiple choice and free-response items. The chained equipercentile

linking functions were also found to be population insensitive for the multiple-choice test to an acceptable degree. However, the Tucker linear linking functions exhibited population dependence for the multiple-choice test.

Yang and Gao (2008) also investigated population invariance of the IRT-based equating for gender subpopulations for the 16 testlet-based forms of the CLEP College Algebra exam. They found that the linking functions were invariant over gender subpopulations across these 16 forms. Likewise, Liu and Holland (2008) extended the application of population invariance of linking functions to subpopulations defined by gender, race/ethnicity, geographic region, whether examinees applied to law schools, and their law school admission status by using LSAT data from a single administration. They explored population invariance for highly equitable tests, less equitable tests, and inequitable tests and found that "results from equating parallel measures of equal reliability showed very little evidence of population dependence of equating function" (p. 41). Similar results were found in Doran's study (2004). He examined population invariance across different forms of the Advanced Placement (AP) Calculus AB exam which were linked under the non-equivalent common item design. The results indicated that the linking function was not sensitive to gender subpopulations.

Thus, once again, the purpose of this study was to investigate population invariance of the Rasch mean/mean linking functions computed on the whole group and on ability subgroups (i.e., AA students, LA students, and students with LD), using the data from five alternate forms of a math problem solving measure. The RMSD, $RESD_j$, and REMSD measures with proportional weights and equal weights were used. The population invariance of the linking functions for the whole group and ability subgroups

was examined at each score level and across score levels. The present study addressed the following research questions:

1. To what extent is there unconditional invariance between the linking function computed on the whole group and the linking function computed on the AA subgroup?
2. To what extent is there unconditional invariance between the linking function computed on the whole group and the linking function computed on the LA/LD subgroup?
3. To what extent is there conditional invariance in the linking function computed on the whole population at any particular score level?
4. To what extent is there unconditional invariance in the linking function computed on the whole population across all score levels?

Chapter Three

Method

Participants

Sample One. The participants in this study were eighth-grade students in the Miami-Dade County Public Schools (M-DCPS) in 2008-2009. M-DCPS is the fourth largest school district in the nation serving approximately 380,000 students. The ethnic distribution is as follows: 9 % white, 26 % African-American, 62 % Hispanic, 1% Asian, and 2 % of other minorities. The M-DCPS district has 57 middle schools. Forty schools representing low, moderate, and high performing schools based on Florida Comprehensive Assessment Test (FCAT) performance level were recruited from the M-DCPS. These schools were participating in a federally funded intervention study which is investigating the efficacy of *Solve It!* (Montague, 2003). *Solve It!* is an intervention designed to teach students with math difficulties how to understand, analyze, solve, and evaluate mathematical problems by developing the processes and strategies that effective problem solvers use. One general education math teacher from each school was nominated by a school administrator to participate in the study. The study involved 779 eighth-grade students attending inclusion, general math, and pre-algebra classes in middle schools. English language learners enrolled as English for Speakers of Other Languages (ESOL) in level one, two, or three were excluded. Students with LD were district-identified (see note for LD district eligibility criteria) and, like students who were categorized as LA, had FCAT math levels of 1 or 2 out of a possible 5. Students categorized as AA had FCAT math levels of 3 or 4.

Sample Two. The participants in this study were seventh-grade students in the M-DCPS in 2009-2010. The criteria for screening schools and students were same as those for the sample one. The study involved 1,082 seventh-grade students attending inclusion, general math, and pre-algebra classes in middle schools. Demographic data for participating students are presented in Table 1.

Table 1 *Student Demographic Data*

	Sample One Eighth Grade (n=779)	Sample Two Seventh Grade (n =1082)
Variable	n (%)	n (%)
Ability Level		
AA	358(46)	274(25)
LA	343(44)	718(67)
LD	78(10)	90(8)
Gender		
Males	359(46)	462(43)
Females	420(54)	620(57)
Ethnicity		
White	55(7)	56(5)
Hispanic	497(64)	688(64)
Black	213(27)	323(30)
Other	14(2)	15(1)
Free/Reduced Lunch		
Yes	563(72)	869(80)
No	216(28)	213(20)

Measure

Seven curriculum-based measures of math problem solving were developed for the study using 30 textbook-type math word problems drawn from the *Solve It!* manual (Montague, 2003). The representation of the item composition across all seven forms was described in detail by Montague et al. (2009). Each measure consisted of 10 problems (i.e., 2 one-step problems, 6 two-step problems, and 2 three-step problems) like the following: A store sells shirts for $13.50 each. On Saturday, it sold 93 shirts. This was 26 more than it had sold on Friday. How much did the store charge for all the shirts sold on both days? In order to solve these problems, students need knowledge of whole numbers and decimals and the four basic operations. They do not need to know specific formulas or have unique math knowledge to solve these problems.

Procedure

Solve It!, the intervention, includes a detailed instructional guide, scripted lessons, class charts, student cue cards, and multiple practice problems. All intervention materials including class sets of practice problems were provided for the entire school year. The intervention implementation occurred over three days in the middle of October. Then teachers provided weekly problem-solving practice sessions for students. Students in the comparison group received typical classroom instructions.

Curriculum-based measures were administered to the entire class seven times for the intervention group classes (i.e., prior to the intervention and then monthly for the remainder of the school year) and three times for the comparison group classes (i.e., prior to the intervention and then the second and seventh administrations). Only five of the

measures were used in the present study because the sample one students did not take CBM 6 and CBM 7 is identical to CBM 4.

Design

The common items across the five forms asked for an IRT common-item nonequivalent groups (NEAT) design to equate the test scores (Kolen & Brennan, 1995). In the NEAT design, examinees are administered different test forms that have a subset of items in common. The mean/mean equating was used to link five alternate forms of the curriculum-based measure under the dichotomous Rasch model (Rasch, 1980). The Rash mean/mean equating refers to an equating relationship established between latent traits on the alternate forms. The Rasch model is a probabilistic measurement model that specifies the probability of correct responses as a function of item difficulty parameters and person abilities. It provides a model of expected responses in which both the performance of the participants and the difficulty of the items are compared in terms of fitting a unidimensional continuum model. In the Rasch model, number-correct scores are considered as sufficient for estimating examinees' underlying abilities and examinees who have the same number-correct scores will have the same estimates of ability (Bond & Fox, 2007).

Data Analysis

The data were analyzed with the Winsteps software (Linacre, 2005) to calibrate all items in the five forms of CBM with the Rasch model. Rasch fit statistics provided useful data to examine the quality of the test items (Bond & Fox, 2007).The infit mean square statistic (infit MNSQ) and the outfit mean square statistic (outfit MNSQ) were evaluated. The outfit MNSQ is not weighted and is sensitive to anomalous outliers for

either person or item parameters, whereas the infit MNSQ is sensitive to residuals close to the estimated person abilities (Bond & Fox, 2007). Fit statistics for items have an expected value of 1.0, and can range from 0 to infinity. The range of 0.5 to 1.5 can be considered as a good fit.

In the following, CBM 1 was chosen to function as the anchor form and the other four forms (CBM 2, 3, 4, and 5) were equated to it using the mean/mean method (Loyd & Hoover, 1980; Kolen & Brennan, 2004) in order to adjust inevitable differences in difficulties, so the metric of the initial item parameter estimates for CBM 1 was used as the metric to which the other four forms were equated. The equating analysis then placed the difficulty parameters of CBM 2 to CBM 5 on the same metric as CBM 1. An initial Rasch calibration was separately conducted for all the five forms. The difficulty parameter estimates of CBM 1 served as the metric and were not transformed. The mean/mean method for the whole group was conducted in the following way. The initial item difficulty parameter estimates of CBM 4 were transformed using the common items (items 1-5) so that they were on the same metric as CBM 1 as following: finding the average difficulty parameter for the five common items in CBM 1 and in CBM 4, taking the difference of the average for CBM 1 and for CBM 4, adjusting the difficulties of the five unique items in CBM 4 by adding the difference of the CBM 1 and CBM 4 common item average difficulties to each of the unique items in CBM 4, and fixing the common items in CBM 4 back to the same values obtained for CBM 1. Therefore, CBM 4 had five items with difficulties identical to those of CBM 1 and five unique items with difficulties that had been adjusted to be on the metric of CBM 1. The initial item difficulty parameter estimates of CBM 2 were transformed using the common items

(items 16-20) in the same way so that they were on the same metric as CBM 4. The initial item difficulty parameter estimates of CBM 5 were also transformed using the common items (items 6-10) in the same way so that they were on the same metric as CBM 1. Likewise, the initial item difficulty parameter estimates of CBM 3 were transformed using the common items (items 21-25) so that they were on the same metric as CBM 5. After the item parameters were on the same scale, the ability estimates on these five forms were considered to be equivalent.

For the present study, the whole group consisted of three subgroups: AA students, LA students, and students with LD. The students with LD and LA students were collapsed in order to increase the sample size. The calibration and mean/mean method were carried out on the AA students and the LA/LD students, respectively.

