THE IMPORTANCE OF TEST VALIDITY: AN EXAMINATION OF MEASUREMENT INVARIANCE ACROSS SUBGROUPS ON A READING TEST

by

Anita Michelle Wilson Rawls

Bachelor of Science
University of South Carolina, 2000

Master of Education
University of South Carolina, 2005

Submitted in Partial Fulfillment of the Requirements

For the Degree of Doctor of Philosophy in

Educational Psychology and Research

College of Education

University of South Carolina

2009

Accepted by:

Huynh Huynh, Major Professor

Tammiee Dickenson, Committee Member

Christine DiStefano, Committee Member

Diane Monrad, Committee Member

James Buggy, Dean of The Graduate School

Dedication

To my maternal grandparents, Patrick Frazier, Sr. and Cora Lee Cohen Frazier; my paternal grandmother, Ida Moultrie; my aunt, Wilhelmina Cohen; and uncle, Julius Moultrie.

Acknowledgements

I give my most sincere gratitude to my husband, parents, sisters, extended family, and friends. I am especially grateful to my husband, Charles Wesley Rawls, III, for his intriguing conversations and unyielding determination to support my quest for more knowledge and skills. It is a reward to have parents, Andrew Wilson and Susan Patricia Frazier Wilson, who are supportive and provide inspiration to balance the mind, body, and soul. Thank you so much. I am appreciative for the encouragement from my sisters, Irma Coraine Wilson Gomillion and Gratiana Anitra Wilson Taylor, and their families. I would also like to thank Marquita Lynn Adger and Sabrina Brown for demonstrating the value of friendship. I love you.

I am also grateful to Johnnie McFadden, Rhittie Gettone, Marva Larrabee and the scholars of the African-American Professors Program for their constant support in many ways to ensure my success. I am honored to be the last doctoral advisee of the award-winning, Huynh Huynh. I am appreciative for the knowledge he has shared with me, for his honesty and the way in which he has mastered the art of providing gentle, yet firm guidance to his students. I would like to thank Tammiee Dickenson, Christine DiStefano, and Diane Monrad for sharing their expertise in their respective areas. I extend my gratitude to Robert Johnson for expressing his confidence in my success. It is only with the foundation set by my family, friends, professors and students in my department, professors and students in my college and professors and students across the university that I was able to attain this goal. Thank you.

Abstract

The study discussed the importance of test validity, often established when making decisions that may affect a student's future. The decisions made by policymakers and educators must not adversely affect any particular subgroups of students (i.e., year of administration, gender, ethnicity, level English proficiency, socioeconomic status, and disability status). The study discussed the testing of measurement invariance across subgroups on an assessment as a process of validation. Methods used to detect measurement invariance at the test, subtest, and item levels were reviewed and three of these methods were applied to a reading test for administrative, gender, and ethnic subgroups. The purpose of this study was to demonstrate how to detect measurement invariance using 1) hierarchical linear modeling at the test level, typically used by policymakers, 2) confirmatory factor analysis at the subtest level for instructional designers, and 3) Rasch item analysis at the item level for psychometricians. The results of the study provided validity evidence that supported the comparison across administration years at the test, subtest, and item levels. Validity evidence also supported the comparison of gender subgroups at the subtest level via partial scalar invariance and at the item level. Finally, the results provided evidence that supported the comparison of ethnic groups at the subtest level via partial scalar invariance.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

Testing has an important role in the society in which we live. Various industries use testing to find people who possess a certain characteristic. The employment sector uses testing to select the most qualified candidates (Mikulay & Goffin, 1998; Sackett, Borneman, & Connelly, 2008; Scroggins, Thomas, & Morris, 2008a, 2008b). Psychologists use testing to understand people and their personality (Kubiszyn, et al., 2000; Meyer, et al., 2001). In the education sector, testing assists in the evaluation of the performance of students or the identification of a students' level of intelligence (Cizek, Rosenberg, & Koons, 2008). The results of these tests lead to important decisions including job placement, medical treatment, promotion, and/or retention. The importance placed on these decisions support the need for evidence of the validity or accuracy of the interpretations and uses of the testing data. The purpose of the validity evidence is to ensure the interpretation and uses of the data are accurate. In particular, it ensures an offer is extended to the most qualified candidate for the job, an accurate diagnosis is given, and an effective treatment plan is developed. This evidence also ascertains the academic performance of students in our nation and provides for an accurate prediction of our future success in the global market.

Since 2001, stricter sanctions have been placed on schools, districts, and states in an effort to hold them accountable for their role in the academic development and performance of students (No Child Left Behind Act, 2002). These recent educational accountability efforts emphasize the need for schools to design instructional targets to meet the needs of all students. Measurement of the schools' success at meeting the needs of all students is established by the performance of subgroups on state assessments. These subgroups are defined by race/ethnicity, socioeconomic status, English proficiency status, and disability status (National Forum on Education Statistics Race/Ethnicity Data Implementation Task Force, 2008). This is slightly different than previous years during which the school performance was evaluated based on the average student performance in the school. "The No Child Left Behind Act (2002) obliges states, school systems, and schools to take steps necessary to ensure that 100% of students in grades three to eight and high school achieve proficient performance on state assessments no later than the 2013 – 2014 school year" (Ferrara & DeMauro, 2006, p. 579), NCLB specifies that each school is to report data disaggregated by subgroup to determine if targets are met (Koretz & Hamilton, 2006). Those groups in a school that meet the minimum group size set by the state for participation are required to report their achievement levels, thus providing evidence of a valid generalization of the proficiency level of students from that subgroup. In an effort to meet the accountability demands, many states are implementing more diagnostic or benchmark testing to foster remedial teaching in the areas of need prior to the summative assessment. Since the inception of NCLB, diagnostic and/or benchmark testing has been coupled with numerous scientifically-based instructional programs designed and implemented to improve the achievement level of students across the nation. One such program at the federal level is the Reading First Initiative, a scientifically-based reading research program. Evidence of validity in the Reading First

Initiative is important due to the instructional, placement, funding, and/or remedial teaching decisions made by educators, policymakers, and other stakeholders.

The Reading First Initiative

The Reading First Initiative is a federally funded reading project designed to provide states with the assistance needed to establish a scientifically-based reading research instructional program for children in primary grades. The Initiative particularly targets low-income, low-performing schools whose districts and states submit their ideas for a scientifically-based reading instructional program through a competitive process. Successful applicants are provided with federal funding, then award subgrants to districts and/or schools. The funds received from the United States Department of Education are allocated to support professional development, the acquisition of instructional resources, and the use of diagnostic and progress monitoring tools. There are 49 states participating in the Reading First Initiative and each state has been given the freedom to develop their own implementation and evaluation plan. Given the variety of plans that may be implemented, it is important to establish the validity of the inferences made from the results presented from each state. One way to establish this evidence is by demonstrating the construct being measured (i.e., reading) is invariant across the reporting subgroups, such as gender and ethnicity. Participating southeastern states include: Alabama, Arkansas, Florida, Kentucky, Louisiana, Mississippi, North Carolina, South Carolina, Tennessee, Virginia, and West Virginia.

The purpose of the Reading First Initiative is to ensure children across the nation are reading at or above grade level by the end of third grade. The acquisition of reading skills and abilities in students is being noted through the use of assessments. Measures of early reading performance capture the wide variation of children's knowledge and skills as they enter school and are the primary determinants of later reading performance (Butler, Marsh, Sheppard, & Sheppard, 1985). Background factors such as age, gender, ethnicity, and family risk characteristics may influence the relationship when measuring early reading performance of students (Chatterji, 2006). Thus, the need for validity evidence across groups is essential. In particular, validity evidence to support the interpretations and score uses across years, or administration date, gender, and ethnicity is necessary. Validity evidence across administrative dates is particularly important because policymakers draw conclusions about the effectiveness of the Reading First Initiative across years to examine the impact of the program (Gamse, Bloom, Kemple, Jacob, & Institute of Education Sciences, 2008). Validity evidence across gender, ethnicity, socioeconomic status, English proficiency status and disability status is also important because policymakers draw conclusions about the effectiveness of the Reading First Initiative and many other programs across the groups in an effort to close the achievement gap (Downey, Steffy, Poston, & English, 2009).

Historical performance of subgroups in reading assessments

Using various data sets, researchers have consistently found gaps in reading achievement based on students' gender/sex, race/ethnicity, English proficiency status, socioeconomic status, and disability status (Foster & Miller, 2007; Montgomery & Hayes, 2005; Risko & Walker-Dalhouse, 2007; Votruba-Drzal, 2006; Yeung & Conley, 2008). The results of some reading assessments such as the National Assessment of Educational Progress, the Programme for International Student Assessment, the Early Childhood Longitudinal Study, and the Progress in International Reading Literacy Study show the following subgroups outperformed their counterparts: White, unsubsidized lunch, and females (Chatterji, 2006; International Reading Association, 2001; Topping, 2006). The demands of NCLB Act also requires students with disabilities to perform at the level of proficiency. Historically, this subgroup has always performed less efficient than students without disabilities (Katsiyannis, Zhang, Ryan, & Jones, 2007; Montgomery & Hayes, 2005). From previous research in differential item functioning, “subgroups are not expected to perform equally well [because they] differ in their experiences, interests, and motivations. Consequently, only groups formed by random assignment should be expected to perform equally well” (Drasgow, Luecht, & Bennett, 2006, p. 499). If data are disaggregated by subgroups to draw conclusions, random assignment no longer exists. Thus, it is important to validate the inferences and actions from the disaggregated data. Sound psychometric practice encourages the inferences made and actions taken as a result of the students’ performance to be valid (Kane, 2006). One aspect of this validation process includes producing evidence of construct validity through the use of measurement invariance across the subgroups of interest. Measurement invariance is a form of construct validity evidence that determines whether or not the construct being measured is perceived in the same manner regardless of group membership (Mellenbergh, 1989). Evidence of this invariance supports the comparisons of students across subgroups to inform policy and practice.

Purpose and educational significance of study

Policymakers and educators have to make decisions that do not adversely affect any particular gender/sex, race/ethnicity, level of English proficiency status, socioeconomic status, and disability status. This study examines the construct validity of the inferences from a reading test using three fundamentally different approaches: hierarchical linear modeling at the test level, confirmatory factor analysis at the subtest level, and item response theory via the Rasch model at the item level. Inferences are made from the reading test that compare the student performance across administrations, gender, and ethnicity. Each of the methods of analysis are capable of producing validity evidence that supports or refutes comparisons across administration date, gender, and ethnicity subgroups.

The hierarchical linear modeling method (Byrk & Raudenbush, 1992) allows for the comparison of the subgroups across the entire examination. This method is particularly interesting because it also simultaneously investigates the data at the item, person, and group levels. The model estimates item coefficients at level 1, person coefficients at level 2, and the subgroup coefficients at level 3. This methodology specifically isolates the variability within and between the groups, thus improving interpretation for decision-making by policymakers.

The confirmatory factor analysis method (Jöreskog, 1969), as presented in this study, is instrumental in establishing evidence of construct validity through the use of measurement invariance at the subtest level. Construct invariance at the subtest level will allow for the interpretation of various components of reading acquisition (i.e., phonics/phonemic awareness, listening vocabulary, vocabulary development, and reading comprehension) across subgroups. This process also determines whether a particular administrative, gender, or ethnic group performs significantly different from their counterparts on a specific subtest. Measurement invariance at the subtest level will provide evidence for the diverse components on which the subgroups may be compared, typically needed for instructional decisions.

The final method, Rasch modeling (Rasch, 1960/1980) within the context of item response theory, is influential in establishing evidence of stability at the item level. The results of this analysis provides person and item parameter estimates for each subgroup. This study specifically examines the stability of the item parameter estimates and fit statistics across subgroups. If the estimates are stable across subgroups, one will be able to conclude the item characteristics are stable across groups. This analysis allows for the interpretation and use of data at the item level, typically used by psychometricians in test development.

The concurrent use of these three methodologies demonstrates how to establish construct invariance at the test, subtest, and item levels as justification for comparisons made across groups in similar contexts. The purpose of this study is to demonstrate how to establish validity evidence through the identification of measurement invariance across subgroups using three methods: hierarchical linear modeling at the test level, confirmatory factor analysis at the subtest level, and item response theory via the Rasch model at the item level. Specifically, this study is designed to answer the following research questions.

Research questions

1. Are the group level coefficients from the hierarchical linear modeling technique invariant across administrative groups, gender, and ethnicity?

2. Are the group measures produced through confirmatory factor analysis (i.e. factor structure, factor loadings, and error variances) invariant across administrative groups, gender, and ethnicity?
3. Are the item difficulty estimates and fit statistics based on Rasch item response theory modeling stable across administrative groups, gender, and ethnicity?

The results of the study will either support or refute the efforts of policymakers, instructional designers, and psychometricians to make comparisons across subgroups on various dimensions of a test. The hierarchical linear modeling will either support or refute the comparison of subgroups across the entire test, the confirmatory factor analysis will either support or refute the comparison of subgroups across the subtests, and the Rasch item analysis will either support or refute the comparisons of subgroups across the items. Given the reading test is an adaptation from a long-standing, norm-referenced test with strong psychometric properties; one would expect the assessment to display measurement invariance at all levels. However, the population used to establish the norms differed from the students in the study by having a wider range of achievement levels, more minorities and students with lower socioeconomic status (Dickenson, Habing, Rawls, & Johnson, 2008). Thus, it is necessary to produce validity evidence to support the actions and interpretations employed. The results from the three methods provide the empirical evidence needed to suggest the use of varying methodologies by test developers and psychometricians to examine group differences at the test, subtest, and item levels. If the three methods provide distinctly different results, these stakeholders should be cautious about making comparisons across subgroups at the appropriate level.

Chapter 2

Literature Review

As early as third grade, policymakers compare the ability of students across the nation in the decision making process. The rate at which students acquire the knowledge, skills, and ability to read is often documented through the use of early reading achievement measures. As with any measure, evidence of construct validity using measurement invariance is necessary to support the interpretations and uses of the data. This chapter begins with a brief review of research on reading acquisition in young children. The following section describes the importance of validity and how the establishment of measurement invariance contributes to the construct validation process. The next sections describe the range of methods used to provide evidence of measurement invariance, such as multiple group confirmatory factor analysis, item response theory via the Rasch model, and more recently hierarchical linear modeling. The final sections describe these methods and studies using these methods to detect measurement invariance.

Reading acquisition among young children

The theory of reading acquisition, as stated by Thompson and Fletcher-Flinn (1993), defined reading as a cognitive skill in which there were "two classes of procedures for word identification responses: recall and generation. The outcomes of generation along with teacher provided associations are the sources of knowledge for recall" (p. 68). From this theory, it was inferred that reading acquisition was a process through which students were exposed to the alphabet, learn to identify letters, recognize the sounds of the letters in the alphabet, comprehend the words created by the combinations of letters and sounds, and eventually come to know words can be combined to form sentences, paragraphs, and stories. Previous research referred to these of components of the developmental process as logographic, alphabetic, orthographic, phonics, phonological awareness, phonemic awareness, phonological decoding, word recognition, fluency, word meaning, vocabulary, and/or sentence development (Adams, 1990; Byrne, Brian, 1992; Chall, 1996; Dally, 2006; Flippo, 2001; Nation, 2008; National Institute of Child Health and Human Development, 2000; Snow, Burns, & Griffin, 1998; Taylor & Pearson, 2002).

The rate of reading acquisition among young children varied depending on the students' background characteristics upon entering school (Biemiller & Boote, 2006; Catts, Fey, Zhang, & Tomblin, 1999; Connor & Craig, 2006; McCoach, O'Connell, Reis, & Levitt, 2006). McCoach, et al. (2006), described the usual growth trajectory for young readers as increasing from fall to spring of kindergarten, decreasing from spring of kindergarten to fall of first grade, then increasing at a rate slightly faster than that of kindergarten from fall to spring of first grade. Researchers completed studies in which they were able to predict the future reading ability of a child based on their performance in the reading components tested in kindergarten and/or first grade (Biemiller & Boote, 2006; Blachman, et al., 2004; Butler, et al., 1985; Catts, 2001; Cunningham & Stanovich, 1997; Sparks, Patton, Ganschow, Humbach, & Javorsky, 2008). Catts (2001) found letter identification, sentence imitation, phonological awareness, rapid naming, and mother's education to be strong predictors of student reading ability in second grade. Biemiller and

Boote (2006) specifically recommended that all students master word decoding in the first grade or they will be at risk of being labeled a struggling reader in second and/or third grade. The Butler, et al. (1985) longitudinal study examined the ability of a comprehensive battery of kindergarten measures to predict early reading achievement of students in grades 1 to 6. The study identified six predictor factors, beyond IQ, that were able to predict or explain the variation in reading achievement scores. These factors included psycholinguistic abilities, figure drawing, language, rhythm, perceptual motor skills, and spatial/form perception. Sparks, et al. (2008) found that students were able to transfer their word decoding, spelling, and reading comprehension skills acquired in early grades to second language acquisition in high school. Overall, reading acquisition among young children varied based on the students' background characteristics and was a strong indicator of future reading performance. The collection of these findings established the need for a strong foundation in federal reading programs to support the development of these skills in young children. One such initiative was the Reading First Initiative.