In order to investigate population invariance of the Rasch mean/mean linking functions computed for the whole group and the ability subgroups, three indices were employed, RMSD, REMSD, and $RESD_j$. The RMSD (Dorans & Holland, 2000) was used to compare linking functions computed on ability subgroups with the functions computed on the whole group at a given score level. The formula for the RMSD is shown in

equation (1)

$$RMSD(x) = \frac{\sqrt{\sum_j w_j \left[e_{P_j}(x) - e_P(x)\right]^2}}{\sigma_{YP}} \qquad (1)$$

where x donates a score level of the test X. P is the whole group of participating students; P_j is a subgroup of the whole group P. $e_P(x)$ represents the overall linking function that

equates X to Y on the whole group P. $e_{P_j}(x)$ donates the linking function that equates X to Y on the subgroup P_j of P. The weight w_j is the relative proportion of P_j in P. The denominator σ_{Y_P} is the standard deviation of Y in P. $e_{P_j}(x) - e_P(x)$ represents the difference between the equated score within the subgroup linking function and the equated score within the whole group linking function at a score level x.

Dorans and Holland (2000) also introduced the REMSD in order to investigate overall differences between the whole group and subgroups linking functions across score levels. The formula for the REMSD is shown in equation (2)

$$REMSD = \frac{\sqrt{\sum_j w_j E_P\left\{\left[e_{P_j}(X) - e_P(X)\right]^2\right\}}}{\sigma_{YP}} \tag{2}$$

where $E_P\{\cdot\}$ describes averaging over the distribution of X in the whole group P and $e_{P_j}(X) - e_P(X)$ tells the difference between equated scores within the subgroup linking function and equated scores within the whole group linking function across all score levels on the test X.

Furthermore, Yang (2004) proposed the $RESD_j$ to compare the linking functions between each subgroup and the whole group across all score levels. The formula for the $RESD_j$ (Yang & Gao, 2008) is shown in equation (3)

$$RESD_j = \frac{\sqrt{E_P\left[e_{P_j}(x) - e_P(x)\right]^2}}{\sigma_{YP}} \tag{3}$$

where E_P denotes averaging over raw score levels weighted by the proportion of examinees at each score level in the whole group P.

The RMSD, REMSD, and $RESD_j$ statistics were computed using the formula given in the (1), (2), and (3). Both the RMSD and REMSD statistics involve w_j weights for each of the subgroups. The REMSD takes the expectation over *P*, so that the scores on *X* are weighted. Two choices for the w_j were used, proportional weights (weights proportional to sample sizes) and equal weights (0.5).

To evaluate magnitudes of the proportionally weighted RMSD, proportionally weighted REMSD, equally weighted RMSD (ewRMSD), equally weighted REMSD (ewREMSD), and $RESD_j$, *difference that matters* (DTM; Dorans & Feigenbaum, 1994) was employed as a benchmark to assess whether differences in linking functions have practical significance (Holland & Dorans, 2006). Computationally, the DTM is defined as half of a reported score unit (Dorans, Holland, Thayer, & Tateneni, 2003). It can also be standardized by dividing by $\sigma_{Y_P}$and denoted as SDTM (von Davier & Wilson, 2008). Nevertheless, the intended use of test scores and the scale units of the test have an impact on its benchmark, so half of a reported score unit is not appropriate and meaningful for some tests (Brennan, 2008). Therefore, the magnitude of the DTM in the present study was defined as follows: averaging half of the differences between Rasch scores of CBM1 because the item parameter estimates of CBM 1 served as the metric to the other four forms. Then the DTM was standardized by dividing by σ_{Y_P} and denoted as the SDTM. If the (ew)RMSD, (ew)REMSD, and $RESD_j$ statistics are less than the SDTM, the differences between the whole group and ability subgroups in linking functions are be considered as negligible.

Chapter Four

Results

In this chapter, the (ew)RMSD, (ew)REMSD, and $RESD_j$ statistics based on the five CBM administrations are reported for the four linking functions across the whole group and subgroups differing in ability. The SDTM was used as a benchmark to assess whether the linking functions were population invariant with the whole group and ability subgroups. By and large, all the (ew)RMSD, (ew)REMSD, and $RESD_j$ statistics were smaller than the SDTM which indicated that the linking functions were population invariant with respect to the ability subgroups across the five forms of CBM.

Regarding Rasch fit statistics, the mean values of the infit MNSQ and outfit MNSQ across the five forms were close to the expected value of 1 with a range (0.99-1.10) well within the conventional acceptable range (0.50-1.50). The summary statistics of the raw scores (e.g., number of correct responses) on all the forms for the whole group, AA group, and LA/LD group are given in Table 2. The AA students had much higher mean raw scores on all the forms than the LA students and students with LD. The differences of the mean raw scores on all the forms between the AA students and the LA students and students with LD were larger than these differences between the whole group and the LA group and the LD group. The LA students only had a much higher mean raw score on CBM 3 than the students with LD. The differences of the mean raw scores on CBM 1 and CBM 4 between the students with LD and the LA students were close to zero. Noticeably, the students with LD even had much higher mean raw scores on CBM 2 and CBM 5 than the LA students. The students with LD had bigger standard deviations of raw scores on all the forms than the AA and LA students.

Table 2 *Raw Score Descriptive Statistics*

CBM1			
Group	N	M	SD
Whole	1578	4.75	2.529
AA	536	6.46	2.203
LA	912	3.88	2.176
LD	130	3.85	2.492

CBM 2			
Group	N	M	SD
Whole	714	4.77	2.615
AA	205	6.00	2.506
LA	445	4.22	2.485
LD	64	4.58	2.537

CBM 3			
Group	N	M	SD
Whole	1422	5.05	2.814
AA	499	6.71	2.472
LA	797	4.20	2.554
LD	126	3.83	2.650

CBM 4			
Group	N	M	SD
Whole	676	6.27	2.522
AA	195	7.26	2.170
LA	413	5.88	2.528
LD	68	5.84	2.669

CBM 5			
Group	N	M	SD
Whole	1315	4.72	2.775
AA	413	6.75	2.355
LA	784	3.72	2.392
LD	118	4.25	2.693

Since the sample size of the students with LD was very small, the LA students and students with LD were collapsed into the LA/LD group. In the Rasch mean/mean equating, all the differences in difficulty across the forms were adjusted and all the item parameters were on the same metric. Table 3 presents the Rasch scores based on the Rasch mean/mean equating on the five forms for the whole group, AA group, and LA/LD group. Consistent with the raw score outcomes, the students in the AA group had higher mean ability estimates across the five forms than the students in the LA/LD group. It indicated that they might have higher abilities of solving math word problems than the other students. They also had smaller standard deviations of ability estimates across the five forms than both the whole group and the LA/LD group. The differences of mean Rasch scores on the five forms between the whole group and the AA group were larger than those differences between the whole group and the LA/LD group.

Table 3

Rasch Score Descriptive Statitics

CBM1			
Group	N	M	SD
Whole	1578	-0.13	1.714
AA	536	0.89	1.392
LA/LD	1042	-0.75	1.622

CBM 2			
Group	N	M	SD
Whole	714	-0.03	1.696
AA	205	1.16	1.564
LA/LD	509	-0.54	1.649

CBM 3			
Group	N	M	SD
Whole	1422	0.41	1.827
AA	499	1.25	1.548
LA/LD	923	-0.13	1.721

CBM 4			
Group	N	M	SD
Whole	676	0.65	1.649
AA	195	1.46	1.437
LA/LD	481	0.29	1.695

CBM 5			
Group	N	M	SD
Whole	1315	0.5	1.867
AA	413	1.54	1.530
LA/LD	902	-0.04	1.727

As one of the five requirements for equating, reliability should be equal. Table 4 lists Cronbach's alpha for summated scores for the whole, AA, and LA/LD groups across the five forms. It also shows person reliability coefficients on the five forms for the whole, AA, and LA/LD groups before and after the Rasch mean/mean equating was conducted. The person reliability under the Rasch model is comparable to Cronbach's alpha. It estimates the replicability of person ordering if the sample of persons takes another set of items measuring the same construct (Bond & Fox, 2007). Rasch person reliability was noticeably lower than Cronbach's alpha for the three groups across the five forms. In the Rasch model, an extreme score is considered as providing little information about that person's location on the infinite latent trait, so a test has an infinitely large standard error. Including persons with extreme scores can increase measurement error and lower reliability. The whole group had higher reliability than the AA and LA/LD groups, which was expected as the same sample sizes in the whole group were larger and more sufficient for conducting the Rasch mean/mean equating. There were no large differences between the unscaled reliability and scaled reliability with all three groups across the four forms. The whole group had close or equal reliability based on summated scores across the five forms. The AA subgroup had close or equal reliability based on summated scores for CBM 2, 3, and 5. The LA/LD subgroup had close or equal reliability based on summated scores for the forms except CBM 1.

Table 4

Reliability Coefficients

		Reliability Coefficients		
		Whole	AA	LA/LD
CBM 1				
	Summated	0.74	0.63	0.68
	Rasch (non-linked)	0.70	0.55	0.66
CBM2				
	Summated	0.75	0.73	0.73
	Rasch (non-linked)	0.70	0.67	0.67
	Linked	0.70	0.67	0.67
CBM3				
	Summated	0.79	0.73	0.74
	Rasch (non-linked)	0.70	0.60	0.60
	Linked	0.70	0.59	0.60
CBM4				
	Summated	0.74	0.67	0.73
	Rasch (non-linked)	0.64	0.45	0.67
	Linked	0.65	0.45	0.68
CBM5				
	Summated	0.77	0.70	0.71
	Rasch (non-linked)	0.71	0.56	0.65
	Linked	0.72	0.57	0.67

CBM 4 → CBM 1

The first Rasch mean/mean linking was conducted by linking CBM 4 to CBM 1 because they had the common items 1 to 5. Figure 2 depicts the proportionally weighted RMSD outcomes at each score level, proportionally weighted REMSD value across score levels, $RESD_j$ values for ability subgroups, and the SDTM. The RMSD values at all score levels ranged from 0.042 to 0.164 and were much smaller than 0.261 (SDTM). This suggested that the Rasch mean/mean linking functions for each subgroup did not differ significantly from the whole group linking function. The linking differences at all score levels were negligible. Figure 2 also shows that the REMSD value was lower than the SDTM. The very small $RESD_j$ values for both the AA and LA/LD subgroups indicated a negligible linking difference between each subgroup and the whole group. In addition, the linking difference between the LA/LD subgroup and the whole group was smaller than that between the AA subgroup and the whole group because the majority of the students were LA students. Figure 3 displays the ewRMSD outcomes at each score level, ewREMSD value across score levels, $RESD_j$ values for the ability subgroups, and SDTM. The ewRMSD values at all score levels ranged from 0.049 to 0.165 and were much smaller than the SDTM. The ewREMSD value was slightly larger than the proportionally weighted REMSD but still well below the SDTM. The same results were found in the equally weighted indices as were found in the proportional weighted indices. The linking functions for each subgroup did not differ significantly from the whole group linking function. The linking differences at all score levels were negligible.

Figure 2. CBM 4-to-CBM 1 Linking Differences between Each Ability Group and Whole Group Using Proportional Weights

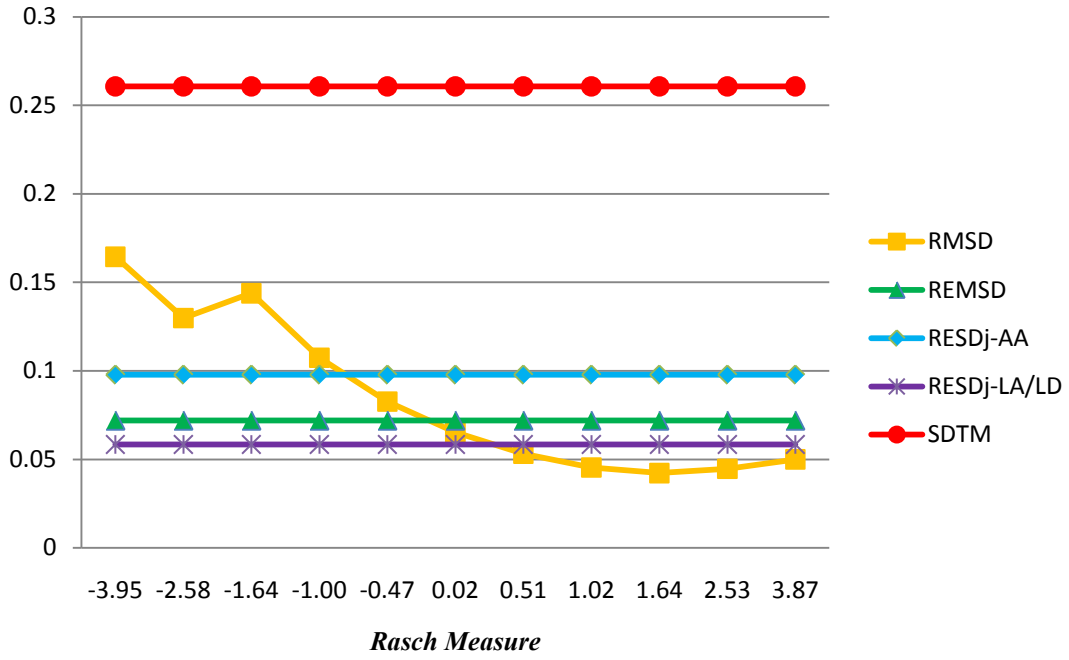
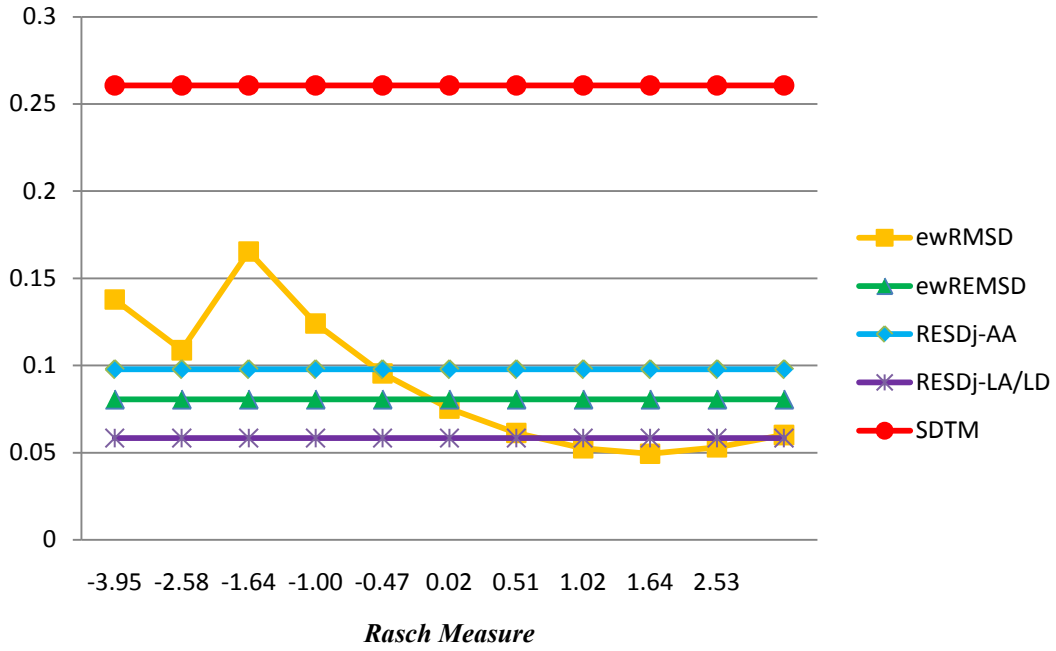


Figure 3. CBM 4-to-CBM 1 Linking Differences between Each Ability Group and Whole Group Using Equal Weights



CBM 2 → CBM 4

Since CBM 2 and CBM 4 had the common items 16 to 20, CBM2 was linked to CBM 4 using the Rasch mean/mean linking method. Figure 4 displays the proportionally weighted RMSD outcomes at each score level, proportionally weighted REMSD value across score levels, $RESD_j$ values for this specific linking, and SDTM. The RMSD values at all score levels ranged from 0.146 to 0.189 and were much smaller than 0.261 (SDTM). This suggested that the Rasch mean/mean linking functions for each subgroup did not differ significantly from the whole group linking function. The linking differences at all score levels were negligible. Figure 4 also represents that the REMSD value was well below the SDTM. The very small $RESD_j$ values for both the AA and LA/LD subgroups indicated a negligible linking difference between each subgroup and the whole group, although the $RESD_j$ value for the AA subgroup was very close to the SDTM. Consistent with the results found in the linking function between CBM 4 and CBM 1, the linking difference between the LA/LD subgroup and the whole group was smaller than that between the AA subgroup and the whole group because the sample size of the LA/LD students doubled that of the AA students. Figure 5 shows the ewRMSD outcomes at each score level, ewREMSD value across score levels, $RESD_j$ values for the ability subgroups, and SDTM. The ewRMSD values at all score levels ranged from 0.173 to 0.222 and were much smaller than the SDTM. The ewREMSD value was slightly larger than the proportionally weighted REMSD value but still well below the SDTM. The same results were found in the equally weighted indices as were found in the proportional weighted indices. The linking functions for each subgroup did not differ

significantly from the whole group linking function. The linking differences at all score levels were negligible.