The Reading First Initiative

The Reading First Initiative was a scientifically-based reading instructional program designed to provide professional development and progress monitoring tools to ensure students are reading on or above grade level by the end of the third grade. The instructional program included the following five components of reading: phonics, phonemic awareness, fluency, vocabulary, and comprehension. Literacy coaches, teachers, reading intervention specialists, and other educators in a Reading First school and/or district received professional development that helped them focus their instruction on these five components of reading instruction. These five components of reading used in the Reading First Initiative were those highlighted by the National Reading Panel as the strongest predictors of reading ability in children that can be included in a formal instructional program (National Institute of Child Health and Human Development, 2000). Educators also received the necessary professional development to analyze and apply results of the progress monitoring tools. The progress monitoring tools allowed educators the opportunity to use data to guide instruction.

The scientifically-based reading instruction of the Reading First Initiative was designed for low-income or low-performing schools and/or districts. Students identified as those at risk for experiencing reading difficulties due to various characteristics, such as, cognitive ability, low-income family background, being from a racial/ethnic group of color, attending a high poverty school, and attending a school with a high enrollment of students of color (Chall, 1996; Flippo, 2001; Kieffer, 2008; McCoach, et al., 2006; Snow, et al., 1998; Taylor & Pearson, 2002). Regardless of background characteristics or situation, the instruction provided to students in the Initiative was designed to meet their needs.

Gender in early reading achievement

Over the years, data from several international and national sources supported the notion that the overall performance of males and females varied with girls outperforming boys in their early reading achievement. Specifically, the data from the Programme for International Student Assessment (PISA), the Progress in International Reading Literacy Study (PIRLS), and the National Assessment of Educational Progress (NAEP) reported a higher percentage of girls as being proficient in reading or girls outperforming boys in reading (Topping, 2006; US Department of Education Center for Education Statistics, 2006). The Digest of Education Statistics stated girls had a slightly higher average reading scale score than boys since 1971 (US Department of Education Center for Education Statistics, 2006). Similarly, the report from the Early Childhood Longitudinal Study suggested the girls perform better than boys (Chatterji, 2006). These differences in gender performance exist at the international and national levels as well as on state level examinations. The performance of third grade students across some of the southeastern states revealed similar gender differences. The third grade girls in Florida had a mean scale score higher than boys on the 2008 Florida Comprehensive Assessment Test for reading (Florida Department of Education, 2008). In North Carolina, the percentage of girls in third grade performing at or above grade level was greater than the percentage of third grade boys performing at or above grade level (North Carolina Department of Public Instruction, 2008). The South Carolina statewide examination of student achievement in English/Language Arts for the Palmetto Achievement Challenge Tests supported the national results with a higher percentage of third grade girls meeting proficiency than third grade boys (South Carolina Department of Education, 2008). In Virginia, the percentage of third grade girls passing the assessments for competency was higher than that for third grade boys (Virginia Department of Education, 2008). The early reading achievement of females was typically higher than that of males in the sample of international, national, and state studies mentioned above.

Race/ethnicity in reading achievement

Race/ethnicity in the United States had an evolving profile due to immigration and the fertility and mortality within the population (Shrestha, 2006). In 1997, the United States Office of Management and Budget officially recognized five racial/ethnic groups in the United States which included: White, Black/African American, American Indian/Alaska Native, Asian, and Native Hawaiian and Other/Pacific Islander (United States Office of Management and Budget, 1997). Researchers identified the diversity of the nation as steadily increasing, and the differences in the achievement of these groups were of concern to many parents and educators (Darling-Hammond, 1998; McCoach, et al., 2006). With the exception of Asian students, almost all minority students were consistently performing under that of White students (Connor & Craig, 2006; McCoach, et al., 2006). Achievement data has shown ethnic differences for years, especially the Black/White gap (Meece & Kurtz-Costes, 2001; Topping, 2006; US Department of Education Center for Education Statistics, 2006). In 2006, data from the Digest of Education Statistics gave the average scale score in reading for White, Black, and Hispanic students. The scales scores for the Black and Hispanic students were below those of the White students for more than 10 years (US Department of Education Center for Education Statistics, 2006). Similar data existed for the 2008 statewide assessments for the state of Florida, North Carolina, South Carolina, and Virginia (Florida Department of Education, 2008; North Carolina Department of Public Instruction, 2008; South Carolina Department of Education, 2008; Virginia Department of Education, 2008). The differences in achievement of these groups, noted by several assessments on the national and state level suggested the need for validity evidence to support the inferences and actions taken.

Test validity

Within the context of schooling, stakeholders were concerned about the purpose, quality, and quantity of testing; thus the demand for validity evidence increased

(Lederman & Burnstein, 2006; Supon, 2008). These stakeholders, in the form of teachers, parents, students, and businesses were interested in making sure the purpose of the testing was clearly defined. Upon a clear exposition of the objective of an assessment, the validity of the scores was evaluated. Messick (1989) defined validity as "an integrated evaluative judgment of the degree to which empirical evidence and theoretical rationales support the 'adequacy' and 'appropriateness' of 'inferences' and 'actions' based on test scores or other modes of assessment" (p. 13). From this definition of validity, the purpose of the assessment was known prior to the evaluation of the resulting inferences and actions. Another author defined validity as the extent to which the evidence supports or refutes the proposed interpretations or uses (Kane, 2006). This need for a clearly defined purpose of the assessment was also conveyed in the 1999 Standards for Educational and Psychological Testing; which suggested the use of validation to develop scientifically sound evidence to support the proposed interpretation of test scores and their intended use (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1999; Goodwin & Leech, 2003).

Given this understanding of the purpose for test validity, researchers began to disseminate their knowledge about the process by which validity evidence may be established. Kane (2006) described validation as the process of evaluating the credibility of interpretations and uses. Several faces of validity that allowed the researchers to establish evidence of the appropriateness of the inferences and actions of a test. These types of validity evidence were categorized by several authors as; criterion validity, content validity, and construct validity (Angoff, 1988; Cureton, 1951; Kane, 2006; Messick, 1989; Pellegrino, 1988). During the beginning of the twentieth century, criterion validity was defined by Cureton (1951) as the correlation between the actual test scores and the 'true' criterion score and was considered the gold standard. The concept of criterion validity was recently divided into two schools of thought; concurrent validity and predictive validity. Two tests given at the same time with a high correlation between their scores can be thought of as having concurrent validity; while predictive validity involved the ability of the test scores to predict future performance (Kane, 2006). The interpretation of validity as defined by Cureton (1951) was later extended to include content validity; which was used to validate academic measures. The idea behind content validity was to provide evidence the content of the measure was relevant and appropriate for the inferences from and uses of the test score (Messick, 1989). The final extension of the concept of validity, specifically construct validity, was used to validate measures of a psychological nature or theoretical attributes. Cronbach and Meehl (1955) described construct validation as the process to follow when criterion and content measures are unavailable. As time progressed in the field of validity studies, the construct validity approach was widely accepted as a general model for validation of a measure (Anastasi, 1986; Embretson, 1983; Guion, 1977; Messick, 1980, 1988, 1989).

Applications of the construct validity model required researchers to clearly define the interpretation and use of the test scores. In the case of NCLB, the federal government required the student scores in grades 3 – 8 and high school to be aggregated to the subgroup level on the statewide assessments (No Child Left Behind Act, 2002). Knowing this level of aggregation was used to make inferences about the type of education students are receiving and whether or not all students received tutoring services, it was imperative to validate the inferences using subgroup level data. In particular, it was important to provide evidence of subgroups being measured on the same latent trait. Identification of the same latent trait for all subgroups, such as gender and ethnicity, was identified by researchers as validity evidence in the form of measurement invariance (Mellenbergh, 1989).

Measurement invariance

Given the historical performance of students in reading achievement by gender and race/ethnicity; sound psychometric practice suggested researchers ensure the construct validity of the test scores used to draw such conclusions. In the context of construct validity, these subgroup differences were examined to determine if the construct was perceived by both categories of gender and all categories of racial/ethnic groups in the same manner. Vandenberg and Lance (2000, p. 4) defined measurement as "the systematic assignment of numbers on variables to represent characteristics of persons, objects, or events." When comparing relevant groups, equivalent measurements were obtained when the relationship between observed scores and latent constructs was identical across relevant groups (Drasgow & Kanfer, 1985). A more formal definition of equivalent measurement or measurement invariance (MI) by Mellenbergh (1989) was $f(Y|\eta, s) = f(Y|\eta)$ where Y represented the observed score, η represented the factor score, and *s* represented the group membership. This equation suggested a person's observed score was the same if the given information included the factor score and the group membership or just the factor score. The person's group membership had no bearing on the observed score; it was simply a function of the factor score. Generally speaking, measurement invariance suggested some properties of a measure were the same regardless of a person's group membership (Millsap, 2007). The general term measurement invariance was applied to various components of measurement models by Little (1997). Specifically, Little (1997) identified these aspects of the measurement model as category 1 invariance and category 2 invariance. Category 1 invariance referred to the psychometric properties such as configural, metric, scalar, and measurement error invariance (Buss & Royce, 1975; Meredith, 1993; Mullen, 1995; Singh, 1995; Steenkamp & Baumgartner, 1998; Suzuki & Rancer, 1994; Vandenberg & Lance, 2000). Category 2 invariance had to do with the differences in the basic group statistics such as mean, variance, and covariance.

Measurement invariance was displayed across several dimensions of a measure, specifically, a single item, a subscale or set of items, and an entire measure (Crocker & Algina, 1986). Regardless of the dimension of measurement invariance being tested; the relationship with a latent variable was expected to be the same within that dimension across groups (Embretson & Reise, 2000). When an item or a set of items were deemed invariant across groups, validity evidence to support the comparison of groups using raw scores was produced. Yoo (2002) noted an examination of group differences was not warranted until it had been established that the measure was invariant across the groups. As a consequence, mean differences between groups were reflected as true differences between the groups and were not attributed to a measurement artifact (Embretson & Reise, 2000). Without measurement invariance, the differences in observed group means, were possibly due to true differences in the way the group perceives a latent construct, the item content, or presentation (Little, 1997; Steenkamp & Baumgartner, 1998). Researchers referred to the differences due to the item content or presentation as measurement bias that led to invalid interpretations of group differences (Ackerman, 1992; Little, 1997; Mellenbergh, 1989). Mellenbergh (1989) defined measurement bias as the possibility that individuals of equal ability on the latent variable from different groups, do not have identical probabilities of observed scores. Little (1997) referred to this as category 2 invariance, which had to do with between-group differences in latent means, variances, and covariances. The elimination of measurement bias was important because measurement bias threatened the validity of interpretation and uses of educational measures (Ackerman, 1992). Other types of measurement bias described by Drasgow (1982, 1987) included external bias and internal bias. External bias was defined as the existence of different test score correlations with nontest variables for two or more groups of examinees (i.e., predictive validity). Internal bias occurred when a test's internal relations (i.e., the covariances among item responses, similar to category 2) differed across two or more groups of examinees. Embretson and Reise (2000) suggested measurement bias, as described here, can lead to a measurement scale not being invariant or equivalent across groups. However, the labeling of measurement bias required caution. Drasgow and Kanfer (1985) cautioned that measurement invariance can be established and there is no guarantee the distributions of individual scores will be equal across groups. This occurrence may be labeled as measurement bias, but people from one group may simply have higher or lower scores than people from other groups. These between group differences were referred to as impact by Ackerman (1992) and Angoff (1993) and do not effect the validity of the group comparisons. If all the items in a test were measuring only the intended constructs, the observed group differences are true differences in the skill being assessed (Ackerman, 1992; Angoff, 1993; Millsap & Everson, 1993). Impact represented score differences caused by true differences in the target ability. Since the groups differed in the attributes measured by the test, observing group differences was unavoidable (Ackerman, 1992; Angoff, 1993; Millsap & Everson, 1993). If a test lacked construct-related evidence of validity, it meant that the test contained items that were measuring constructs other than those intended to be measured, indicating there was a potential for bias against or for a certain group of examinees (Atar, 2007).

Previous literature revealed the concept of measurement invariance was also similar to differential item functioning (DIF). The term measurement invariance was most often used by validity theorists (Cronbach & Meehl, 1955; Embretson, 1983; Guion, 1980; Vandenberg & Lance, 2000); while DIF was the term most popular among item response theorists (Dorans & Holland, 1993; Drasgow, 1982; Mislevy, 1983; Thissen, Steinberg, & Gerrard, 1986; Woods, 2008). The previously mentioned validity and item response theorists described the detection of measurement invariance and DIF as two fundamentally different processes.

Testing for measurement invariance

As the concept of validity and the process of validation became more prevalent in social science research, so had the interest in testing for measurement invariance (Bowden, et al., 2008; Byrne, Barbara, Baron, & Balev, 1996; Carle, Millsap, & Cole, 2008; Cheung & Rensvold, 1999, 2002). Testing for measurement invariance was necessary to make valid inferences about differences among groups. The examination of measurement invariance tested the hypothesis that the set of latent variables derived from a set of observed variables was the same for persons from different groups; and the numerical relationships between observed scores and the corresponding latent variables were the same (Bowden, et al., 2008). After establishing the hypothesis, the next step in the process of testing for measurement invariance was to evaluate whether the same general factor structure of the measure (i.e., configural invariance) was followed by other psychometric properties, such as metric invariance, and scalar invariance in both groups (Campbell, Barry, Joe, & Finney, 2008). If measurement invariance was established for the instruments, or the hypothesis was true, evidence of the validity of the scale score uses and interpretations was provided and researchers placed confidence in their group comparisons (Hong, Malik, & Lee, 2003). If measurement invariance was not evident, the accuracy of group comparisons and the validity of interpretations and uses of the data were questioned (Hong, et al., 2003; Little, 1997; Yoo, 2002).

Methods used to detect measurement invariance

Several methods used to detect measurement invariance, or group differences, included analysis of variance (ANOVA), multiple regression, odds ratios, and log-linear models. One of the methods used to detect group differences was multiple regression. Multiple regression was a useful method for detecting group differences between observed variables. Typically, a raw test score was created for an individual by summing up the item scores and the overall test score was predicted using several observed variables (Bowey, 1995; Rock, Werts, & Flaugher, 1978). This approach worked well when one guaranteed the predictors contained no measurement error. The lack of measurement error in the predictors was a necessary assumption for multiple regression, however, it was not always valid when the predictors were observed variables. Rock, et al (1978) noted the consequences of allowing the error variances of the predictors to vary may lead to biased multiple regression estimates. Stone-Romero, Alliger, and Aguinis

(1994) also noted through a simulation study that statistical power increased as the group sample size increased and the difference between the within-group correlation coefficients increased when using multiple regression. This was a benefit, but it was heavily dependent on the data. Given the heavy dependence on the quality of the data (i.e., predictors contain no measurement error) in multiple regression, there was some concern about it being the best method for detecting measurement invariance.

Despite the assumption about the measurement error in the predictors, Kamata (2001) extended the multiple regression methodology to account for the hierarchical nature of the data and proposed the use of a multilevel item analysis model or hierarchical generalized linear model (HGLM; herein referred to as HLM) to detect group differences. Through a simulation study, Kamata (2001) compared the estimation of coefficients to show a three-level latent regression model (one-step approach) with group effects provided estimates less affected by the number of items on a test than an alternative two-step approach. The three-level latent regression model was presented as mathematically equivalent to the Rasch item response theory model. However, the purpose of the model was not to estimate Rasch parameters (Kamata, 2001). The model sought to examine group differences and the study gave examples on how the model was used with and without predictors at the individual and group levels.

Other methods for examining group differences, similar in application to multiple regression, but less traditional were odds ratios and log-linear models. Odds ratios and log linear models allowed one to calculate the probability or odds of a person from a particular group exhibiting a particular characteristic and were especially useful when the sample was small. The odds ratio, commonly cited in the DIF literature, was used to detect the size of the DIF and was combined with a significance test to create the Mantel-Haenszel statistics (Fidalgo, Hashimoto, Bartram, & Muniz, 2007; Holland & Thayer, 1988; Innabi & Dodeen, 2006; Mantel & Haenszel, 1959; Zwick, Donoghue, & Grima, 1993). Despite its popularity, Kristjansson, Aylesworth, and McDowell (2005), suggested the Mantel-Haenszel statistic was not the best method for detecting differences when nonuniform DIF is present. Nonuniform DIF occured when there was an interaction between the ability level and group, specifically, when group A in the low ability level performed better than group B, but in the high ability level group B performed better than group A. The log-linear models were also used to examine group differences. Dancer, Anderson, and Derlin (1994) suggested log-linear models were similar to multiple regression and analysis of variance, which made them relatively easy for traditional researchers to incorporate in their studies. In particular, log-linear models were used in studies to examine the effect of a combination of variables on item responses. However, the log-linear model was most advantageous in determining the interactions among a variety of categorical outcome variables and was not considered useful for ordinal and interval outcome variables (Dancer, et al., 1994).

Methods for detecting measurement invariance, discussed by Millsap (1995, 1997), were those that allowed a structural relationship between the factors and test scores. These methods were recommended in contrast to the multiple regression, odds ratios, and log-linear models because of their ability to eliminate the bias associated with using predictors that were free of measurement error (Lubke, Dolan, Kelderman, & Mellenbergh, 2003). Given the predictors were free of error, the concept of prediction bias was contradictory. If the predictors were invariant, one would expect an equation using them or their outcome to be invariant. Millsap (1995) suggested solving the invariant predictor problem by modeling the structural relations between the latent variables and test scores instead of using the observed variables. These structural relations between factors incorporated methods for testing measurement invariance such as confirmatory factor analysis (CFA) and item response theory (IRT) (Embretson & Reise, 2000; Waller, Thompson, & Wenk, 2000).

Confirmatory factor analysis

Researchers suggested a prerequisite for the study of factorial invariance was the equal dimensionality of the measures in the groups compared; thus, multi-group CFA was used only if equal dimensionality was indicated (Jöreskog, 1971; Vandenberg & Lance, 2000; Vassend & Skrondal, 1999). Confirmatory factor analysis, defined mathematically as $Y = \Lambda\xi + \theta$; where Y was the outcome variable or test score, Λ was the vector of observed item responses; ξ was the latent variable being measured and θ represents a vector of error variances associated with each item response. Campbell, et al. (2008) suggested the exploration of the concept of measurement invariance using CFA by the terms configural, metric, and scalar invariance. Configural invariance tested the overall structure of two or more groups. It sought to determine whether the each group has the same number of factors and whether the items were loading on the same factor across groups (Campbell, et al., 2008). Only if configural invariance was established can metric invariance exist. Metric invariance tested the extent to which the relationships between the factors and the items were equivalent across the two groups. Similarly, if metric invariance was supported, it may be concluded that the groups were interpreting the items in the same way (Byrne, Barbara, 1998). A lack of metric invariance may imply that some items were more important for one population than for the other or that some items were more ambiguous for one group than for another (Campbell, et al., 2008; Chan, 2000). Metric invariance, sometimes referred to as generalizability, was conceived basically as a matter of factorial invariance (Vassend & Skrondal, 1999). The third step of measurement invariance testing, scalar invariance, tested the equality of error variances or intercept terms. At this step, it may be determined whether both groups used the response scale in a similar way (Hong, et al., 2003; Steenkamp & Baumgartner, 1998; Vandenberg & Lance, 2000). The need for scalar invariance provided the evidence necessary to interpret mean differences between the groups, if mean level group comparisons were the objective of the study (Meredith, 1993; Steenkamp & Baumgartner, 1998). Unless scalar invariance was supported, the validity of inferences pertaining to group comparisons was deemed questionable by Campbell, et al. (2008). In the measurement literature, configural invariance was referred to as the least restrictive or congeneric model; metric invariance was referred to as essentially tau-equivalent model in which errors may be different across measures; and scalar invariance was referred to as the parallel model (Graham, 2006; Raykov, 1997). For the context of this study, the terms configural, metric, and scalar were used.

Rasch item analysis

Rasch item analysis, a form of item response theory (IRT) used to detect group differences on measures, was based on the idea that the probability of a person getting an item correct or obtaining a particular score on an item given their ability can be modeled (Lord, 1980; Rasch, 1960/1980). IRT was based on the logit model and simultaneously estimated person abilities and item parameters on the same scale. Crocker and Algina

(1986) described three IRT models that related the probability of a correct response to an item to an examinee's ability. All of the models estimated person ability. In addition, the Rasch model estimated the item difficulty (Rasch, 1960/1980). In this model, all items had the same discrimination parameter, the parameter used to distinguish those high performing examinees from the lower performing examinees. The equation for the one-parameter item characteristic curve (ICC) defined by (Rasch, 1960/1980) is

$P_g(\theta) = \frac{e^{Da(\theta - b_g)}}{1 + e^{Da(\theta - b_g)}}$, where *D* was a constant which was typically set to 1.0 or 1.7, *a* was the index of discrimination which was assumed to be constant across groups and b_g was the item difficulty for a particular group.

The two-parameter model recognized the changes of the item difficulty and the item discrimination when measuring the examinee's performance or estimating a person's ability. This model increased in value as items became more difficult and it was not dependent on the ability level. The equation for the two-parameter ICC was defined by Lord (1980) as $P_g(\theta) = \frac{e^{Da_g(\theta - b_g)}}{1 + e^{Da_g(\theta - b_g)}}$, where a_g was the discrimination index for a particular group. The three-parameter model recognized the changes in item difficulty, the changes in item discrimination, and considered the possibility of students' guessing the correct response through the use of a guessing parameter. This model specifically recognized the strength in the relationship between ability and item response. The equation for the three-parameter ICC defined by Lord (1980) as

$P_g(\theta) = c_g + \frac{(1 - c_g)e^{Da_g(\theta - b_g)}}{1 + e^{Da_g(\theta - b_g)}}$ where c_g was the pseudo guessing parameter for each group. DeMars (2001) suggested the 3PL model is best, but for some group comparisons, the Rasch model will suffice. Due to its simplicity and popularity in achievement tests (i.e., Arkansas, Maryland, and South Carolina) the Rasch model was examined in this study along with evidence of model fit (Huynh & Rawls, 2009).

In the item response theory literature, the existence of group differences on an item were sometimes referred to as DIF (Embretson & Reise, 2000). In this literature, DIF was said to occur when a test item did not have the same relationship to a latent variable (or a multidimensional latent vector) across two or more examinee groups or an item parameter differs across groups (Dorans & Holland, 1993; Embretson & Reise, 2000; van der Linden & Hambleton, 1997). In many DIF studies the item parameters for a reference group and a focal group were compared. This level of analysis allowed some items to be invariant while other items from the same measure were distinctly different for the reference and focal groups. A test of this type was referred to as being partially invariant across the groups (Embretson & Reise, 2000). An item is said to be biased if the probability of getting a correct answer for an item was not the same for different groups of examinees with the same ability level; the cause of which was described by Ackerman (1992) as nuisance abilities. There were several methods used to detect DIF that were not based on item response theory such as the Mantel-Haenszel statistic and standardization (Dorans & Holland, 1993; Holland & Thayer, 1988; Mantel & Haenszel, 1959) or logistic regression (Swaminathan & Rogers, 1990). In IRT based methods used to detect DIF, or measurement invariance, Embretson and Riese (2000) compared the value of the item parameters for the reference group and the focal group. In a multiple group IRT calibration, item parameters for two or more groups were estimated simultaneously as well as group mean differences on the latent variable (Embretson & Reise, 2000).

Embretson and Riese (2000) suggested multiple group IRT modeling was often helpful given the following: 1) the chosen IRT model fits the data for both groups, 2) the anchor items and their parameters are available, and 3) an appropriate number of anchor items are chosen thus ensuring the parameters are on the same scale (Embretson & Reise, 2000). The authors also found the results of the DIF analysis may be statistically significant yet posses no practical significance (Embretson & Reise, 2000).

Hierarchical linear modeling

Leeuw and Kreft (1995) described several classic examples of hierarchical data, such as but not limited to, students within classes; individuals in census tracts or political districts; and time points within individuals. Any data in which the observations can be classified into groups was considered hierarchical. Hierarchical models were used in the analysis of educational data because they allowed one to model the data giving consideration for its hierarchical nature (Anguiano, 2004; Draper, 1995; McCoach, et al., 2006; Morris, 1995). The use of a hierarchical model gave one the opportunity to separate the within-group variance from the between-group variance. Hierarchical models have become better understood and received more support in diverse fields, such as education, health and medicine, quality assurance, demography, and remote sensing (Morris, 1995). Hierarchical generalized linear modeling was proposed by Kamata (2001) as a method for detecting measurement invariance or subgroup differences. Mathematically speaking, the general form of a two level hierarchical model was described as:

Level 1 (Within unit effects): $Y_{ij} = \beta_{0j} + \beta_{1j}X_{ij} + r_{ij}$ where

β_{0j} was the estimated mean of Y_{ij} when independent variable, X = 0

β_{1j} was the change in Y_{ij} relative to a one unit change in X or slope.

r_{ij} was the level 1 error

Level 2 (Between unit effects): $\begin{aligned}\beta_{0j} &= \gamma_{00} + \gamma_{01}W_j + u_{0j} \\ \beta_{1j} &= \gamma_{10} + \gamma_{11}W_j + u_{1j}\end{aligned}$ where

γ_{00} was the grand mean

γ_{01} was the change of β_{0j} relative to a one unit change in W, or slope.

γ_{11} was the change of β_{1j} relative to a one unit change in W, or slope.

$u_{0j} ... \mu_{1j}$ was the level 2 error

This study specifically used the hierarchical generalized linear model, a type of HLM model that allows for a non-normal distribution of errors. The generalized linear model was used for prediction just as in regression. HGLM models allowed for the relationship between the continuous outcome variable and predictors to differ across individuals; this was referred to as a model with random effects. If variation across individuals was ignored; this was identified as a model with fixed effects (Kamata, 2001; Raudenbush & Bryk, 2002). There was the possibility of having an HGLM model with fixed effects on level one and random effects on the second or third levels. The level-1 model interpreted as an item-level model and the level-2 model a person-level model that allowed hypothesizing that the item coefficients were constant across people. If the model were extended to a level-3 model, a group-level model, could show that item effects were constant across gender and/or race (Kamata, 2001).

Previous detection of group differences

Numerous studies have applied the CFA method to detect measurement invariance across subgroups in the measurement of personality traits, creative thinking, pay satisfaction, mental ability, nonverbal intelligence, and psychological tests (Campbell, et al., 2008; Gustavsson, Eriksson, Hilding, Gunnarsson, & Ã–Stensson, 2008; Guttmannova, Szanyi, & Cali, 2008; Kim, K. H., Cramond, & Bandalos, 2006; Lievens, Anseel, Harris, & Eisenberg, 2007; Lubke, et al., 2003; Maller & French, 2004; Millsap, 1995, 2007; Richardson, Ratner, & Zumbo, 2007; Utsey, Brown, & Bolden, 2004; Wicherts, Dolan, & Hessen, 2005; Yin & Fan, 2003). All of the previously mentioned studies found the measure to be invariant across groups. Other studies used IRT methods to detect methods to detect measurement invariance across subgroups on measures of psychological status, achievement tests, and those of varying modes of delivery (Cauffman & MacIntosh, 2006; Cook, Eignor, & Taft, 1988; Fan, 1998; Ferrando & Lorenzo-Seva, 2005; Miller & Linn, 1988).

The strong relationship between CFA and IRT has been established over the years (McDonald, 1982; Reise, Widaman, & Pugh, 1993; Stark, Chernyshenko, & Drasgow, 2006). Studies used both CFA and IRT to determine the ability of the methods to demonstrate the measure was invariant across subgroups (Carle, et al., 2008; Kim, D. & Huynh, 2008; Reise, et al., 1993). The Carle, et al. (2008) study used a psychological test to determine if there were significant differences in the performance of girls and boys in the third and sixth grades on the Children’s Depression Inventory. In particular, the study confirmed through the use of a five-factor model for the CFA and the Rasch model from IRT; the groups were not statistically different. The Kim and Huynh (2008) study compared students taking a paper-and-pencil version of an English test to students taking a computer-based version of the English test. Using CFA and the Rasch model from IRT, the authors found the measure to be invariant across modes of administration. The Riese, et al. (1993) study used the CFA and graded response model, a model for polytomous items, from IRT to determine if the mood ratings were the same when collected from a group in Minnesota and a group in China. The authors also found the measure to be invariant across groups. A more recent study used to detect group differences in data was that of Willse and Goodman (2008). In this study, the authors demonstrated the advantages of using IRT over structural equation modeling (SEM; similar to CFA) or raw scores when assessing group differences. However, the results of this study suggested a researcher can use the methodology with which she or he was most familiar. These results were also found in similar studies that compared raw scores to IRT and classical test theory to IRT (DeMars, 2001; Fan, 1998). These findings supported the continued interchange of CFA and IRT.

There were fewer studies demonstrating the similarities in CFA and HLM. One study compared CFA and HLM and presented marginally different estimates for each method on a measure of school-based substance use (Chou, Bentler, & Pentz, 1998). Results in another study showed that both approaches offered nearly identical parameter estimates and standard errors; thus leading to identical conclusions about the data centered around married people (Wendorf, 2002). Although few in number, these studies supported the decision of a researcher to implement CFA or HLM.

Significance of study

Previous research shows the ability of CFA and IRT to be interchanged, and more recent studies show some support for the use of CFA and HLM for group comparisons. None of the previously mentioned literature addressed the use of HLM, CFA, and Rasch to detect group differences. This study seeks to adds to the psychometric literature by demonstrating the use of HLM, CFA, and Rasch item analysis models in establishing test validity via measurement invariance across subgroups at the test, subtest, and item levels using empirical data from a primary reading assessment. The use of HLM will support the comparison of scores at the test level, often done by policymakers. CFA will support the use of group comparisons at the subtest level, often used by instructional designers. The Rasch item analysis will support the comparison of groups across items, which is important to the psychometric community. Examining the various levels of groups comparisons on a single assessment will remind policymakers, instructional designers, and psychometricians of the importance of test validity.

Chapter 3

Methodology

The purpose of this study is to demonstrate how construct validity evidence is established using three methods to detect measurement invariance: hierarchical linear modeling, confirmatory factor analysis, and Rasch item response theory. The use of three methods is to provide validity evidence for comparisons across administrative, gender, and ethnic groups at the test, subtest, and item levels. In particular, this study will answer the following research questions:

Research questions

1. Are the group level coefficients from the hierarchical linear modeling technique invariant across administrative, gender, and ethnicity groups?
2. Are the group measures produced through confirmatory factor analysis (i.e. factor structure, factor loadings, and error variances) invariant across administrative, gender, and ethnicity groups?
3. Are the item difficulty estimates and fit statistics based on Rasch item response theory modeling stable across administrative, gender, and ethnicity groups?

General methodology

This section of the dissertation outlined the procedures used to answer the research questions. It begins with a description of the assessment. Next, the descriptive statistics of the data are presented. The descriptive data gives a count of the students who took the test by year, gender, and ethnicity. The remaining sections describe the procedures used to address the research questions in detail. The data set identifies the test date, gender, ethnicity, and item responses for each student. The initial comparison of student subgroups was across administration dates. If the measure is deemed invariant across administration dates; comparison of the gender and ethnic subgroups can collapse the data across dates. If the administration dates are not invariant, the gender and ethnicity subgroup comparisons will be conducted by administration date. For the gender subgroups, females are compared to males; for the racial/ethnic groups the achievement of White, Black, and Hispanic students will be compared. This choice of ethnicities is due to their historical performance in relation to White students and the small number of observations available for the remaining ethnic groups.

Instrument

The instrument was a multiple choice reading assessment for students in primary grades. The items for this assessment were chosen from the item bank of a long-standing reading test with strong psychometric characteristics. The item responses were from the spring of 2005, 2006, and 2007 administrations of the summative assessment which had a Cronbach's alpha (1951) of 0.91 for each administration. The results of the spring administrations determined the effectiveness of the reading intervention. Given the varied rates at which students become proficient in reading, this study used the data from the third grade reading assessment. The grade 3 test specifications identified four content areas within the assessment. The specifications were covered with at least nine items per content area, thus supporting the use of confirmatory factor analysis (Marsh, Hau, & Balla, 1995). Table 1 lists the number of items and the reliability coefficient, Cronbach's alpha (1951), per content area with most of the 72 items addressing phonemic awareness/phonics or reading comprehension. The reliability coefficient alpha suggested the subtest score variances due to true score variances ranged from 76% to 80%, which met the expected values of 0.7 or greater defined by Nunnally (1978).

Table 1.

Number of Items and Reliability per Content Area.

Content Area	Number of items	Cronbach's alpha
Phonemic awareness/Phonics	24	0.78
Listening vocabulary	9	0.85
Reading vocabulary	9	0.77
Reading comprehension	30	0.76

Participants

The participants in the data set represented grade 3 students from schools in a southeastern state. Table 2 gives the number of participants by year of administration, gender, and ethnicity. There were eight students missing demographic data, all of which were from the 2005 administration and were excluded from the analyses. The number of males in the sample slightly outweighed the number of females for each of the administrations. The number of Black students in the sample was approximately three times as large as the number of White students in the sample across each administration. The number of Hispanic students in the sample was less than 100 for each of the administrations.