Figure 4. CBM 2-to-CBM 4 Linking Differences between Each Ability Group and Whole Group Using Proportional Weights

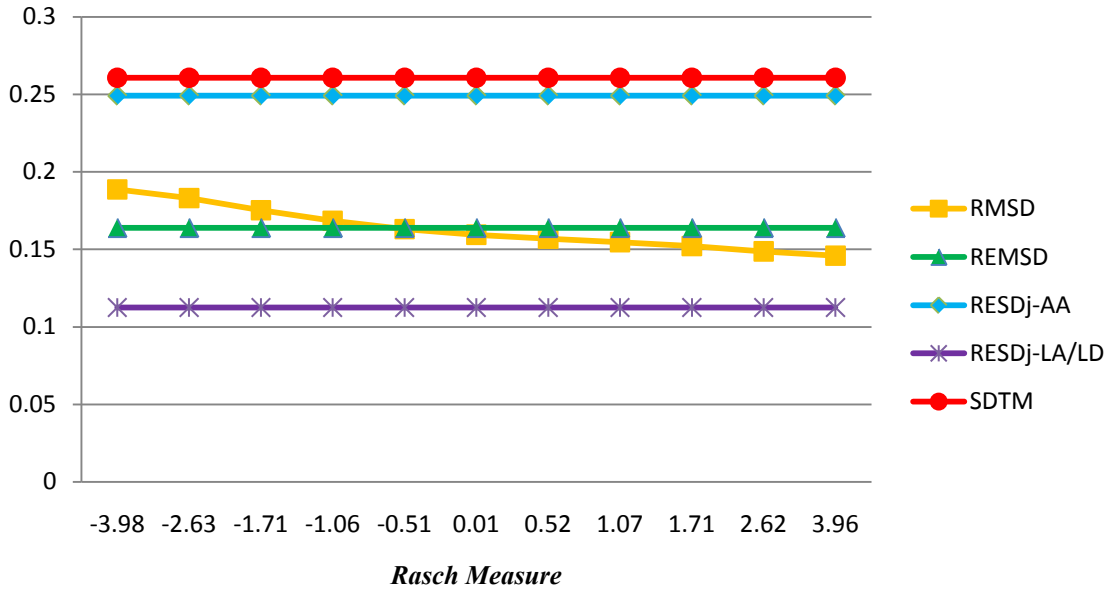
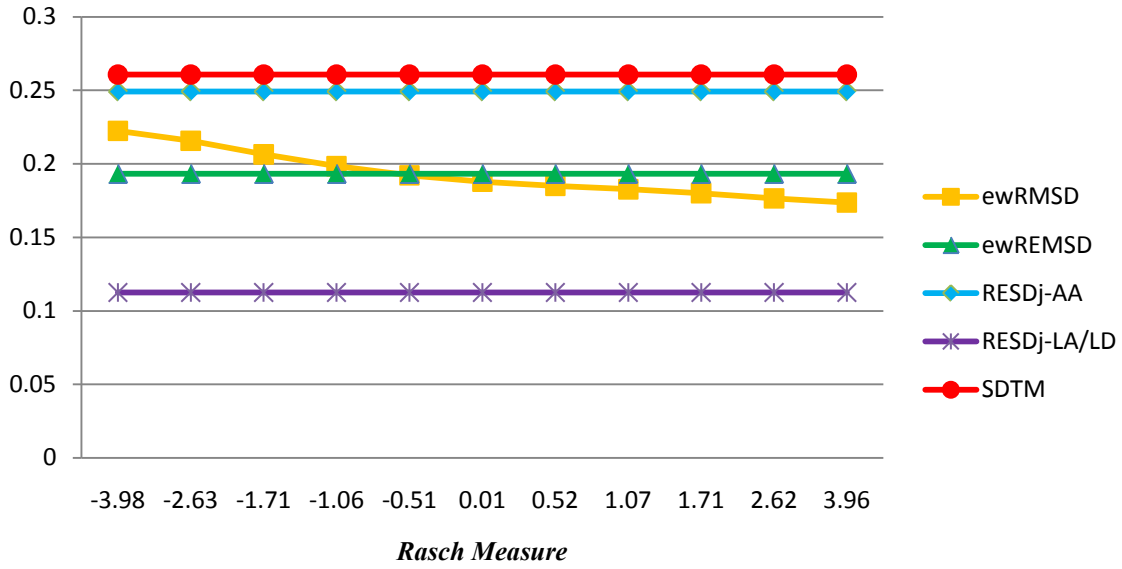


Figure 5. CBM 2-to-CBM 4 Linking Differences between Each Ability Group and Whole Group Using Equal Weights



CBM 5 → CBM 1

CBM 5 was linked to CBM 1 because they shared the items 6 to 10. Figure 6 depicts the proportionally weighted RMSD outcomes at each score level, proportionally weighted REMSD value across score levels, $RESD_j$ values for this linking, and SDTM. The RMSD values at all score levels ranged from 0.015 to 0.144 and were much smaller than 0.261 (SDTM). This indicated that the Rasch mean/mean linking functions for each subgroup did not differ significantly from the whole group linking function. The linking differences at all score levels were negligible. Figure 6 also shows that the REMSD value was well below the SDTM. The very small $RESD_j$ values for both the AA and LA/LD subgroups indicated a negligible linking difference between each subgroup and the whole group. As expected, the linking difference between the LA/LD subgroup and the whole group was smaller than that between the AA subgroup and the whole group mainly because there were more LA students and students with LD in the whole group. Figure 7 shows the ewRMSD outcomes at each score level, ewREMSD value across score levels, $RESD_j$ values for the ability subgroups, and SDTM. The ewRMSD values at all score levels ranged from 0.016 to 0.162 and were much smaller than the SDTM. The ewREMSD value was slightly larger than the proportionally weighted REMSD value but still well below the SDTM. The same results were found in the equally weighted indices as were found in the proportional weighted indices. The linking functions for each subgroup did not differ significantly from the whole group linking function. The linking differences at all score levels were negligible.

Figure 6. CBM 5-to-CBM 1 Linking Differences between Each Ability Group and Whole Group Using Proportional Weights

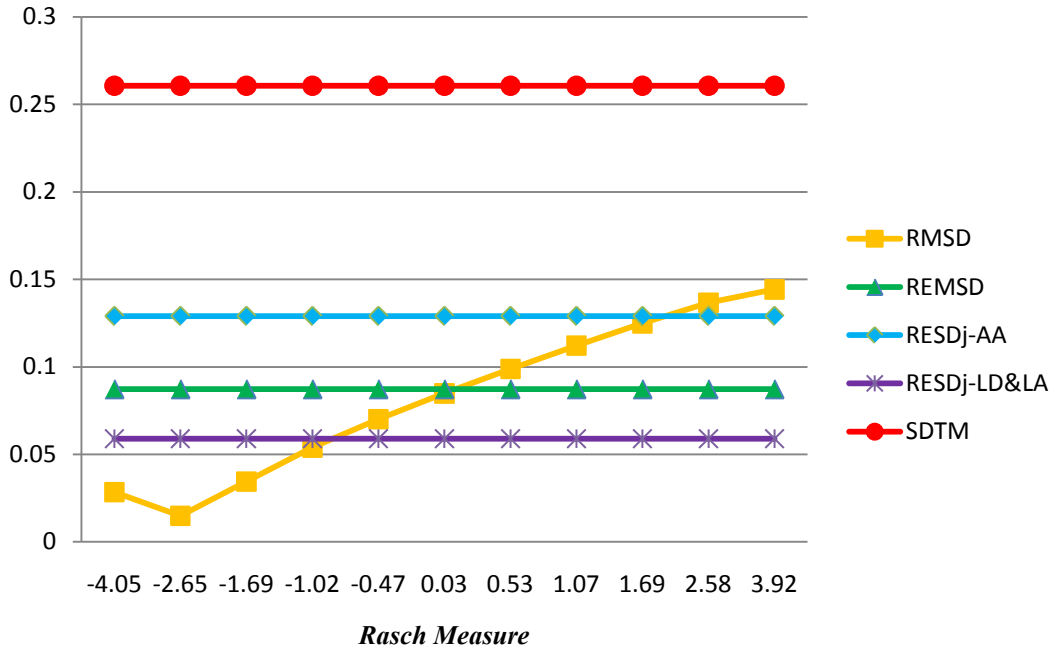
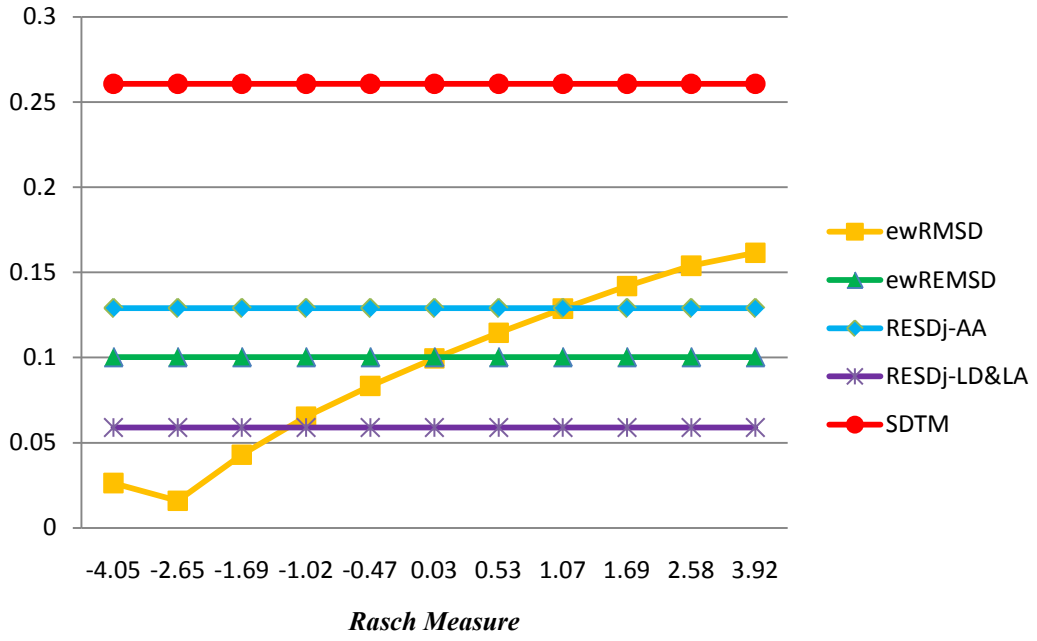