The 2005 administration had 2,908 grade 3 student responses, 48% of which were females and 52% of which were males. This group of students included 73% Black or African American, 22% White, 3% Hispanic, and less than 1% from the Asian, American Indian, and Other ethnic categories. The 2005 data set had five observations missing values for the gender demographic and three observations missing values for the ethnicity demographic.

Table 2.

Number of Participants per Administration by Gender and Ethnicity.

Demographics	2005	2006	2007	Total
Gender				
Male	1,500	1,382	1,472	4,354
Female	1,403	1,349	1,349	4,101
Missing	5	0	0	5
Ethnicity				
Black/African American	2,129	2,015	1,960	6,104
White	652	595	710	1,957
Hispanic	84	70	91	245
Asian	13	13	6	32
American Indian	8	10	12	30
Other	19	28	42	89
Missing	3	0	0	3
Total Participants	2,908	2,731	2,821	8,460

The 2006 administration had 2,731 grade 3 student responses, 49% of which were females and 51% of which were males. This administration also included the item responses of 74% Black or African American students, 22% White students, 3% Hispanic students, and 1% or less of the students responses were from the Asian, American Indian, and Other ethnic groups.

The 2007 administration had 2,821 grade 3 students, 48% of which were females and 52% of which were males. The 2007 administration also had 69% Black or African American student responses, 25% White student responses, 3% Hispanic student responses, and 1% or less of the students responses were from the Asian, American Indian, and Other ethnic groups.

Method for question 1: Hierarchical linear modeling

Hierarchical linear modeling (HLM), a type of generalized linear model, typically used to account for non-independent or nested data was used in this study. This model was similar to an ordinary least squares (OLS) regression model that allowed for more accurate estimates and the ability to separate the variability due to within group differences from that of the between group differences. The major difference was the effect of nested data. Perhaps one had an interest in modeling the effect of school on student achievement in a three level hierarchical linear model; the student characteristics modeled at the first level, the classroom characteristics modeled at the second level and the school characteristics at the third level. Suppose, the general structure of this three level hierarchical model was written by level using the following equations:

Level 1(students): $Y_{ijk} = \beta_{0jk} + r_{ijk}$

where Y_{ijk} was the outcome variable of student *i* in classroom *j* and school *k*;

β_{0jk} was the average of the outcome variable of classroom *j* in school k; and

r_{ijk} was the random "student effect."

Level 2 (classroom): $\beta_{0jk} = \gamma_{00k} + u_{0jk}$

where γ_{00k} was the mean achievement in school *k*; and

u_{0jk} was the random "classroom effect."

Level 3 (schools): $\gamma_{00k} = \pi_{000} + u_{00k}$

where π_{000} was the grand mean ;

u_{00k} was the random "school effect."

This methodology, in the HLM 6.0 software, used the restricted maximum likelihood estimation procedure, which is not conditional on point estimates of the fixed effects (Kamata, 2001; Raudenbush & Bryk, 2002). This study specifically used the hierarchical generalized linear model, a type of HLM model that allowed for a non-normal distribution of errors. In ordinary least squares (OLS) regression, the residuals or the distribution of errors were assumed to have a normal distribution with a mean of zero and some variance; also applicable in HLM (Agresti & Finlay, 1997). However, in the hierarchical generalized linear model, this assumption about the residuals was relaxed, thus the prediction errors may have a non-normal distribution (Kamata, 2001). In this study, level one represented the fixed items, level two represented the person characteristics and level three represented the administrative, gender, and ethnic group characteristics. Specifically, the item level model, or level 1 model, represented the effect of item difficulty on overall score. The level 1 model was:

$$\log\left(\frac{p_{ijm}}{1-p_{ijm}}\right) = \beta_{01jm} + \beta_{1jm}X_{1ijm} + \beta_{2jm}X_{2ijm} + \ldots + \beta_{(k-I)jm}X_{(k-1)ijm},$$

where items were represented by *i* (*i* = 1, …, k); persons were represented by *j* (*j*=1, …, n); and subgroups were represented by *m*. In addition, p_{ijm} was the probability that person *j* from subgroup *m* answers item *i* correctly, X_{qijm} was defined as the *q*th dummy variable $(X = -1$ when $q = i$ and $X = 0$ when $q \neq i)$, q = 1, …, *k*-1 for the *i*th item for person *j* in subgroup *m*, β_{0jm} was the expected effect of the reference item for person *j* from subgroup *m*, and β_{qjm} was the expected effect of the *q*th item for person *j* from subgroup *m* compared to the reference item. “For the design matrix of the model to achieve full rank, one of the dummy variables in the equation [was] dropped. This constraint [resulted] in an interpretation of β_{0jm} as the expected item effect in absolute value of the dropped item for person *j*. The individual item effect, in this context, [was] defined as the *difference* of effect from β_{0jm}” (Kamata, 2001, p. 82). Structuring the model in this way allowed for examination of item effects.

The equation for the person-level model, level 2, represented the effect of person ability and item difficulty on the overall score. This model was algebraically equivalent to the Rasch model, $p_{ijm} = \dfrac{1}{1+\exp\left[-(\theta_{jm}-\delta_{im})\right]}$, where $\theta_{jm} = r_{0jm}$ and $\delta_{im} = -\gamma_{q0m} - \gamma_{oom}$ and *i*=*q*(*i*=1, …, *k*-1). This model was expressed as

$$p_{ijm} = \frac{1}{1+\exp\left\{-\left[r_{0jm}-\left(-\gamma_{q0m}-\gamma_{00m}\right)\right]\right\}} \text{ or}$$

$$\begin{array}{ll} \beta_{0jm} = \gamma_{00m} + r_{0jm} & \\ \beta_{1jm} = \gamma_{10m} & \\ \beta_{2jm} = \gamma_{20m} & \text{where } r_{0jm} \sim N\left(r_{00m}, \tau_{\gamma}\right), \\ \ldots & \\ \beta_{(k-1)jm} = \gamma_{(k-1)0m} & \end{array}$$

$-\gamma_{q0m} - \gamma_{00m}$ was the item difficulty for item *i* for *i*=*q*(*i*=1, …, *k*-1), γ_{00m} was the item difficulty for the *k*th item, and r_{0jm} was the random person ability and indicated how much person *j* from subgroup *m* was deviated from the mean, r_{00m}, within subgroup *m*. The variance of r_{0jm} within subgroup was denoted τ_{γ} and assumed identical for all groups. At this level, the item parameters were fixed across person and vary across items, thus there is only one random term for person ability.

The equation for the subgroup level, level 3, represented the effect of person ability, item difficulty, and group membership on overall score and was also algebraically equivalent to the Rasch model, $p_{ijm} = \dfrac{1}{1+\exp\left[-(\theta_{jm}-\delta_{im})\right]}$, where $\theta_{jm} = u_{00m} + r_{0jm}$ and

$\delta_{im} = -\pi_{q00} - \pi_{000}$. This model was expressed as

$$p_{ijm} = \frac{1}{1+\exp\{-[(u_{00m}+r_{0jm})-(-\pi_{q00}-\pi_{000})]\}} \text{ or}$$

$$\begin{aligned}
\gamma_{00m} &= \pi_{000} + u_{00m} \\
\gamma_{10m} &= \pi_{100} \\
\gamma_{20m} &= \pi_{200} \qquad \text{where } u_{00m} \sim N(0, \tau_{\pi}), \\
&\ldots \\
\gamma_{(k-1)0m} &= \pi_{(k-1)0m}
\end{aligned}$$

$-\pi_{q00} - \pi_{000}$ was the item difficulty for item *i* for *i*=*q*(*i*=1, …, k-1), π_{000} was the item difficulty for item *k*, $u_{00m} + r_{0jm}$ was the ability for person *j* in subgroup *m*, u_{00m} was the average ability of students in subgroup *m*, and r_{0jm} indicates how much the ability of person *j* deviated from the average ability of students in school *m*. The significance of the random coefficients for level three were used to detect measurement invariance (Kamata, 2001).

Model assumptions and model fit

The key assumptions for the hierarchical generalized linear model identified by Hoffman, Griffin, and Gavin (2000, p. 490) were: 1) the lower-level units are nested within identifiable higher-level units, 2) the lower-level units were exposed to and influenced by characteristics and/or processes of higher-level units, 3) the outcome variable was measured at the lowest level of interest to the researcher, and 4) the outcome variable varied both within the lower-level units and between the higher-level units. Determining whether the data met these assumptions did not require any statistical tests; the qualitative nature of the data was evaluated to determine if these assumptions were met.

Other statistical assumptions identified by Hoffman, et al. (2000, p. 490) and Raudenbush and Bryk (2002, p. 255) for a two level model included 1) level 1 residuals were independent and normally distributed with mean zero and variance σ^2 for every level 1 unit within each level 2 unit; 2) level 1 predictors were independent of level 1 residuals, 3) random errors at level 2 were multivariate normal (each with a mean of zero and variance of τ_{qq}, and a covariance of $\tau_{qq'}$) and were independent among level 2 units, 4) the set of level 2 predictors were independent of every level 2 residual, 5) residuals at level 1 and level 2 were also independent, and 6) the predictors at each level were not correlated with the random effects at the other level. The independence assumptions for the level 1 predictors and level 1 residuals were checked by evaluating the Pearson correlation, which has an expected value of zero in this case. The assumptions of normality and model fit were checked using the normal probability plots in PROC UNIVARIATE in SAS (SAS Institute, 2003). Finney and DiStefano (2006) also suggested that skewness vales less than 2.0 and kurtosis values less than 7.0 meet the normality assumption.

Model testing

Item analysis for the HLM technique followed the model outlined by Kamata (2001) where level one represented the fixed items, level two represented the person characteristics and level three represented the administrative, gender, and ethnic group characteristics. This method estimated the coefficients simultaneously for each level of the model, therefore the model testing was completed in one step. The HLM 6.0 software (Raudenbush, Bryk, & Congdon, 2008) was used to test the model (See Appendix A for syntax). The significance of the coefficients for level three were used to detect measurement invariance (Kamata, 2001).

Method for question 2: Confirmatory factor analysis

Figure 1 shows the model tested for each administrative, gender, and ethnic subgroups represented by *g* in the figure. The figure has one latent variable, reading, and four observed variables/factors: phonics/phonemic awareness, listening vocabulary, reading vocabulary, and reading comprehension. The observed variables, also known as components, factor scores, or item parcels, were created by summing up the item scores within each content area (Little, Cunningham, Shahar, & Widaman, 2002). An item parcel was defined by Little (2002, p. 152)as "an aggregate-level indicator comprised of the sum (or average) of two or more items, responses, or behaviors." This method of item parceling improved the stability of the factor solution by improving the item to subject ratio and continuity of data (Marsh & Hocevar, 1988). The phonics/phonemic awareness factor score was created by summing up the scores for the related 24 items. The factor scores for listening vocabulary, reading vocabulary, and reading comprehension were

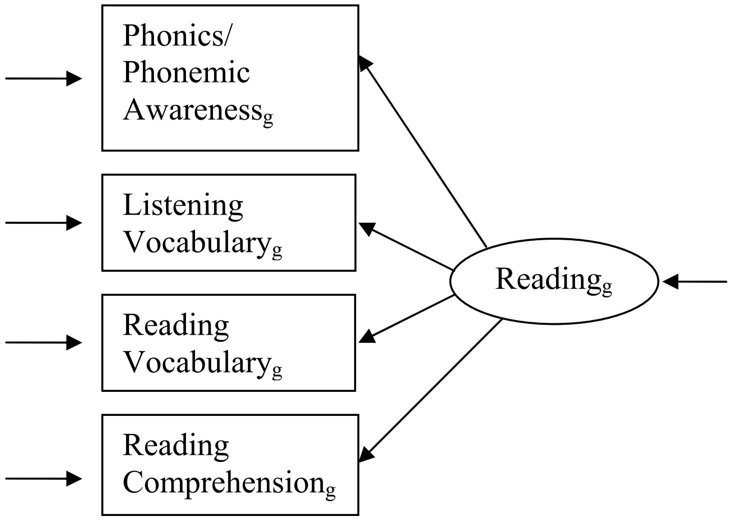

Figure 1. Model used to test for measurement invariance across *g* groups.

similarly created by summing up the scores for the related 9, 9, and 30 items, respectively.

Model assumptions

Confirmatory factor analysis (CFA) was used to partition the variance among the four item parcels for each of the following subgroups: 2005, 2006, and 2007; males and females; and White, Black, and Hispanic. The process began by checking the assumption of multivariate normality of the item parcels using the criteria outlined by Finney and DiStefano (2006), which suggested skewness values less than 2.0 and kurtosis values less than 7.0 meet this assumption. The next step to checking assumptions for measurement invariance, was an omnibus test of the homogeneity of the variance-covariance matrices (Box, 1949), $\Sigma_{Group1} = \Sigma_{Group2}$, where

$$\Sigma = \begin{bmatrix} \sigma_{11} & & & \\ \sigma_{21} & \sigma_{22} & & \\ \sigma_{31} & \sigma_{32} & \sigma_{33} & \\ \sigma_{41} & \sigma_{42} & \sigma_{43} & \sigma_{44} \end{bmatrix},$$

σ_{ij} represented covariances and when i=j, σ_{ij} represented the variances.

If this test was not significant, one assumed the groups had configural, metric, and scalar invariance. This test was conducted using the PROC DISCRIM procedure in SAS (Ritz & Brockhoff, 2005). This test compared the variance-covariance matrices for administration dates of 2005, 2006, and 2007. The variance-covariance matrix for males was compared to that of females. The variance-covariance matrices of Whites, Blacks, and Hispanics were compared, yielding the following null hypotheses:

$$\Sigma_{2005} = \Sigma_{2006} = \Sigma_{2007}, \ \Sigma_{Male} = \Sigma_{Female}, \text{ and } \Sigma_{White} = \Sigma_{Black} = \Sigma_{Hispanic}.$$

If the hypotheses were true, one assumed that the measure was invariant for the subgroups. If the variance-covariance matrices of the groups were significantly different, the Vandenberg and Lance (2000) sequence of steps was used to identify the source of invariance.

Model testing

Vandenberg and Lance (2000) outlined a sequence of model comparisons needed to provide evidence of measurement invariance across subgroups. This sequence of steps was necessary if and only if the omnibus test was significant or the variance-covariance matrix was not the same across groups. Lisrel 8.5 (Joreskog & Sorbom, 2001) software was used to conduct the Vandenberg and Lance (2000) methodology; which had the following sequence with steps 4 and 7 being optional:

1. Freely estimate the baseline model
2. Constrain factor loadings to equality
3. Constrain error variances to equality
4. Constrain item means to equality (optional)
5. Constrain factor variances to equality
6. Constrain factor covariances to equality
7. Constrain factor means to equality (optional).

The Vandenberg and Lance (2000) methodology was ideal for addressing the invariance of the mean and covariance structures (MACS). This study was based on covariance structures only because the latent means were not of interest; thus, eliminated the need to examine the mean structures in steps 4 and 7. If at any point during the procedure the constrained model was significant, partial invariance was tested. If the model with partial invariance was not significant, the model had some invariance. However, if all possible models with partial invariance were significant, the sequence terminated. The chi-square difference test was used to determine if the model with more constraints significantly improved model fit (Jöreskog, 1978). If the result was significant, the stricter model was considered the best model to describe the data.

Step1: Configural invariance (equal factor structures)

Configural invariance tested the overall factor structure of the groups by gender, ethnicity, and administrative date. In this step, the baseline model was freely estimated for each of the groups (See Appendix B for Lisrel syntax). Configural invariance was useful to determine whether each group had an equal number of factors and whether the items were loading on the same factor across groups (Campbell, et al., 2008). Mathematically, the configural invariance test for the administrative subgroups had the following hypothesis: $\Lambda\xi_{2005} = \Lambda\xi_{2006} = \Lambda\xi_{2007}$. where Λ was the vector of observed item responses and ξ was the latent variable being measured. Similarly, the hypotheses for the configural test for the gender subgroups was $\Lambda\xi_{Male} = \Lambda\xi_{Female}$ and the ethnic subgroups was $\Lambda\xi_{White} = \Lambda\xi_{Black} = \Lambda\xi_{Hispanic}$.

The test or assessment in this measure was expected to assess the single latent variable of reading. This latent variable was described by the factor scores of four content areas: phonics/phonemic awareness, listening vocabulary, vocabulary development, and reading comprehension. If the hypothesis of unidimensionality was not true for each of the groups, one assumed the groups were not being tested on the same concept and none of the remaining steps were conducted. If this hypothesis was rejected, there was not enough evidence to support across group comparisons.

Step 2: Metric invariance (equal factor loadings)

If configural invariance was established, metric invariance was the next test. Metric invariance tested the extent to which the relationships between the factors and the items were equivalent across the groups. If metric invariance was supported, Byrne (1998) suggested that the groups were interpreting the items in the same way. A lack of metric invariance may imply that some items were more ambiguous for one group than for another (Campbell, et al., 2008; Chan, 2000).

Before metric invariance was tested, a referent item was chosen to set the metric for each factor or to allow the comparison of loadings and error variances. For an item to serve as a referent item for a factor, it must be invariant across the groups. To ensure the referent item was invariant across groups, each of the other items on the subscale can be used as a temporary referent item to ensure that the target item remains invariant across samples (Cheung & Rensvold, 1999). The hypotheses tested with metric invariance for administrative, gender, and ethnicity, were:

$$\begin{bmatrix}\lambda_1 & \xi\\ \lambda_2 & \xi\\ \lambda_3 & \xi\\ 1 & \xi\end{bmatrix}_{2005} = \begin{bmatrix}\lambda_1 & \xi\\ \lambda_2 & \xi\\ \lambda_3 & \xi\\ 1 & \xi\end{bmatrix}_{2006} = \begin{bmatrix}\lambda_1 & \xi\\ \lambda_2 & \xi\\ \lambda_3 & \xi\\ 1 & \xi\end{bmatrix}_{2007}, \begin{bmatrix}\lambda_1 & \xi\\ \lambda_2 & \xi\\ \lambda_3 & \xi\\ 1 & \xi\end{bmatrix}_{Male} = \begin{bmatrix}\lambda_1 & \xi\\ \lambda_2 & \xi\\ \lambda_3 & \xi\\ 1 & \xi\end{bmatrix}_{Female}\text{, and}$$

$$\begin{bmatrix}\lambda_1 & \xi\\ \lambda_2 & \xi\\ \lambda_3 & \xi\\ 1 & \xi\end{bmatrix}_{White} = \begin{bmatrix}\lambda_1 & \xi\\ \lambda_2 & \xi\\ \lambda_3 & \xi\\ 1 & \xi\end{bmatrix}_{Black} = \begin{bmatrix}\lambda_1 & \xi\\ \lambda_2 & \xi\\ \lambda_3 & \xi\\ 1 & \xi\end{bmatrix}_{Hispanic},$$

where $\lambda_1, \lambda_2, \text{and } \lambda_3$ represented the factor loadings of the observed variables on the latent variable, ξ and 1 is the invariant referent item (See Appendix C for syntax).

Cheung and Rensvold (2002) recommended comparing the difference in the comparative fit index (CFI) value between a model imposing equality constraints on the factor loadings between two groups versus a model estimating a particular factor loading separately to evaluate metric invariance. If the difference in CFI was less than 0.01

(ΔCFI< 0.01), then strengths of factor loadings were considered invariant across groups (i.e., the scale demonstrates metric invariance) (Karazsia, van Dulmen, & Wildman, 2008). This procedure was used to evaluate metric invariance in this study. If this evidence was not significant, the test for scalar invariance was conducted.

Step 3: Scalar invariance (error variances constrained to be equal)

The third step of measurement invariance testing, scalar invariance, tested the equality of error variances or intercept terms. This step, allowed one to determine whether the groups used the response scale in a similar way (Hong, et al., 2003; Steenkamp & Baumgartner, 1998; Vandenberg & Lance, 2000). Unless scalar invariance was supported, the validity of inferences pertaining to group comparisons was questionable (Campbell, et al., 2008). Mathematically, scalar invariance was represented by the equality of the following vectors for the gender, ethnic and administrative groups, respectively:

$$\begin{bmatrix}\delta_1\\ \delta_2\\ \delta_3\\ \delta_4\end{bmatrix}_{2005} = \begin{bmatrix}\delta_1\\ \delta_2\\ \delta_3\\ \delta_4\end{bmatrix}_{2006} = \begin{bmatrix}\delta_1\\ \delta_2\\ \delta_3\\ \delta_4\end{bmatrix}_{2007}, \begin{bmatrix}\delta_1\\ \delta_2\\ \delta_3\\ \delta_4\end{bmatrix}_{Male} = \begin{bmatrix}\delta_1\\ \delta_2\\ \delta_3\\ \delta_4\end{bmatrix}_{Female}, \text{ and}$$

$$\begin{bmatrix}\delta_1\\ \delta_2\\ \delta_3\\ \delta_4\end{bmatrix}_{White} = \begin{bmatrix}\delta_1\\ \delta_2\\ \delta_3\\ \delta_4\end{bmatrix}_{Black} = \begin{bmatrix}\delta_1\\ \delta_2\\ \delta_3\\ \delta_4\end{bmatrix}_{Hispanic}$$

where $\delta_1, \delta_2, \delta_3,$ and δ_4 represented the error variances associated with each observed variable or item parcel (See Appendix D for syntax).

Step 4: Equal item means

Scalar invariance provided the evidence necessary to interpret item mean differences between the groups (Meredith, 1993; Steenkamp & Baumgartner, 1998). The evaluation of mean level differences required the item means to be constrained to equality for the groups. However, given this study only examined the covariance structures as support of measurement invariance; this step was not conducted.

Step 5: Equal factor variances

Prior to beginning this sequence of steps, an omnibus test of homogeneity of the variance-covariance matrices for each group was conducted which conceptually addressed the fifth and sixth steps. The sequence of steps was followed if and only if the omnibus test was significant. If the omnibus test was significant and there was evidence of configural, metric, and scalar invariance; the next step was to determine if the differences in the groups were within the factor variances. Factor variances represented the variability of the reading construct within each group. The hypotheses tested for the administrative, gender, and ethnic groups, respectively, in this step were:

$$\sigma_{\xi\,2005} = \sigma_{\xi\,2006} = \sigma_{\xi\,2007},\ \sigma_{\xi\,Male} = \sigma_{\xi\,Female},\text{ and }\sigma_{\xi\,White} = \sigma_{\xi\,Black} = \sigma_{\xi\,Hispanic}.$$

These null hypotheses assumed the variability of the reading construct was the same for each group. If this variability was the same for each group, it was assumed the groups used equivalent ranges of the reading construct continuum to respond to the items. If the null hypothesis was rejected, it was concluded that at least one of the groups was using a smaller range of the construct continuum to respond to the items.

Step 6: Equal factor covariances

The sixth step examined the possibility of the factor covariances being equal across the groups, similar to the omnibus test. Given this test had one latent variable, the covariance between latent variables was not applicable. However, if there was more than one latent variable, and all of the previous tests did not show significance, the final test of the following administrative, gender, and ethnic hypotheses was expected to be significant:

$$\begin{bmatrix} & & & \\ \sigma_{21} & & & \\ \sigma_{31} & \sigma_{32} & & \\ \sigma_{41} & \sigma_{42} & \sigma_{43} & \end{bmatrix}_{2005} = \begin{bmatrix} & & & \\ \sigma_{21} & & & \\ \sigma_{31} & \sigma_{32} & & \\ \sigma_{41} & \sigma_{42} & \sigma_{43} & \end{bmatrix}_{2006} = \begin{bmatrix} & & & \\ \sigma_{21} & & & \\ \sigma_{31} & \sigma_{32} & & \\ \sigma_{41} & \sigma_{42} & \sigma_{43} & \end{bmatrix}_{2007},$$

$$\begin{bmatrix} & & & \\ \sigma_{21} & & & \\ \sigma_{31} & \sigma_{32} & & \\ \sigma_{41} & \sigma_{42} & \sigma_{43} & \end{bmatrix}_{Male} = \begin{bmatrix} & & & \\ \sigma_{21} & & & \\ \sigma_{31} & \sigma_{32} & & \\ \sigma_{41} & \sigma_{42} & \sigma_{43} & \end{bmatrix}_{Female} \text{, and}$$

$$\begin{bmatrix} & & & \\ \sigma_{21} & & & \\ \sigma_{31} & \sigma_{32} & & \\ \sigma_{41} & \sigma_{42} & \sigma_{43} & \end{bmatrix}_{White} = \begin{bmatrix} & & & \\ \sigma_{21} & & & \\ \sigma_{31} & \sigma_{32} & & \\ \sigma_{41} & \sigma_{42} & \sigma_{43} & \end{bmatrix}_{Black} = \begin{bmatrix} & & & \\ \sigma_{21} & & & \\ \sigma_{31} & \sigma_{32} & & \\ \sigma_{41} & \sigma_{42} & \sigma_{43} & \end{bmatrix}_{Hispanic}.$$

Vandenberg and Lance (2000) suggested the interpretation of significant results would be hard to explain if the hypothesis of a more stringent test, such as configural invariance, was not rejected. Thus, if the factor covariances were equal, this was detected with the omnibus test.

Step 7: Equal factor means

The final step of constraining the factor means to be equivalent across the groups provided additional evidence to declare the measure invariant across groups and also make comparisons across groups. However, this study used the covariance structure to compare groups, not the MACS; therefore this step was not conducted.

Model fit

As the various steps were completed, the chi-square difference test was used with caution for model comparisons (Yuan & Bentler, 2004). Other fit statistics that noted included the root mean square error of approximation (RMSEA), comparative fit index (CFI), non-normed fit index (NNFI) and the standardized root mean residual (SRMR). RMSEA was described by Marsh, Balla, and McDonald (1988) as an estimate of the error associated with the model per degree of freedom which was sensitive to misfit among latent variables with values less than 0.05 indicating a good fit and values around 0.08 indicating adequate fit. The CFI and NNFI were described by Hu and Bentler (1999) as sensitive to misspecification among the measurement model with values greater than or equal to 0.95 indicating good fit. Hu and Bentler (1999) described the SRMR as an average measure of the differences between the observed variances and covariances in the model based on standardized residuals with values less than 0.05 indicating a good fit, and values less than 0.08 indicating adequate fit.

Method for question 3: Item response theory via the Rasch model

There were several IRT-based logistic models discussed in the literature review that related the probability of a correct item response to an examinee's ability. The Rasch item response theory (IRT) model was used in this study to express the probability of a person of a given ability successfully responding to an item. The Rasch model or the one parameter logistic model (1PL), assumed all items have the same discrimination parameter and participants did not guess when responding to items (Rasch, 1960/1980). In 1960, Rasch defined the 1PL as $P(\theta) = \frac{e^{Da(\theta-b)}}{1+e^{Da(\theta-b)}}$, where D = 1 or 1.7, a was the item discrimination parameter, θ represented the ability level and b represented the item difficulty which increases in value as the item became more difficult, but it was not dependent on ability level. The two-parameter model recognized the changes of the item difficulty and the item discrimination when measuring the examinee's performance. The three-parameter model recognized the changes in item difficulty, item discrimination and a guessing parameter, when measuring the examinee's performance. The DeMars (2001) study reviewed other studies that found the 3PL model had a better fit than the 1PL for multiple choice items, but questioned whether this fit mattered when comparing test scores are used to compare conditions or groups. The author found the effect sizes from 3PL model estimates and 1PL model estimates were similar, thus suggesting one would not improve the results by much choosing one model over the other. Due to its simplicity and popularity in state achievement tests (i.e., Arkansas, Maryland, and South Carolina), the Rasch model was used in this study to simultaneously estimate the person and item characteristics (Huynh & Rawls, 2009). The estimate of item characteristics was that of the item difficulty and the estimate of person characteristics was the probability that a person of a given ability level accurately respond to an item. The estimates were examined for each group at the item level to determine if the estimates for one group were significantly different from the estimates from the other groups. This process was done for administrative, gender, and ethnic subgroups.

Model assumptions

The Rasch model predicted a person's probability of success on an item given their ability level and the item difficulty. A person's ability level and the item difficulty values were set on the same scale, logits. When person and item parameters were estimated, the average item difficulty was set to zero. The result represented the probability of a person with average ability successfully answering an item given the item's difficulty. The item difficulty and person ability levels were simultaneously estimated using Winsteps software (Linacre, 2006) for each administrative, gender, and ethnic group. Prior to estimation, the assumptions of models based on item response theory was checked. These assumptions included the monotonically, increasing (S-shaped) item characteristic curves (ICC) and the existence of local independence (Embretson & Reise, 2000). Item characteristics curves represented the probability that an individual will get an item correct conditional on their ability. A monotonically, increasing (S-shaped) ICC suggested that persons with a lower ability have a lower probability of getting an item correct and this probability increased as the person ability increased. The local independence assumption suggested an item is independent of the other items and it did not cue responses to other items. Checking of this assumption or examination of item content was not conducted in an effort to keep the operational items secure. Other assumptions of IRT included unidimensionality and speededness. The unidimensionality assumption suggested the construct measured was best described via a single latent trait. Hambleton and Swaminathan (1985) suggested the assumptions of unidimensionality can be checked using factor analysis. Speededness suggested that the test was not administered under timed conditions (Hambleton & Swaminathan, 1985). Thus, it was assumed, items to which persons do not respond was due to their limited ability and not because they lacked the time. This assumption suggested missing items in this analysis were incorrect responses and they were treated as such. The response patterns were scanned to ensure the items near the end of the test were not blank, this pattern may suggest the test was timed.

The next assumption necessary to model the data using the Rasch model was local independence, the probability of a person responding to an item without the use of other items or no use of contextual cueing. Embretson and Reise (2000) gave the following mathematical definition: $P\left(X_{is} = 1 \middle| X_{i's} = 1, \xi_k, \theta_s\right) = P\left(X_{is} = 1 \middle| \xi_k, \theta_s\right)$ where the probability of a person *s* correctly solving item *i*, was independent of how the person responded to any other items i' controlling for the item characteristics and person characteristics, ξ_k, θ_s, respectively. However, this assumption was not checked for this data. Items were not examined on a contextual level to ensure the security of the operational items was not compromised.

The unidimensional assumption for this data was checked using the results of the confirmatory factor analysis from research question 1. If the data fit the model from Figure 1, with a single latent variable, it was concluded the assumption of unidimensionality was met. If the data did not meet the assumption of unidimensionality, the item response theory methodology was not conducted and the expected results of the CFA are those of a multidimensional structure or an imperfect fit. When in this situation, other multidimensional models may fit the data better.

If the data met the speededness and unidimensional assumptions, a graphical representation of the item characteristic curves (ICC) was used to determine if the data met the final assumption of monotonically, increasing ICCs. A graphical representation of the ICCs was created for all items using Winsteps software (Linacre, 2006). Theoretically, one would expect the ICCs to produce monotonically, increasing (S-shaped) curves, similar to Figure 2, however, the empirical data may not create such smoothe curves. Figure 2 represents the theoretically based ICCs for 10 dichotomous items; the y-axis represents the item difficulty measure and the y-axis represents the probability a person will correctly answer the item or a person's score on the item. The item characteristic curves for dichotomous items based on the Rasch model have the same slope or discrimination and only change in difficulty (Rasch, 1960/1980).

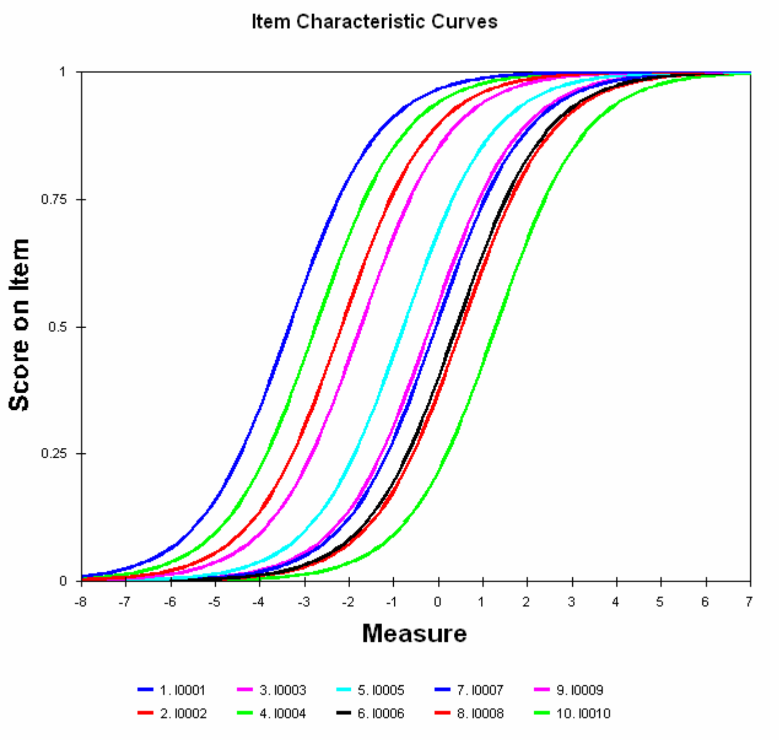

Figure 2. Theoretical representation of item characteristic curves.