Figure 7. CBM 5-to-CBM 1 Linking Differences between Each Ability Group and Whole Group Using Equal Weights



CBM 3 → CBM 5

The last Rasch mean/mean linking was conducted in order to link CBM 3 to CBM 5 through their common items 21 to 25. Figure 8 shows the proportionally weighted RMSD outcomes at each score level, proportionally weighted REMSD value across score levels, $RESD_j$ values for this particular linking, and SDTM. The RMSD values at all score levels ranged from 0.016 to 0.123 and were much smaller than 0.261 (SDTM). This suggested that the Rasch mean/mean linking functions for each subgroup did not differ significantly from the whole group linking function. The linking differences at all score levels were negligible. Figure 8 also displays that the REMSD value was well below the SDTM. The very small $RESD_j$ values for both the AA and LA/LD subgroups indicated a negligible linking difference between each subgroup and the whole group. Additionally, the linking functions in the LA/LD and AA subgroups were similar to that in the whole group. Same as the other three linking functions, the linking difference between the LA/LD subgroup and the whole group was smaller than that between the AA subgroup and the whole group. Figure 9 depicts the ewRMSD outcomes at each score level, ewREMSD value across score levels, $RESD_j$ values for the ability subgroups, and SDTM. The ewRMSD values at all score levels ranged from 0.015 to 0.136 and were much smaller than the SDTM. The ewREMSD value was slightly larger than the proportionally weighted REMSD value but still well below the SDTM. The same results were found in the equally weighted indices as were found in the proportional weighted indices. The linking functions for each subgroup did not differ significantly from the whole group linking function. The linking differences at all score levels were negligible.

Figure 8. CBM 3-to-CBM 5 Linking Differences between Each Ability Group and Whole Group Using Proportional Weights

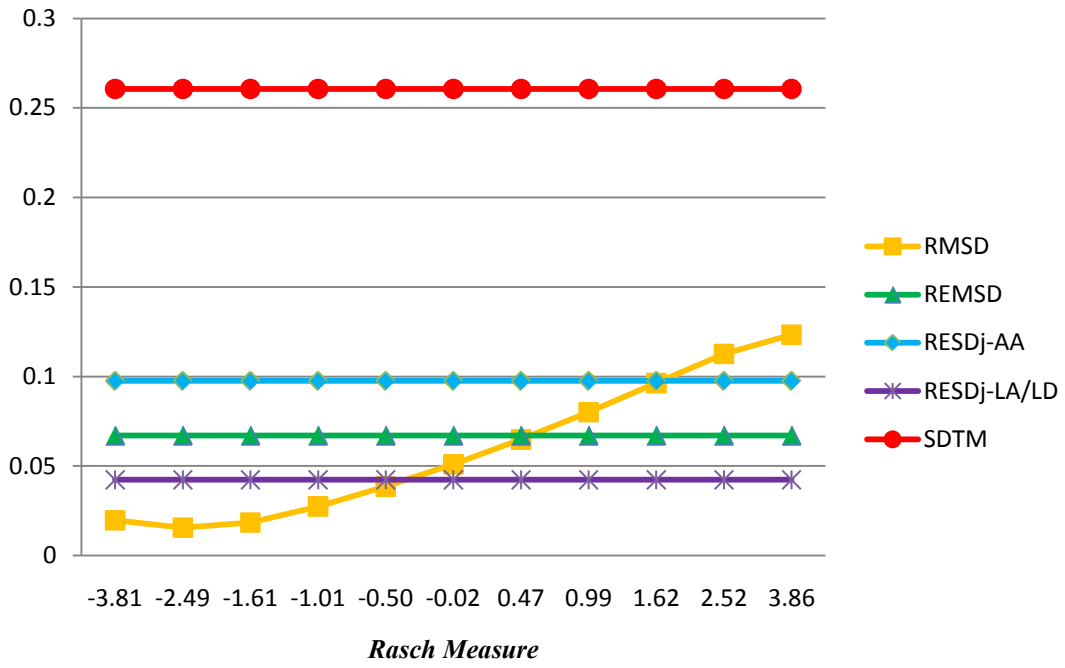


Figure 9. CBM 3-to-CBM 5 Linking Differences between Each Ability Group and Whole Group Using Equal Weights

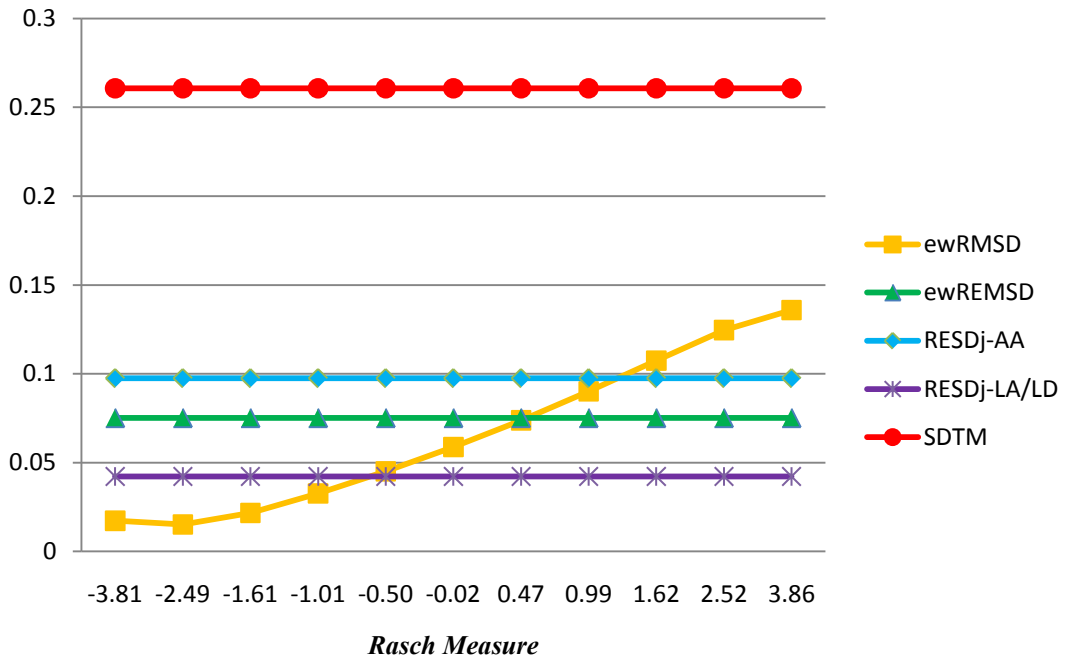


Table 5 shows the REMSD, ewREMSD, and $RESD_j$ statistics for all the linking functions, respectively. As with the RMSD statistics presented earlier, the REMSD values ranged from 0.054 to 0.164, well below the SDTM. The ewREMSD statistics ranging from 0.075 to 0.193 were also smaller than the SDTM. The results suggested that linking functions were population invariant with respect to the ability subgroups across the five forms.

Table 5 *Equatability Measures for Each Linking Function on Each Subgroup.*

	Equatability Index			
	$RESD_j$ for Each Subgroup			
	AA	LA/LD	REMSD	ewREMSD
CBM 4 → CBM1	0.0158	0.0162	0.0720	0.0806
CBM 2 → CBM4	0.0392	0.0290	0.1639	0.1933
CBM 5 → CBM1	0.0268	0.0117	0.0543	0.1003
CBM 3 → CBM5	0.0219	0.0080	0.0671	0.0752

Chapter Five

Discussion

The present study illustrated the procedure of examining population invariance of the Rasch mean/mean equating method in the NEAT design. The (ew)RMSD, (ew)REMSD, and $RESD_j$ statistics were used to explore the degree of population dependence of linking functions of curriculum-based measures on the whole, AA, and LA/LD groups. As described earlier, the five forms of CBM were linked as follows: CBM 4 was linked to CBM 1, CBM 2 to CBM 4, CBM 5 to CBM 1, and CBM3 to CBM5. This study demonstrated how population invariance could be applied to evaluate linking functions involving the Rasch mean/mean equating for five alternate forms of the math word problem solving measure.

Findings

Overall, the linking functions were population invariant with respect to the subgroups defined by ability for all of the five forms of CBM. Consistent patterns were found in linking differences between ability subgroups across the forms. All the effect sizes of invariance exhibited acceptable levels of population dependence, in that the dependence was minimal enough to be considered not practically problematic. Results from equating these five parallel forms of CBM showed very little evidence of population sensitivity of linking functions across the subgroups defined by ability. The differences between the linking functions based on ability groups and the linking function based on the whole group were negligible for all of the five forms. The equatability indices were small enough to suggest negligible linking differences. However, it is apparent that the $RESD_j$ value for the AA subgroup was very close to the SDTM when linking CBM 2 to

CBM 4. It might be due to second-order linking. When more than one linking is conducted, more errors are introduced in the linking procedure.