Model testing and model fit

Model testing and model fit for the Rasch item response methodology was conducted simultaneously, similar to the simultaneous estimation of person ability and item difficulties. This calibration procedure was conducted in Winsteps software (Linacre, 2006) by subgroup: 2005, 2006, and 2007; male and female; and White, Black, and Hispanic. Embretson and Reise (2000) identified the most common method for item calibration as maximum likelihood, which is centered around finding the value of person's ability, θ, that maximizes the probability of an person's response pattern. The mean person ability and mean item difficulty was set to zero for ease of interpretation of the estimates. The item calibration was run on the full sample and then run separately for the 2005, 2006, 2007, male, female, White, Black, and Hispanic subgroups (See Appendix E for Winsteps control file details).

Model fit was examined using the standardized mean square infit statistic, INFIT-ZSTD, and the standardized mean square outfit statistic, OUTFIT-ZSTD. The expected value of INFIT- ZSTD and OUTFIT-ZSTD statistics was zero with productive values for measurement ranging from -0.5 to 0.5 (Linacre & Wright, 1999). If the INFIT-ZSTD and OUTFIT-ZSTD values were acceptable, it was assumed the Rasch model fit the data well. If these values were not within the expected range, it was concluded the Rasch model was not the best fit for the data. Upon completion of the item calibration and examination of model fit the sequence of steps outlined by Kim and Huynh (2009) were used to determine the stability of the item difficulty estimates and fit statistics across the administrative, gender, and ethnic subgroups:

1. Correlation of item difficulty estimates and fit statistics
2. Mean differences in item difficultly estimates and fit statistics
3. Absolute differences in item difficulty estimates
4. Robust *Z* statistic for differences in item difficulty estimates and fit statistics.

Steps 1 to 4 were followed for item difficulty estimates and model fit statistics. The correlation was expected to be strong and positive, the mean of the differences was expected to be minimal, less than 20% of the item absolute differences were expected to be greater than or equal to 0.3 in absolute value, and less than 20% of the item robust *Z* statistics were expected to be greater than or equal to 1.645 in absolute value. The criteria for steps 3 and 4 were set from empirical work using the Rasch linking protocol (See Appendix F for details).

Step1: Correlation between item difficulty estimates and fit statistics

Upon completion of the item calibration, the item difficulty estimates and fit statistics were exported to SAS via Microsoft Excel in preparation to calculate the Pearson product moment correlation coefficient, *r*. The pairwise correlations for the administrative, gender, and ethnic subgroups was examined using PROC CORR in SAS. The subgroups were to have a strong positive correlation. The INFIT-ZSTD and OUTFIT-ZSTD statistics were expected show a strong, positive correlation between the subgroups. If these subgroups showed strong, positive correlations for the item difficulty estimates and fit statistics; one may conclude the items were stable across groups for the first step in the Rasch item analysis. If the correlation was strong and negative for the item difficulty estimates and/or the fit statistics, one assumed the items were easier for one subgroup than the other and the Rasch model fit one subgroup well and the other poorly.

Step 2: Mean differences in item difficulty estimates and fit statistics

The second step in the process began by finding the differences between the item difficulty estimates and the fit statistics using Microsoft Excel. The mean of the differences in the item difficulty estimates and the mean of the differences in the fit statistics between pairwise administrative subgroups were calculated. These mean differences were also calculated for the pairwise gender and ethnic subgroups. During item calibration, the mean for the item difficulty estimates were set to zero, which suggested the mean of the differences in item difficulty estimates and mean of the differences in fit statistics between subgroups was expected to be zero. Kim and Huynh (2009) suggested this step produce mean differences close to zero for the item difficulty estimates and the fit statistics to endorse the notion of the items being stable across groups for step 2 in the Rasch item analysis.

Step 3: Absolute differences in item difficulty estimates

The third step in the procedure examined the data exported during the second step. The second step used the data to examine the mean of the differences, while this step examined the individual item differences. The differences in item difficulty estimates for the administrative, gender, and ethnic subgroups was calculated for each item. The differences in item difficulty estimates by item were expected to be less than or equal to 0.3 in absolute value for each of the groups (Huynh & Rawls, 2009). From Rasch linking protocol, it was suggested that no more than 20% of the items display an absolute difference in item difficulty estimates larger than 0.3 in absolute value (H. Huynh, personal communication, March 22, 2009). Removal of more than 20% of the items may be interpreted as a change in the test specification and/or the construct being measured (H. Huynh, personal communication, March 22, 2009).

Step 4: Robust Z statistic for differences in item difficulty estimates and fit statistics

The fourth step was completed using the data exported from the second step and the item level differences in difficulty estimates and item level differences in fit statistics from the third step. This final step used the item level differences in difficulty estimates and item level differences in fit statistics for the administrative, gender, and ethnic subgroups to calculate the robust *Z* statistic. Huynh defined this statistic as:

$$Robust\ Z = \frac{D - Median}{0.74(IQR)}$$

where *D* was the difference in the item difficulty estimates or differences in fit statistics between subgroups, *Median* was the median of the *D*'s, and *IQR*, was the interquartile range of the *D*'s (Huynh, Gleaton, & Seaman, 1992). These calculations resulted in item level robust *Z* statistics for the gender and ethnic groups. The value of the item level robust *Z* statistic was expected to be less than or equal to 1.645 in absolute value for items that were invariant across groups. Those items that were not stable across groups were expected to have values outside of this range. With successful completion of steps 1 to 4, one was able to declare the measure stable across groups or identify specific items that were not stable across groups. Using the guidelines for the Rasch linking protocol, no more than 20% of the items were expected to have a Robust Z statistic greater than 1.645 in absolute value (H. Huynh, personal communication, March 22, 2009).

Summary of methodology

The methods presented in this section sought to demonstrate how construct validity evidence for measurement invariance was produced using hierarchical linear modeling, confirmatory factor analysis, and Rasch item analysis. Each methodology was unique in that it supported the various levels of interpretations across subgroups. If the measure was invariant using hierarchical linear modeling, it supported the comparison of the groups across the entire measure. If the measure was invariant using CFA, it supported the comparison of groups across factors or item parcels. If the measure was invariant using Rasch item analysis, it supported the comparison of groups across items. If the hierarchical methodology refuted the comparison of the measure across groups, one expected to find partial or no invariance using CFA and/or significant differences across groups at the item level.

Chapter 4

Results

The establishment of validity evidence for construct invariance used three methods: hierarchical linear modeling (HLM), confirmatory factor analysis (CFA), and Rasch modeling through item response theory. These methods approached measurement invariance in three conceptually different ways. The HLM procedure determined the existence of measurement invariance at the test level using the variance components between the subgroups. The CFA determined the existence of measurement invariance at the subtest level using the variance-covariance matrix. The Rasch item analysis determined the existence of measurement invariance at the item level using the item difficulty estimates and fit statistics. In particular, the results of the study answered the following research questions:

Research questions

1. Are the group level coefficients from the hierarchical linear modeling technique invariant across administrative, gender, and ethnic subgroups?
2. Are the group measures produced through confirmatory factor analysis (i.e. factor structure, factor loadings, and error variances) invariant across administrative, gender, and ethnic subgroups?
3. Are the item difficulty estimates and fit statistics based on the Rasch item response theory model stable across administrative, gender, and ethnic subgroups?

Overall results

Overall, the reading test displayed invariance across administrative subgroups at the test, subtest, and item levels. The measure did not support invariance across gender and ethnicity subgroups at the test level and showed partial invariance at the subtest level There was support for invariance across administrative and gender subgroups at the item level. There was not enough evidence to support invariance of ethnicity at the item level.

Results for research question 1

The first research question used hierarchical linear modeling to determine measurement invariance across the subgroups at the test level. Hoffman, et al. (2000) suggested four key assumptions for hierarchical linear modeling (HLM). These assumptions were met for this data due to the 1) nested structure of the data (i.e., items, individuals, and groups); 2) the data at level 1, the items, were exposed to and influenced by individuals and their respective subgroups; 3) the outcome variable, Rasch logit score, was a combination of the item difficulty and person ability both of which were measured at the item level; and 4) the outcome variable, Rasch logit score, varied within the items, across individuals, and possibly across subgroups. Overall, the data met the key assumptions.

The five statistical assumptions for HLM suggested by Hoffman, et al. (2000) were also examined for this data. These included 1) independent level 1 residuals and level 1 residuals normally distributed with mean zero and variance σ^2 for every level 1 unit within each level 2 unit; 2) independence of level 1 predictors and level 1 residuals, 3) multivariate normal random errors at level 2 (each with a mean of zero and variance of τ_{qq}, and a covariance of $\tau_{qq'}$) and independence among level 2 units, 4) independence of the set of level 2 predictors and every level 2 residual, 5) residuals at level 1 and level 2 were also independent, and 6) the predictors at each level were not correlated with the random effects at the other level. Table 3 addresses the first statistical assumption via the presentation of the mean, standard deviation, skewness, and kurtosis of the level 1 residuals. The values of the skewness and kurtosis for each of the subgroups met those set by Finney and DiStefano (2006) to indicate normality. The skewness values were approximately -0.6 for each of the subgroups and the kurtosis values were 0.11 for the administrative and gender models and 0.25 for the ethnicity model. The mean of the level 1 residuals was zero for each of the subgroups and the standard deviation was approximately 0.17 for the administrative model and 0.16 for the gender and ethnicity models. The normal probability plots, (i.e., histogram, box plot, and Q-Q plot), from PROC UNIVARIATE in SAS showed the data were not significantly different from being normal for the level1 residuals for administration date, gender, and ethnicity.

Table 3.

Descriptive Statistics for Level 1 Residuals.

Level 1 Residuals	Mean	Standard Deviation	Skewness	Kurtosis
Administration Model	0	0.17	-0.63	0.11
Gender Model	0	0.16	-0.62	0.11
Ethnicity Model	0	0.16	-0.68	0.25

Note: N = 8,302

The second statistical assumption required the level 1 predictors, the items, to be independent of the level 1 residuals. From Table 4, the data for this study met this assumption. The correlation between the level 1 residuals and the level 1 predictors was zero as expected when two variables are independent. The first and second statistical assumptions presented by Hoffman et al. (2000) suggested the mean of the residuals was always zero and the correlation between the residuals and level 1 predictors was zero.

These assumptions were standard mathematical properties of residuals and the Pearson correlation in linear regression (Agresti & Finlay, 1997). From this data, the property was extended to hierarchical linear modeling.

Table 4.

Correlation between Level 1 Residuals and Level 1 Predictors by Subgroup.

Level 1 Residuals and Level 1 Predictors (Items)	Correlation
Administration Model	0
Gender Model	0
Ethnicity Model	0

The remaining statistical assumptions did not apply to this model because there were no level 2 or level 3 predictors. The third assumption examined the level 2 and level 3 random errors. In the fourth assumption, level 2 and level 3 predictors were examined. The fifth assumption explored the properties of the level 2 and level 3 residuals However, the random errors, predictors, and residuals did not exist for the third, fourth and fifth assumptions. In the sixth assumption, the model only had predictors at level 1 and no random errors at level 2 or level 3; thus, this assumption was not checked.

The mathematical model for the administrative, gender, and ethnicity groups were similar for levels 1, 2, and 3. The level 1 model represented item effects, the level 2 model represented individual or person effects and the subgroup effects were represented in the level 3 model, each of which are listed below:

Level 1 (item effects):

$$\log\left(\frac{p_{ijm}}{1-p_{ijm}}\right) = \beta_{01jm} + \beta_{1jm}X_{1ijm} + \beta_{2jm}X_{2ijm} + \ldots + \beta_{(k-I)jm}X_{(k-1)ijm},$$

where i (i = 1, …, k) represented items,

j (j=1, …, n) represented persons or individuals, *m* (m = 1, 2) represented gender or (m= 1, 2, 3) administrative and ethnic subgroups,

p_{ijm} was the probability that person *j* from subgroup *m* answers item *i* correctly,

X_{qijm} was defined as the *q*th dummy variable for item *i*, person *j*, from group *m* $(X = -1 \text{ when } q = i \text{ and } X = 0 \text{ when } q \neq i)$, for q = 1, …, *k*-1 for the *i*th item for person *j* in subgroup *m*,

β_{0jm} was the expected effect of the reference item for person *j* from subgroup *m*,

and β_{qjm} was the expected effect of the *q*th item for person *j* from subgroup *m* compared to the reference item.

Level 2 (individual effects):

$$\begin{aligned}
&\beta_{0jm} = \gamma_{00m} + r_{0jm} \\
&\beta_{1jm} = \gamma_{10m} \\
&\beta_{2jm} = \gamma_{20m} \\
&\ldots \\
&\beta_{(k-1)jm} = \gamma_{(k-1)0m}
\end{aligned}$$

where $r_{0jm} \sim N(r_{00m}, \tau_{\gamma})$ and τ_{γ} was assumed to be identical for all groups,

γ_{00m} was the item difficulty for the *k*th item, and

r_{0jm} was the random person ability and indicated how much person *j* from subgroup *m* is deviated from the mean, r_{00m}, within subgroup *m*.

At this level, the item parameters were fixed across person and varied across items, thus there was only one random term for person ability. Thus, there were no level 2 random errors or predictors.

Level 3 (subgroup effects):

$$\begin{aligned}&\gamma_{00m} = \pi_{000} + u_{00m}\\&\gamma_{10m} = \pi_{100}\\&\gamma_{20m} = \pi_{200}\\&\ldots\\&\gamma_{(k-1)0m} = \pi_{(k-1)0m}\end{aligned}$$

where $u_{00m} \sim N(0, \tau_{\pi})$,

π_{000} was the item difficulty for item *k*, the reference item, and

u_{00m} was the average ability of students in subgroup *m*,

The outcome variable for the models was the Rasch logit score, $\log\left(\frac{p_{ijm}}{1-p_{ijm}}\right)$, produced in SAS PROC PROBIT. The level 1 predictors were the dummy variables, X_{qijm}, for the *i*th item for person *j* in subgroup *m*; where *q* = 1 to 71 (one less than the number of items), *i* = 1 to 72 (the number of items), *j* = 1 to 8302 (the number of persons), and *m* = 1 to 3 (number of subgroups, except for the gender subgroups *m* = 1 to 2). In this matrix of dummy variables the value of the dummy variable was -1 when *i* = *q*, and 0 otherwise. Also, the last item of the test served as a reference item. For the second and third level, there were no predictors, the significance of the variance components at level 3 were used to determine invariance across administrative, gender, and ethnicity subgroups. The random effect of level 1 and level 2 was represented by r_{0jm}, which indicated how much person *j* from subgroup *m* is deviated from the mean of r_{00m} within subgroup *m*. The random effect associated with subgroup *m* was represented by u_{00m}, and represented the parameter by which measurement invariance was determined.

The estimation of the fixed effects for the administrative, gender, and ethnic subgroups are presented in Appendix G, H, and I, respectively. The fixed effects for the administrative, gender, and ethnicity models were not significant. The random effects were presented in Table 5. The level 3 random effect for administration date was not significant, suggesting there was no difference between the administrative dates (See Table 5). The level 3 random effects for gender and ethnicity were significant, suggesting there was a significant difference between male and females and there was a significant difference between Whites, Blacks, and Hispanics. However, the value of the variance component is close to zero, suggesting the difference is minimal (See Table 5).

Table 5.

Estimation of Random Effects for Administrative, Gender, and Ethnicity Models.

Random Effects	Variance Component (x 10^{-4})	*df*	χ^2	P-value
Administrative Date				
u_{00}	0	2	0.50	>.500
r_0	14	8299	227.99	>.500
Gender				
u_{00}	3.4	1	4.83	0.026
r_0	14	8300	223.67	>.500
Ethnicity				
u_{00}	9.4	2	9.73	0.008
r_0	14	8299	219.80	>.500

Results for research question 2

The assumptions for confirmatory factor analysis included multivariate normality of the item parcels and an omnibus test of the equivalence of the variance-covariance matrices. Table 6 gives the descriptive statistics and reliability for the item parcels. The skewness values ranged from -0.15 to -0.26 and the kurtosis values ranged from -0.85 to 0.11, thus meeting the criteria set by Finney and DiStefano (2006) was met. This confirmed the use of confirmatory factor analysis using maximum likelihood estimation, the most common form of estimation (Netemeyer, Bearden, & Sharma, 2003).

Table 6.

Descriptive Statistics for Item Parcels.

	N	Mean	Standard Deviation	Skewness	Kurtosis
Phonics	24	14.98	3.97	-0.15	0.11
Vocabulary Development	9	5.42	1.75	-0.19	-0.11
Reading Vocabulary	9	5.48	2.31	-0.26	-0.78
Reading Comprehension	30	17.30	6.38	-0.17	-0.85

Note: *N* = 8,302

The omnibus test of the homogeneity of the variance-covariance matrices suggested the groups were equivalent, $\Sigma_{Group1} = \Sigma_{Group2}$ (Box, 1949; Morrison, 1976). According to the results listed in Table 7; the variance-covariance matrices for the administrative groups were not significantly different. The variance-covariance matrices for gender and ethnicity were statistically significant. The p-value for the chi-square value was less than 0.001. The significance of the chi-square value suggested the variance-covariance matrices were different for the gender and ethnicity subgroups and the variance-covariance matrices were evaluated separately.