As discussed before, population invariance is one of the important prerequisites for equating. However, "population invariance is a necessary but not sufficient condition for equating" (Brennan, 2008, p.102). When linking functions are population invariant and meet the other important prerequisites for equating, alternate forms of a test are equated and differences in the difficulty of alternate forms are adjusted. Then the equated scores can be used interchangeably regardless of the form administered or sample of population tested. In the present study, the five forms of CBM had close or equal reliability based on summated scores with the whole group, although the reliability coefficients were slightly below 0.8. Nevertheless, Liu and Holland (2008) addressed that "it is difficult to prescribe precisely how much reliability is needed for equating to be population invariant" (p. 40). Brennan (2008) also mentioned that "high reliability is not a necessary condition for population invariance" (p. 106). In conclusion, the five forms of CBM measured the same construct, had symmetry and equal reliability, and population invariance held appropriately and reasonably well with the ability subgroups, so the equated scores can be used interchangeably across forms.

Traditional assessment for identifying students with LD relies on the IQ-achievement discrepancy. In 2004, the IDEA included RtI as an alternative method to identify students with LD. RtI employs response to intervention at various tiers to identify students with LD. Students are provided effective instruction in the general education setting. Students who are at risk of academic failure are identified by a percentile cutoff on a screening measure: a norm-referenced test or a cutoff point on a

CBM test. The at-risk students are assessed using progress monitoring. Students unresponsive to primary prevention receive research-based preventative treatment during which progress is monitored frequently. Students who are unresponsive to secondary treatment receive interventions. Student progress is continually monitored to determine effectiveness of instructional programs. Progress monitoring is a vital aspect of the RTI model. With the implementation of RtI models in schools, educational measurement plays an increasingly important role. Typically, curriculum-based measures are used to identify the students who are not meeting expected levels of performance at various tiers. Teachers assess students' academic performance using CBM. CBM takes place frequently and each alternate test form assesses performance of what is expected at the end of the school year. Teachers determine what additional instruction students might need to address the learning gap by monitoring students' progress. Scores on CBM are viewed as an indicator of overall student performance. CBM can produce accurate, meaningful information about students' academic levels and their rates of improvement, and CBM corresponds well with high-stakes tests. Therefore, how to accurately interpret and use CBM scores should be receiving much more attention than before. To evaluate students' progress, curriculum-based measures can be used efficiently, easily, and frequently. That is, equivalent forms of CBM are undoubtedly needed. Creating equivalent forms of CBM is the challenge. When researchers develop and administer alternate forms of curriculum-based measures, they should not only consider the technical adequacy such as reliability and validity, but also address whether the alternate forms of curriculum-based measures are equivalent or whether the test scores obtained from the alternate forms can be used comparably and interchangeably.

The NCLB (2001) requires that all students be assessed and that schools show academic growth and progress on these assessments. To accomplish this goal, educators should use tools to identify students who are at risk of academic failure and adjust instruction to better meet individual students' needs. Curriculum-based measures can be employed to help teachers monitor student progress over time, evaluate the effectiveness of their teaching, and make reasonable and accurate instructional decisions. CBM information can be used for school accountability and specific goals and objects for IEPs. At the beginning of NCLB implementation, schools assess every student using CBM to identify the number of students who initially meet CBM benchmarks, which represents the school's initial proficiency status. To derive the discrepancy between initial proficiency status and universal proficiency, schools subtract the initial proficiency status from the total number of students in the school. It is important to also note that the CBM benchmark for proficiency becomes more stringent each year as students advance through the grades. Therefore, a school must assess and document the number of proficient students and its corresponding AYP each year based on the current student body. Although studies on CBM investigated reliability and validity of the measures, most did not attend to equating alternate forms of CBM. Equating and linking methods will adjust for minor differences in form-to-form difficulty and produce equivalent or comparable scores. Therefore, schools can use these scores to explain what students have learned, what they are still missing, and how instruction can be improved.

Limitations

In the present study, there are some limitations that need to be mentioned. First of all, reliability was moderate at best. There are some factors that may have an impact on

reliability including test length, item type and quality, conditions of administration, and the group of examinees (Traub & Rowley, 1991). The number of items can exert an influence on reliability because the error variance increases at a slower rate than the true score variance when the number of items increases. Based on test information function, item information for Rasch item is $I_i(\theta) = P_i(\theta)[1 - P_i(\theta)]$ and maximum information is at $P_i(\theta) = 0.5$, in which case $I_i(\theta) = 0.25$. For 10 items, maximum possible test information is given by $\sum_i I_i(\theta) = 10 \times 0.25 = 2.5$. And $SEM = \frac{1}{\sqrt{2.5}} = 0.63$ which is really big but this is the best-case scenario. Then, if assuming $Var(\hat{\theta}) \approx 2.2$, $P^2_{\hat{\theta}} = 1 - \frac{SEM^2}{Var(\hat{\theta})} = 1 - \frac{(0.63)^2}{2.2} = 0.82$. Nevertheless, this is a completely best-case scenario of all items having the same *b* and θ = *b*. Therefore, reliability increases when the number of items increases. In addition, the properties of items can have an effect on reliability. Items with poor properties can reduce reliability. If items in a test have low discrimination or are extremely easy or difficult, more errors will be produced and reliability will be low. As a result, Rasch fit statistics need to be carefully reviewed. For example, the outfit MNSQ of the first item in CBM 1 and CBM 4 was 2.86 for the whole group and 2.72 for the LA/LD group, respectively. That is to say, this particular item did not fit the Rasch model. Nonetheless, the overall infit MNSQ and outfit MNSQ statistics had a good fit across the five forms. Furthermore, the conditions of administration varied from one class to another. Some teachers gave the participating students enough time to finish the test but some teachers did not. Some classes had a good testing environment but some did not. The test scores unavoidably differed for some reasons other than

differences among examinees in their abilities to solve math word problems. Thus, reliability decreased by introducing unpredictable sources of random variation and measurement error into the test scores (Bond & Fox, 2007). Besides, the population homogeneity might decrease the reliability because the majority of the students were low achieving students. Their true differences in abilities in the population are small. Second, the sample size of the students with LD is not big, so population invariance of linking functions for the LD group could not be conducted. Since there are differences between the LA students and students with LD, these differences might be ignored when these two groups of students were collapsed. Third, only five alternate forms of CBM were included in this study rather than the full six forms. The linking procedures for CBM 2 and CBM 6 and for CBM 6 and CBM 3 were missing. Therefore, it is hard to have the full picture of linking across all these alternate forms of CBM.

Implications for Future Research

More extensive research is needed to compute the standard errors of equating in order to investigate whether the differences found in the study can be attributed to random error. In addition, if there is a big enough sample of students with LD available, researchers can compare the linking functions for the students with LD to the linking functions for other students. Researchers can also investigate population dependence of the linking functions with subgroups defined by gender and ethnicity because some studies have found that linking functions may be more similar for gender subgroups than ethnicity subgroups (e.g., Liu & Holland, 2008; Yang & Gao, 2008). Therefore, the gender and ethnicity variable might be interesting for linking functions in the future

research. So far, most studies on population invariance have investigated the linking procedure conducted with a single step, which is linking test A to test B or linking parallel forms of a test to a reference form. However, the linking procedures in the present study were conducted in multiple steps (i.e., linking CBM 4 to CBM 1, CBM 2 to CBM 4, CBM 5 to CBM 1, and CBM 3 to CBM 5). There is no study that conducts linking processes in such a complex way, so the quality of linking procedures with multiple steps needs to be examined in the future.

To summarize, the present study found that the Rasch mean/mean linking functions were population invariant with respect to the AA students and the LA/LD students. There were no significant differences between the linking function defined by the whole group and the linking functions defined by the ability subgroups. Therefore, the equated scores are comparable and can be used interchangeably for monitoring students' progress over time.

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

Betts, J., Pickart, M., & Heistad, D. (2009). An investigation of the psychometric evidence of CBM-R passage equivalence: Utility of readability statistics and equating for alternate forms. *Journal of School Psychology, 47,* 1-17.

Bond, T.G. & Fox, C.M. (2007). *Applying the Rasch Model: Fundamental measurement in the human sciences* (*2nd*). Mahwah, NJ: Lawrence Erlbaum.

Brigance, A. (1985). *Brigance Preschool Screen*. North Billerica, MA: Curriculum Associates, Inc.

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

Brennan, R. L. (2008). A discussion of population invariance. *Applied Psychological Measurement, 32*(1), 102-114.

Busch, T. W, & Reschly, A. L. (2007). Progress monitoring in reading. *Assessment for Effective Intervention, 32*, 223-230.

Carver, R. P. (1992). What do standardized tests of reading comprehension measure in terms of efficiency, accuracy, and rate? *Reading Research Quarterly, 27*, 347-359.

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

Christ, T. J., & Ardoin, S. P. (2009). Curriculum-based measurement of oral reading: Passage equivalence and probe-set development. *Journal of School Psychology, 47,* 55-75.

Clarke, B., & Shinn, M. R. (2004). A preliminary investigation into the identification and development of early mathematics curriculum-based measurement. *School Psychology Review, 33*, 234-248.

Compton, D. L., Fuchs, D., Fuchs, L. S., & Bryant, J. D. (2006). Selecting at-risk readers in first grade for early intervention: A two-year longitudinal study of decision rules and procedures. *Journal of Educational Psychology, 98*, 394-409.

Daly, E. J. III, Wright, J. A., Kelly, S. Q., & Martens, B. K. (1997). Measures of early academic skills: Reliability and validity with a first grade sample. *School Psychology Quarterly, 12*, 268-280.

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

Deno, S. L. (2003a). Developments on curriculum-based measurement. *The Journal of Special Education, 37,* 184-192.

Deno, S. L. (2003b). Curriculum-based measures: Development and perspectives. *Assessment for Effective Intervention, 28,* 3-12.

Deno, S. L., & Fuchs, L. S. (1987). Developing curriculum-based measurement systems for data-based special education problem solving. *Focus on Exceptional Children, 19*, 1-15.

Deno, S. L., Fuchs, L. S., Marston, D. & Shin, J. (2001). Using curriculum-base measurement to establish growth standards for students with learning disabilities. *School Psychology Review, 30,* 507-524.

Deno, S. L., Mirkin, P., & Marston, D. (1980). Valid measurement procedures for continuous evaluation of written expression. *Exceptional Children Special Education and Pediatrics: A New Relationship, 48*, 368-371.

Deno, S. L., Marston, D., Mirkin, P., Lowry, L., Sindelar, P., & Jenkins, J. (1982). The use of standard tasks to measure achievement in reading, spelling, and written expression: A normative and developmental study (Vol. IRLD-RR-87). University of Minnesota, Institute for Research on Learning Disabilities.

Deno, S. L., & Mirkin, P. K. (1977). *Data-based program modification: A manual*. Reston. VA: Council for Exceptional Children.

Deno, S. L., & Mirkin, P. K., & Chiang, B. (1982). Identifying valid measures of reading. *Exceptional Children, 49*, 36-45.

Dorans, N.J. (2004). Using subpopulation invariance to assess test score equity. *Journal of Educational Measurement, 41*, 43-68.

Dorans, N. J., & Feigenbaum, M. D. (1994). Equating issues engendered by changes to the new SAT and PSAT/NMSQT. In I. M. Lawrance, N. J. Dornas, M. D. Feigenbaum, N. J. Feryok, A. P. Schmitt, & N. K. Wrights (eds.), *Technical issues related to the introduction of the new SAT and PSAT/NMSQT* (ETS Research Memorandum No. RM-94-10). Princeton, NJ: Educational Tresting Service.

Dorans, N. J., & Holland, P. W. (2000). Population invariance and equitability of tests: Basic theory and the linear case. *Journal of Educational Measurement, 37*, 281-306.

Dorans, N. J., Pommerich, M., & Holland, P. W. (2007). *Linking and Aligning Scores and Scales.* New York: Springer-Verlag.

Embretson, S. E., & Hershberger, S. L. (Eds.). (1999). *The new rules of measurement: What every psychologist and educator should know*. Mahwah, NJ: Lawrence Erlbaum.

Embretson, S. E., & Reise, S. P. (2000) *Item Response Theory for Psychologists*. Mahwah, NJ: Lawrence Erlbaum.

Espin, C. A., De La Pza, S., Scierka, B. J., & Roelofs, L. (2005). The relationship between curriculum-based measures in written expression and quality and completeness of expository writing for middle school students. *The Journal of Special Education, 38*, 208-217.

Espin, C. A., Shinn, J., Deno, S. L., Skare, S., Robinson, S., & Benner, B. (2000). Identifying indicators of written expression proficiency for middle school students. *The Journal of Special Education, 34,* 140-153.

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

Foegen, A. (2000). Technical adequacy of general outcome measures for middle school students. *Diagnostique, 25*, 175-203.

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

Fuchs, L. S. (2004). The past, present, and future of curriculum-based measurement research. *School Psychology Review, 33*, 188-192.

Fuchs, L. S., & Deno, S. L. (1992). Effects of curriculum within curriculum-based measurement. *Exceptional Children, 58*, 232-242.

Fuchs, L. S., Deno, S. L., & Maston, D. (1982). Use of aggregation to improve the reliability of simple direct measures of academic performance (Vol. IRLD-RR-94). University of Minnesota, Institute for Research on Learning Disabilities.

Fuchs, L. S., & Fuchs, D. (1986). Effects of systematic formative evaluation: A meta-analysis. *Exceptional Children, 53*, 199-208.

Fuchs, L. S., & Fuchs, D. (1992). Identifying a measure for monitoring student reading progress. *School Psychology Review, 21*, 45-58.

Fuchs, L. S., & Fuchs, D. (2006). Introduction to responsiveness-to-intervention: What, why, and how valid is it? *Reading Research Quarterly, 41*, 92-99.

Fuchs, L. S., Fuchs, D., & Compton, D. L. (2004). Monitoring early reading development in first grade: Word identification fluency versus nonsense word fluency. *Exceptional Children, 71*, 7-21.

Fuchs, L. S., Fuchs, D., & Courey, S. J. (2005). Curriculum-based measurement of mathematics competence: From computation to concepts and applications to real-life problem solving. *Assessment for Effective Intervention, 30,* 33-46.

Fuchs, L. S., & Fuchs, D., & Hamelett, C. L. (1989). Effects of alternative goal structures within curriculum-based measurement. *Exceptional Children, 55*, 429-438.

Fuchs, L. S., & Fuchs, D., & Hamelett, C. L., & Allinder, R. M. (1989). The reliability and validity of skills analysis within curriculum-based measurement. *Diagnostique, 14*, 203-221.

Fuchs, L. S., & Fuchs, D., Hamelett, C. L., & Ferguson, C. (1992). Effects of expert system consultation within curriculum-based measurement using a reading maze task. *Exceptional Children, 58*, 436-450.

Fuchs, L. S., Fuchs, D., Hamelett, C. L., & Stecker, P. M. (1990). The role of skills analysis in curriculum-based measurement in math. *School Psychology Review, 19*, 6-22.

Fuchs, L. S., Fuchs, D., Hamelett, C. L., & Stecker, P. M. (1991). Effects of curriculum-based measurement and consultation on teacher planning and student achievement in mathematics operations. *American Educational Research, 28*, 617-641.

Fuchs, L. S., Fuchs, D., Hamelett, C. L.,Thompson, A., Roberts, P. H., Kupek, & Stecker, P. M. (1994). Technical features of a mathematics concepts and applications curriculum-based measurement system. *Diagnostique, 19,* 23-49.

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

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

Fuchs, D., Fuchs, L. S., McMaster, K., & Al Otaiba, S. (2003). Identifying children at risk for reading failure: Curriculum-Based Measurement and the dual discrepancy approach. In L. Swanson, K. R., Harris, & S. Graham (Eds.), *Handbook of Learning Disabilities* (pp. 431-449). New York: Guilford.

Fuchs, L. S., Hamlett, C. L., & Fuchs, D. (1998). *Monitoring basic skills progress: Basic math concepts and applications* [computer program manual]. Austin, TX: PRO-ED.

Fuchs, L. S., Hamlett, C. L., & Fuchs, D. (1999). *Monitoring basic skills progress: Basic math concepts and applications* (2nd ed.) [computer program manual]. Austin, TX: PRO-ED.

Ginsburg, H. P., & Baroody, A. J. (1990). *Test of early mathematics ability* (2nd ed.). Austin, TX: Pro-ed.

Good, R. H., III, & Kaminski, R. A. (1996) Assessment for instructional decisions: Toward a proactive/prevention model of decision-making for early literacy skills. *School psychology Quarterly, 11,* 326-336.

Hambleton, R. K. & Jones, R. W. (1993). Comparison of classical test theory and item response theory and their applications to test development. *Educational Measurement: Issues and Practice, 12*, 38-47.

Hammill, D. D., & Larsen, S. C. (1978). *The test of written language*. Austin, TX: PRO-ED.

Harcourt Brace & Company. (1996). *Stanford achievement test* (9th ed.). San Antonio, TX: Harcourt Brace Jovanovich.

Helwig, R., Anderson, L., & Tindal, G. (2002). Using a concept-grounded, curriculum-based measure in mathematics to predict statewide test scores for middle school students with LD. *The Journal of Special Education, 36,* 102-112.

Helwig, R., & Tindal, G. (2002) Using general outcomes measures in mathematics to measure adequate yearly progress as mandated by Title I. *Assessment for Effective Intervention, 28,* 9-18.

Hoover, H. D., Hieronymus, A. N., Fribie, D. A., & Dunbar, S. B. (1993). *Iowa Tests of Basic Skills*. Chicago: Riverside.

Individuals with Disabilities Education Improvement Act, 20 U.S.C 1400 et seq. (2004).

Jenkins, J. R., & Jewell, M. (1993). Examining the validity of two measures for formative teaching: Reading aloud and maze. *Exceptional Children, 59*, 421-432.

Jitendra, A. K., Sczesniak, E., & Deatline-Buchman, A. (2005) An exploratory validation of curriculum-based mathematical word problem-solving tasks as indicators of mathematics proficiency for third graders. *School Psychology Review, 34*, 358-371.