Table 7.

Omnibus Test for Homogeneity of Variance-Covariance Matrices by Subgroup.

	Chi-Square	*df*	p-value	Pooled
$\Sigma_{2005} = \Sigma_{2006} = \Sigma_{2007}$	5934.93	6162	0.98	Yes
$\Sigma_{Male} = \Sigma_{Female}$	5689.26	3160	<0.001	No
$\Sigma_{White} = \Sigma_{Black} = \Sigma_{Hispanic}$	8989.57	6320	<0.001	No

The Vandenberg and Lance (2000) sequence of steps were conducted for the gender and ethnicity subgroups to identify the source of greatest variance between the groups. The sequence of steps are below:

1. Freely estimate the baseline model
2. Constrain factor loadings to equality
3. Constrain error variances to equality
4. Constrain item means to equality (optional)
5. Constrain factor variances to equality
6. Constrain factor covariances to equality
7. Constrain factor means to equality (optional).

Steps 4 and 7 were not conducted and steps 5 and 6 were addressed with the omnibus test. Tables 8 and 9 present the results of steps 1, 2, and 3 for the gender and ethnicity subgroups, respectively. The first gender model was the baseline model and it sought to determine if configural invariance existed. In the baseline model, all of the parameters were freely estimated for males and females. In the model 2, all of the factor loadings were constrained to be equal to find evidence of metric invariance. This model was not significantly different from the baseline model $(\chi^2_{0.05}(3) = 7.815)$, thus allowing one to conclude, the model had metric invariance or the factor loadings were equivalent across gender. In the model 3, scalar invariance was examined by constraining the error variances to be equal. Given the model was significant when compared to the model with equal factor loadings $\left(\chi^2_{0.05}(4) = 9.488\right)$, the test for partial invariance began by allowing

Table 8.

Model testing for Measurement Invariance of Gender Subgroups.

Model	χ^2	df	RMSEA	$\Delta\chi^2$	Δdf	Significant
1. Baseline	29.59	4	0.04			
2. Equal factor loadings	35.11	7	0.03	5.52	3	No
3. Equal error variances	52.45	11	0.03	17.34	4	Yes
a. Listening free	38.87	10	0.03	3.76	3	No

the error variances to be freely estimated in a stepwise fashion. The modification indices were examined to determine which of the error variances when freely estimated for each group improved the model. The modification index for listening vocabulary was the highest, with a value of 13.51. Model 3a represents the model with the error variance of listening vocabulary freely estimated for males and females. This model was not significantly different from the model with equal factor loadings $\left(\chi^2_{0.05}(3) = 7.815\right)$. The error variance of listening vocabulary for males was different from the error variance for females. The model for gender subgroups had configural invariance, metric invariance and partial scalar invariance. Overall, the measure was partially invariant across gender subgroups and had good fit values. The final estimates of factor loadings and error variances are in Figure 3. The root mean squared error of approximation (RMSEA) was 0.03, which was less than 0.05 (Marsh, et al., 1988). The comparative fit index (CFI) and the non-normed fit index (NNFI) values for models 1 to 3a were 0.99, which were greater than the recommended value of 0.95 for good fit (Hu & Bentler, 1999). The value of the standardized root mean residual (SRMR)

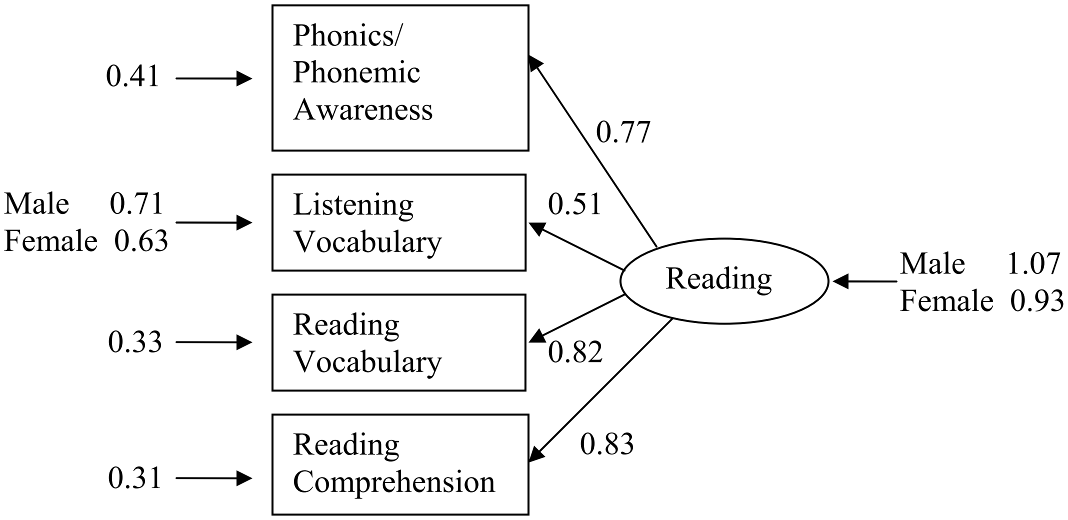

Figure 3. Standardized solution for the gender model.

for males was 0.01 and 0.02 for females, both of which were less than 0.05 as recommended by Hu and Bentler (1999). With the exception of the error variance for listening vocabulary and the variance of the latent variable, reading, the estimates were the same for males and females.

The results from the Vandenberg and Lance (2000) sequence of steps for the ethnicity sample are presented in Table 9. The baseline model suggested configural invariance existed for the White, Black, and Hispanic subgroups. The model to test for metric invariance was estimated and was not significant when compared to the baseline model, $\left(\chi^2_{0.05}(6) = 12.592\right)$. Next, the model with equal error variances was used to test for scalar invariance. This model was significant when compared to model 2, the model with equal factor loadings, $\left(\chi^2_{0.05}(8) = 15.507\right)$. Examination of the modification indices revealed the error variance of reading vocabulary be freely estimated to determine partial scalar invariance. The modification index for reading vocabulary was 31.87.

Table 9.

Model testing for Measurement Invariance of Ethnicity Subgroups.

Model	χ^2	df	RMSEA	$\Delta\chi^2$	Δdf	Significant
1. Baseline	31.88	6	0.04			
2. Equal factor loadings	35.27	12	0.03	3.39	6	No
3. Equal error variances	83.37	20	0.03	48.10	8	Yes
a. Vocabulary free	42.60	18	0.02	7.33	6	No

Model 3a, the model with equal error variances except reading vocabulary, was not significant when compared to the model with equal factor loadings $\left(\chi^2_{0.05}(6) = 12.592\right)$. The ethnicity model had configural invariance, metric invariance, and partial scalar invariance. The final estimates for the ethnicity model are in Figure 4. With the exception of the error variance of reading vocabulary and the variance of the latent variable, reading, the final estimates were the same for the White, Black, and Hispanic subgroups. Overall, the model was partially invariant across ethnic subgroups. The RMSEA value indicated good fit, 0.02 (Marsh, et al., 1988). The CFI and NNFI values for models 1 to 3a were 0.99, which indicated good fit (Hu & Bentler, 1999). The SRMR value was 0.02 for Whites, 0.01 for Blacks, and 0.03 for Hispanics, also indicating good fit (Hu & Bentler, 1999).

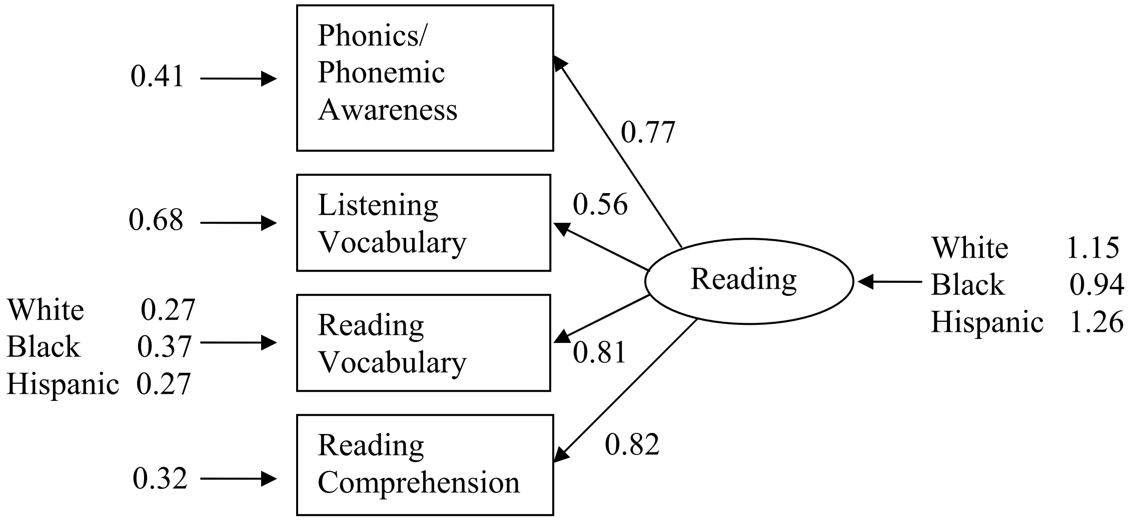

Figure 4. Standardized solution for the ethnicity model.

Results for research question 3

The Rasch item analysis process began with a check of the model assumptions. The first assumption examined the shape of the item characteristic curves (ICC), which were expected to be monotonically, increasing. Most of the items had a monotonically increasing item characteristic curve with a few anomalies. Figure 5 highlights the ICCs for the last 10 items, the remaining items on the test had similar ICCs. Overall, the data met the assumption for the item characteristic curves. However, the anomalies seemed to occur in the extreme scores where the sample size was small. Thus, the study proceeded with the use of Rasch item analysis with caution.

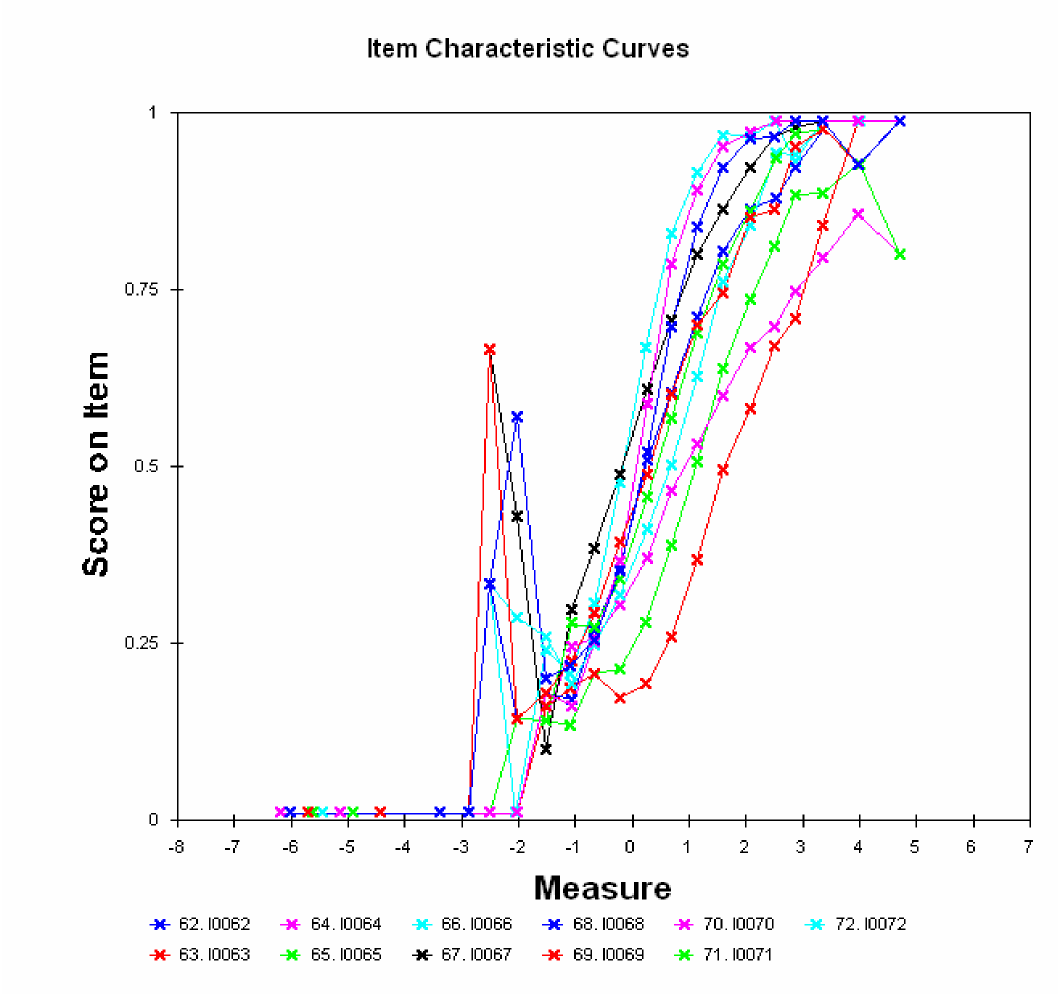

Figure 5. Empirical item characteristic curves for a select number of items.

Other assumptions for the use of Rasch item analysis were local independence, unidimensionality, and speededness. The local independence assumption was not checked for this data and the unidimensionality assumptions were confirmed through the configural invariance presented in research question 2 for each subgroup. Given the good model fit values presented in research question 2, it was assumed the model used for confirmatory factor analysis was unidimensional, with the single latent trait of reading. For the speededness assumption, approximately 32 respondents omitted questions near the end of the test across the three administrations. This represented less than 1% of the respondents, thus the speededness assumption was met.

Item calibration using Rasch analysis was conducted in Winsteps (Linacre, 2006) by subgroup: 2005, 2006, 2007, male, female, White, Black, and Hispanic. The mean, median, standard deviation and interquartile range for the item difficulty estimates, the standardized mean square infit statistics and the standardized mean square outfit statistics are in Table 10. These statistics were calculated for the whole group, administrative groups, gender groups, and ethnicity groups. The mean of the item difficulty estimates were set to zero for each of the groups for ease of interpretation. The mean of the standardized mean square infit statistics ranged from -0.38 to 0.05 and the mean of the standardized mean square outfit statistics ranged from -0.32 to 0.06. These values were within the productive values for measurement, given the theoretical mean was zero and standard deviation was one (Linacre & Wright, 1999). The median values for the item difficulty estimates for each group were greater than the mean; suggesting the data was left-skewed, or there were very few easy items for the examinees. The median values of the item difficulty estimates for each subgroup ranged from 0.08 to 0.21. The median values for the standardized mean square infit and outfit statistics were less than the mean values, suggesting the data was right-skewed, or there were outliers in the item infit and outfit. The median values of the standardized mean square infit statistics for the groups ranged from -1.44 to -0.37. The median values of the standardized mean square outfit statistics for the groups ranged from -2.01 to -0.40. The standard deviation and interquartile range for the item difficulty estimates, infit and outfit statistics are also presented in Table 10.

Table 10.

Descriptive Statistics for Item Difficulty Estimates, Infit, and Outfit Statistics by Group.

	Statistics			
Group	Mean	Median	Standard deviation	Interquartile range
Item Difficulty Estimates				
Combined	0	0.17	1.07	0.96
Administration				
2005	0	0.19	1.06	0.91
2006	0	0.16	1.07	0.88
2007	0	0.11	1.09	0.95
Gender				
Male	0	0.08	1.01	0.93
Female	0	0.21	1.16	0.91
Ethnicity				
White	0	0.11	1.13	1.07
Black	0	0.13	1.07	0.94
Hispanic	0	0.08	0.99	0.98
INFIT – ZSTD				
Combined	-0.38	-1.44	6.32	8.52
Administration				
2005	-0.15	-0.93	4.66	5.39
2006	-0.05	-0.71	4.35	5.61
2007	-0.06	-0.76	4.75	4.80
Gender				
Male	-0.26	-1.20	5.55	7.46
Female	-0.03	-0.80	5.12	5.42
Ethnicity				
White	0.05	-0.58	4.10	4.72
Black	-0.22	-1.15	5.74	7.95
Hispanic	-0.06	-0.37	1.82	2.13
OUTFIT - ZSTD				
Combined	-0.32	-2.01	6.61	9.67
Administration				
2005	0.01	-1.22	4.78	6.10
2006	0.06	-0.75	4.50	5.65
2007	0.01	-0.93	4.72	5.79
Gender				
Male	-0.03	-0.88	5.61	7.50
Female	-0.08	-1.37	5.25	6.69
Ethnicity				
White	-0.10	-1.22	4.13	5.45
Black	0.05	-1.47	6.10	8.52
Hispanic	0.01	-0.40	1.98	2.92

Note: The 16 persons were deleted from the Rasch item analysis for having an extreme raw score .