Kaminski, R. A., Good, R. H., III. (1996). Toward a technology for assessing basic early literacy skills. *School Psychology Review, 25*, 215-227.

Kolen, M.J. (2004). Population invariance in equating and linking: Concept and history. *Journal of Educational Measurement, 41*, 3-14.

Kolen, M. J., & Brennan, R. L. (1987). Linear equating models for the common-item nonequivalent-populations design. *Applied Psychological Measurement, 11*, 263-277.

Kolen, M.J., & Brennan, R.L. (2004). *Test equating, scaling, and linking: Methods and practices* (2nd Ed.). NY: Springer.

Lee L., & Canter, S. M. (1971). Developmental sentence scoring. *Journal of Speech and Hearing Disorders, 36*, 335-340.

Leh, J. M., Jiterdra, A. K., Caskie, G. I. L., & Griffin, C. C. (2007). An evaluation of curriculum-based measurement of mathematics word problem-solving measures for monitoring third-grade students' mathematics competence. *Assessment for Effective Intervention, 32*, 90-99.

Liu, M., & Holland, P.W. (2008). Exploring population sensitivity of linking functions across three law school admissions test administrations. *Applied Psychological Measurement, 32*, 27-44.

Lord, F.M. (1980). Chapter 13: Equating. In, *Application of item response theory to practical testing problems*. Hillsdale, N.J.: Lawrence Erlbaum.

Loyd, B. H., & Hoover, H. D. (1980). Vertical equating using the Rasch model. *Journal of Educational Measurement, 17*, 179-193.

MacGinitie, W. H., Kamons, J., Kowa;ski, R. L., MacGinitie, R. K., & McKay, T. (1978). *Gates-MacGinitie reading tests* (2nd ed.). Chicago: Riverside.

Macmillan/McGraw-Hill (1989). *Comprehensive Test of Basic Skills* (4th ed.). Monterey, CA: Macmillan/McGraw-Hill School Publishing Co.

Madden, R., Gardner, E. F., Rudman, H. C., Karlsen, B., & Merwin, J. C. (1978). *Stanford Achievement Test*. New York: Harcourt Brace Jovanovich.

Madelaine, A., & Wheldall, K. (1999). Curriculum-based measurement of reading: A critical review. *International Journal of Disability, Development, and Education, 46*, 71-85.

Marston, D. (1989). A curriculum-based measurement approach to assessing academic performance: What it is and why do it. In M. Shinn (Ed.), *Curriculum-based measurement: Assessing special children* (pp. 18-78). New York: Guildford.

Marston, D., & Deno, S. (1981). The reliability of simple, direct measures of written expression (Vol. IRLD-RR-50). University of Minnesota, Institute for Research on Learning Disabilities.

Meherns, W. A., & Clarizio, H. F. (1993). Curriculum-based measurement: Conceptual and psychometric considerations. *Psychology in the Schools, 30*, 241-254.

Montague, M. (2003). *Solve It! A practical approach to teaching mathematical problem solving skills.* VA: Exceptional Innovations, Inc.

Montague, M., Penfield, R. D., Enders, C., & Huang, J. (2010). Curriculum-based measurement of math problem solving: A methodology and rationale for establishing equivalence of scores. *Journal of School Psychology*, *48*(1), 39-52.

Musial, D., Nieminen, G.,Thomas, J., & Burke, K. (2009). *Foundations of Meaningful Educational Assessment.* Monterey, CA: McGraw-Hill Higher Education.

National Council of Teachers of Mathematics (1989). *Curriculum and evaluation NCTM Standards for school mathematics.* Reston, VA: The National Council of Teachers of Mathematics, Inc.

National Council of Teachers of Mathematics (2000). *Principles and NCTM Standards for school mathematics.* Reston, VA: The National Council of Teachers of Mathematics, Inc.

No Child Left Behind Act. Reauthorization of the Elementary and Secondary Education Act. Pub. L. 107-110 2102(4) (2001).

Okamoto, Y., Case, R. (1996). Exploring the microstructure of children's central conceptual structures in the domain of number. *Monographs of the Society for Research in Child Development, 61*, 27-59.

Osterlind, S. J. (2006). *Modern measurement: Theory, principles, and applications of mental appraisal*. New Jersey: Prentice Hall, Inc.

Parker, R. I., Tindal, G., Hasbrouck, J. (1991a). Countable indices of writing quality: Their suitability for screening-eligibility decisions. *Exceptionality, 2*, 1-17.

Parker, R. I., Tindal, G., Hasbrouck, J. (1991b). Progress monitoring with objective measures of writing performance for students with mild disabilities. *Exceptional Children, 58*, 61-73.

Phopham, W. J. (2000). *Modern educational measurement* (3rd edition). Boston, MA: Allyn & Bacon, Inc.

Prescott, G. A., Balow, I. H., Hogan, T. P., & Farr, R. C. (1984). Metropolitan Achievement Tests (MAT-6). San Antonio, TX: The Psychologist Corporation.

Rasch, G. (1980). *Probabilitic models for some intelligence and attainment tests* (Expanded ed.). Chicago: University of Chicago Press.

Reynolds, C. R., Livingston, R. B., & Willson V. (2009). *Measurement and assessment in education*. New Jersey: Perason Education, Inc.

Shapiro, E. S., Edwards, L., & Zigmond, N. (2005). Progress monitoring of mathematics among students with learning disabilities. *Assessment for Effective Intervention, 30*, 15-32.

Shinn, M.R., Good, R. H., Knutson, N., Tilly, W. D., & Collins, V. L. (1992). Curriculum-based measurement of oral reading fluency: A confirmatory analysis of its relation to reading. *School Psychology Review*, 21, 459-479.

Speece, D. L., Case, L. P., & Molloy, D. E. (2003). The validity of a response-to-instruction paradigm to identify reading disabilities: A longitudinal analysis of individual differences and contextual factors. *School Psychology Review, 32*, 557-582.

Stecker, P.M., Fuchs, L.S., & Fuchs, D. (2005). Using curriculum-based measurement to improve student achievement: Review of research. *Psychology in the Schools, 42,* 795-819.

Stecker, P. M., Lembke, E. S., & Foegen, A. (2008). Using progress monitoring data to improve instructional decision making. *Preventing School Failure 52*(2), 48-58.

Tindal, G., & Parker, R. (1989a). Assessment of written expression for students in compensatory and special education programs. *The Journal of Special Education, 23*, 169-183.

Tindal, G., & Parker, R. (1989b). Development of written retell as a curriculum-based measure in secondary programs. *School Psychology Review, 13*, 328-343.

Traub, R.E., & Rowley, G. L. (1991). Understanding reliability. *Educational Measurement: Issues and Practice, 10*(1), 37-45.

VanDerHeyden, A. M., Broussard, C., Fabre, M., Stanley. J., LeGendre, J., & Creppell, R. (2004). Development and validation of curriculum-based measures of math performance for preschool children. *Journal of Early Intervention, 27*, 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*, 363-382.

Videen, J., Marston, D., & Deno, S. L. (1982). Correct Word Sequences: A Valid Indicator of Proficiency in Written Expression (Vol. IRLD-RR-84, pp. 61). Minnesota Univ, Minneapolis Inst for Research on Learning Disabilities.

von Davier, A.A., & Wilson, C. (2008). Investigating the population sensitivity assumption of item response theory true-score equating across two subgroups of examinees and two test formats. *Applied Psychological Measurement, 32*(1), 11-26.

Wallace, T., Espin, C. A., McMaster, K., Deno, S. L., & Foegen, A. (2007). CBM progress monitoring within a standards-based system. *The Journal of Special Education, 41*(2), 66-67.

Wayman, M. M., Wallace, T., Wiley, H. I., Ticha, R., & Espin, C. A. (2007). Literature synthesis on curriculum-based measurement in reading. *The Journal of Special Education, 41*, 85-120.

Weissenburger, J. W., & Espin, C. A. (2005). Curriculum-based measures of writing across grade levels. *Journal of School Psychology, 43*, 153-169.

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

Wright, B. D., & Tennant, A. (1996). Sample size again. *Rasch Measurement Transactions*, *9:4,* 468.

Yang, W.L. (2004). Sensitivity of linkings between AP multiple-choice scores and composite scores to geographical region: An illustration of checking for population invariance. *Journal of Educational Measurement, 41*, 33-41.

Yang, W.L., & Gao, R. (2008). Invariance of score linkings across gender groups for forms of a testlet-based college-level examination program examination. *Applied Psychological Measurement, 32*(1), 45-61.

Yi, Q., Harris, D.J., & Gao, X. (2008). Invariance of equating functions across different subgroups of examinees taking a science achievement test. *Applied Psychological Measurement, 32*(1), 62-80.

Appendix A

District Eligibility Criteria for Placement in the Learning Disabilities Program

(a) a disorder in one or more of the basic psychological processes including visual, auditory, or language processes,

(b) academic achievement significantly below the student's level of intellectual functioning,

(c) learning problems that are not due primarily to other disabling conditions, and

(d) ineffectiveness of general educational alternatives in meeting the student's educational needs.