The results to determine invariance for the administrative, gender, and ethnic subgroups followed the sequence of steps outlined by Kim and Huynh (2009):

1. Correlation of item difficulty estimates and fit statistics
2. Mean differences in item difficultly estimates and fit statistics
3. Absolute differences in item difficulty estimates
4. Robust *Z* statistic for differences in item difficulty estimates and fit statistics.

Results for step 1: Correlation

The Pearson product moment correlation coefficients for each of the subgroups were strong and positive as expected. Table 11 lists the values of the correlation coefficient for the pairwise administrative groups, gender, and pairwise ethnicity groups. The values were strong and positive with a range from 0.75 to 0.99. The correlations for the administrative and gender subgroups were strong and positive, 0.9 and above. The correlations for the ethnic subgroups were also strong and positive, but not as strong as the administrative and gender subgroups. The lowest correlations were between the White-Hispanic subgroups on the infit and outfit statistics, with values of 0.77 and 0.75, respectively.

Table 11.

Correlation for Item Difficulty Estimates, Infit, and Outfit Statistics by Group.

		Difficulty	Infit - ZSTD	Outfit - ZSTD
Administration				
	2005/2006	0.998	0.957	0.953
	2005/2007	0.996	0.969	0.958
	2006/2007	0.996	0.967	0.969
Gender				
	Male/Female	0.983	0.927	0.939
Ethnicity				
	White/Black	0.977	0.912	0.885
	White/Hispanic	0.945	0.768	0.754
	Black/Hispanic	0.959	0.826	0.833

Note: The total sample size was 8,286 due to the removal of 16 persons with extreme scores.

Results for step 2: Mean differences

This step examined the mean of the differences in item difficulty estimates between the groups as well as the mean of the differences in standardized mean square infit statistics and standardized mean square outfit statistics. Table 12 lists the value of the mean pairwise differences for administration date, gender, and ethnicity.

The mean of the differences for the item difficulty estimates was zero for each group comparison. This result was expected given the mean of the item difficulty estimates was set to zero for each subgroup. The mean of the differences for the standardized mean square infit statistics ranged from -0.24 to 0.27. The mean of the differences for standardized mean square outfit statistics ranged from -0.15 to 0.05. Table 12 also includes the values for the median, standard deviation, and interquartile range. These values were used to calculate the Robust Z statistic in the fourth step.

Results for step 3: Absolute differences

In the second step, the summary statistics were presented for the item differences in difficulty and fit statistics. In the third step, each item was individually examined to reveal the absolute differences between the item difficulty estimates. The absolute differences of the item difficulty estimates in the third step was expected to be less than or equal to 0.3 in absolute value for items that were stable across groups (Huynh & Rawls, 2009). The percentage of items with an absolute difference greater than 0.3 in absolute value ranged from 0% to 26% for the administrative, gender, and ethnic pairwise differences (See Table 13).

Table 12.

Descriptive Statistics for the Pairwise Differences in Item Difficulty Estimates, Infit, and Outfit Statistics by Administration Date, Gender and Ethnicity.

		Mean	Median	Standard Deviation	Interquartile Range
Administration					
(2005 – 2006)	Difficulty	0.00	0.00	0.08	0.09
	Infit – ZSTD	-0.10	-0.12	1.35	1.70
	Outfit - ZSTD	-0.05	-0.08	1.45	2.10
(2005 - 2007)	Difficulty	0.00	0.00	0.10	0.11
	Infit-ZSTD	-0.09	-0.01	1.18	1.22
	Outfit -ZSTD	0.00	-0.02	1.38	1.47
(2006 - 2007)	Difficulty	0.00	0.00	0.10	0.13
	Infit-ZSTD	0.02	0.14	1.23	1.36
	Outfit -ZSTD	0.05	0.18	1.17	1.68
Gender					
(Male – Female)	Difficulty	0.00	-0.02	0.25	0.36
	Infit-ZSTD	-0.24	-0.41	2.08	2.19
	Outfit -ZSTD	0.04	0.00	1.94	2.57
Ethnicity					
(White – Black)	Difficulty	0.00	0.03	0.24	0.24
	Infit-ZSTD	0.27	0.25	2.62	3.25
	Outfit -ZSTD	-0.15	-0.32	3.11	4.29
(White – Hispanic)	Difficulty	0.00	0.05	0.38	0.32
	Infit-ZSTD	0.11	-0.02	2.94	2.67
	Outfit -ZSTD	-0.10	-0.63	2.94	3.22
(Black - Hispanic)	Difficulty	0.00	0.04	0.30	0.30
	Infit-ZSTD	-0.16	-0.61	4.36	6.49
	Outfit -ZSTD	0.05	-0.92	4.58	6.68

Note: The total sample size was 8,286 due to the removal of 16 persons with extreme scores.

Less than 20% of the items from the administrative and gender subgroups had an absolute difference that exceeded the expected value of 0.3. Comparing this result to the rule of thumb from the Rasch linking protocol, the construct was expected to have a similar meaning across groups (H. Huynh, personal communication, March 22, 2009). Eighteen percent of items for the White-Black comparison for absolute differences exceeded 0.3 in absolute value, the percentage for the White-Hispanic comparison and the Black-Hispanic comparison was 22% and 26%, respectively. The percentage of items with absolute differences greater than 0.3 exceeded the 20% rule of thumb from the empirical Rasch linking protocol (H. Huynh, personal communication, March 22, 2009). These results suggested there were a number of items that were not stable across the ethnicity groups in terms of item difficulty.

Results for Steps 4: Robust Z Statistics

The Robust Z statistics for the item difficulty estimates, standardized mean square infit statistics, and standardized mean square outfit statistics were expected to be less than or equal to 1.645 in absolute value. From Table 13, the percentage of items with a Robust Z value greater than 1.645 in absolute value ranged from 1% to 18% for item difficulty estimates, ranged from 3% to 28% for standardized mean square infit statistic, and ranged from 7% to 22% for the standardized mean square outfit statistic. The percentage of items with a Robust Z value greater than 1.645 in absolute value across administrative and gender groups ranged from 8% to 19%. This data met the 20% rule of thumb recommended by Huynh from Rasch linking protocol (personal communication, March 22, 2009). Therefore, it was inferred that stability existed across administrative and gender subgroups at the item level. The percentage of items across the White-Black and

Black-Hispanic comparison of Robust Z values greater than or equal to 1.645 in absolute value ranged from 3% to 18%. However, the White-Hispanic comparison of Robust Z values greater than 1.645 in absolute value ranged from 1% on the difficulty estimates to 22% and 28% on the fit statistics. Thus, the measure did not display stability across ethnicity at the item level.

Overall, at the test level, evidence from the hierarchical linear modeling procedure supported the comparison of the administrative subgroups. At the subtest level, confirmatory factor analysis showed the administrative groups were invariant based on the results of the omnibus test of homogeneity of variance-covariance. The gender and ethnic subgroups at the subtest level displayed configural invariance, metric invariance, and partial scalar invariance. All of the subgroups had excellent model fit. The administrative and gender subgroups met the rule of thumb for item level comparison. There was not enough evidence to conclude the test was invariant across ethnicity at the item level. Thus, the evidence supported the comparison of the administrative subgroups across all components; the gender subgroups across all components except listening vocabulary; and the ethnic subgroups across all components except reading vocabulary. Using the Rasch item analysis, the items were stable across administrative, and gender subgroups but lacked enough evidence to support stability across ethnicity.

Table 13.

Number and Percentage of Item Absolute Differences and Robust Z Statistics greater than or equal to Target Value by Administration Date, Gender, and Ethnicity.

	Absolute Differences	Robust Z		
	Difficulty	Difficulty	Infit-ZSTD	Outfit-ZSTD
Administration				
2005 - 2006				
N	0	7	11	8
Percentage	0%	10%	15%	11%
2005 - 2007				
N	0	12	11	11
Percentage	0%	17%	15%	15%
2006 - 2007				
N	1	8	9	6
Percentage	1%	11%	13%	8%
Gender				
Male - Female				
N	14	6	14	8
Percentage	19%	8%	19%	11%
Ethnicity				
White - Black				
N	13	13	10	7
Percentage	18%	18%	14%	10%
White - Hispanic				
N	16	1	20	16
Percentage	22%	1%	28%	22%
Black - Hispanic				
N	19	9	2	5
Percentage	26%	13%	3%	7%

Note: Items in this table have absolute differences greater than 0.3 in absolute value or robust Z statistics greater than or equal to 1.645 in absolute value.

Chapter 5

Discussion and Conclusions

Test validity is an important concept for which evidence produced supports the inferences and actions taken based on a persons' performance. Policymakers and educators desire to make decisions that will not adversely affect any particular gender, ethnicity, level of English proficiency, socioeconomic status, and disability status. In an effort to ensure the achievement gap is narrowing; the achievement data is disaggregated by these subgroups. This validity evidence is provided by demonstrating the measure is invariant across groups. Evidence of measurement invariance supports the interpretations and score uses at the disaggregated level.

Several methods for detecting measurement invariance were reviewed and three of these methods were applied to a reading test for administrative, gender, and ethnic subgroups. The study demonstrated the use of hierarchical linear modeling (HLM), confirmatory factor analysis (CFA), and Rasch item analysis in the detection of measurement invariance across subgroups at the test, subtest, and item levels, respectively. Specifically, the study used HLM to produce validity evidence of interpretations across subgroups at the test level. These test level analyses, used by policymakers for decision making; either supported or refuted the comparisons of subgroups on the overall score. The study used confirmatory factor analysis to provide evidence of the validity of interpretations across subgroups at the subtest level. Subtest level analyses, used for decision making by instructional designers, provided validity evidence to support the comparison of subgroups on various subtests. The study used Rasch item analysis to provide validity of interpretations across subgroups at the item level. Item level analyses, used in psychometric decision making, provided validity evidence to support the comparison of subgroups on the items. The results of the study produced strong validity evidence for the comparison of administrative subgroups at the test, subtest, and item levels. There was no validity evidence produced to support the comparison of gender and ethnicity at the test level. Weak validity evidence, or partial scalar invariance, was produced to support the comparison of gender and ethnicity at the subtest levels. Some validity evidence was produced to support interpretations or comparison of males and females at the item level. There was no validity evidence produced to support comparison of ethnic subgroups at the item level.

To what can this weak evidence or lack thereof across gender and ethnicity at the test, subtest, and item levels be contributed? One characteristic of the study to which the production of weak evidence or the lack thereof was contributed to the use of one test to determine if the subgroups were comparable in their level of reading acquisition. A single measure was not an accurate depiction the full range of reading acquisition skills a student may possess. It was known from previous literature that the construct of reading acquisition was very complex and had many components, only four of which were measured on this assessment, thus the use of this assessment alone was not sufficient to make inferences about a students' reading ability (Adams, 1990; Byrne, Brian, 1992; Chall, 1996; Dally, 2006; Flippo, 2001; Nation, 2008; National Institute of Child Health and Human Development, 2000; Snow, et al., 1998; Taylor & Pearson, 2002). Thus, there was a need to increase the validity evidence via analysis of content analysis via sensitivity review, concurrent validity studies, predictive validity studies, external bias studies, and impact studies. Many of the students in our nation were given only one reading test to provide evidence of progress or achievement. States such as Arkansas, Arizona, and Tennessee, give students additional tests or use the students' past performance on other assessments to improve their inferences about the level of improvement in the student knowledge and skills (United States Department of Education, 2007). Overall, the there was not enough evidence to support the decisions about the class in which a child may enroll or the reading group to which the student maybe assigned based on one assessment.

The second characteristic of the study that contributed to the production of weak evidence or the lack thereof was the structure of the subgroups. From previous research in differential item functioning, "subgroups are not expected to perform equally well [because they] differ in their experiences, interests, and motivations. Consequently, only groups formed by random assignment should be expected to perform equally well (Drasgow, et al., 2006). None of the subgroups in this study were created by random assignment; schools and students were chosen by application to participate in the program. Given the established expectation, why are comparisons made and inferences drawn, and actions taken based on the gender and ethnicity subgroups, groups not formed by random assignment? The disaggregated data was a strategy to help close the achievement gap, but is it valid for our federal accountability laws and regulations to demand our administrators, teachers, students, and parents to disaggregate assessment data? Were the results of all assessments intended for disaggregation? Is the strategy of disaggregating the data valid for all assessments?

The final characteristic of the study to which the production of weak evidence or the lack thereof is contributed was the detailed level at which invariance was studied. Although it contributed to the weak invariance or the lack thereof, the level of detail also served as a unique aspect of the study. It can be noted from the invariance results of administrative groups and partial invariance of the gender and ethnic subgroups at the subtest level that test developers and psychometricians addressed invariance at some levels. The results of the confirmatory factor analysis suggested that test developers and psychometricians may have completed similar analyses during the development stages of the assessment. Including this type of analysis in the test development process supported the level of inferences and actions taken at the subtest level. However, it was noted less attention was received for item level comparisons in the development process. The steps outlined by Kim and Huynh (2009) go beyond the test and subtest level to compare of mean item difficulty across groups, the other unique comparison across groups was that of the fit statistics. The Kim and Huynh (2009) methods identified a significant percentage of items with extreme absolute differences and extreme Robust Z statistics that may not be noticed when comparisons were examined across test or subtest level. However, the 1999 *Standards for Educational and Psychological Testing* suggested "…differential item functioning is not always a flaw or weakness. Subsets of items that have specific characteristics in common (e.g., specific content, task representation) may function differently for different groups of similarly scoring examinees. This indicates a kind of multidimensionality that may be unexpected or may conform to the test framework (American Educational Research Association, et al., 1999).

Future Research

Hierarchical linear modeling, confirmatory factor analysis, and Rasch item analysis methods are effective when used in combination to detect measurement invariance across subgroups at the test, subtest, and item level. Future studies, empirical or simulated, may show further strength of the combination of methods to detect measurement invariance across subgroups at the test, subtest, and item levels. Simulation studies will be particularly helpful given their ability to vary the type of student in the sample, the sample sizes, the test content, and subgroup structure. The level of student ability in this study was not representative of all students seeking reading acquisition skills given the nature of the Reading First Initiative. Students in this study are served due to the low performance of the schools or demonstrated financial need. The instrument in this study addresses four components of reading acquisition for young children. Other content areas, such as mathematics, and science, may be explored as well as a wider range of age groups in these areas. Simulation studies also allow for comparison of results with different sample sizes. The proportion of students across administrative and gender subgroups in this study are relatively equal. However, the Black students significantly outnumber the White students and there are even fewer Hispanic students. How would the results change if there were more White students than Black students or if there were more Hispanic students than White and Black students? How would the results change if each of the groups were equally represented? An examination of more empirical data or a simulation study may strengthen the use of these methodologies. Upon the examination of results of the various studies, the effectiveness of the HLM, CFA, and Rasch item analysis methods can be examined to determine the effectiveness of the combination of these procedures for producing construct validity evidence at the test, subtest, and item levels through measurement invariance.

References

Ackerman, T. A. (1992). A didactic explanation of item bias, item impact, and item validity from a multidimensional perspective. *Journal of Educational Measurement, 29*(1), 67-91.

Adams, M. J. (1990). *Beginning to read : Thinking and learning about print*. Cambridge, MA: MIT Press.

Agresti, A., & Finlay, B. (1997). *Statistical methods for social sciences* (3rd ed.). Upper Saddle River, NJ: Prentice Hall, Inc.

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

Anastasi, A. (1986). Evolving concepts of test validation. *Annual Review of Psychology, 37*, 1-15.

Angoff, W. (1988). Validity: an evolving concept. In H. Wainer & H. I. Braun (Eds.), *Test Validity* (pp. 19 - 32). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Angoff, W. (1993). Perspectives on DIF methodology. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Anguiano, R. P. V. (2004). Families and schools: The effect of parental involvement on high school completion. *Journal of Family Issues, 25*(1), 61-85.

Atar, B. (2007). *Differential item functioning analyses for mixed response data using IRT likelihood-ratio test, logistic regression, and GLLAMM procedures.* Unpublished Ph.D., The Florida State University, United States -- Florida.

Biemiller, A., & Boote, C. (2006). An effective method for building meaning vocabulary in primary grades. *Journal of Educational Psychology, 98*(1), 44-62.

Blachman, B. A., Schatschneider, C., Fletcher, J. M., Francis, D. J., Clonan, S. M., Shaywitz, B. A., et al. (2004). Effects of intensive reading remediation for second and third graders and a 1-year follow-up. *Journal of Educational Psychology, 96*(3), 444-461.

Bowden, S. C., Gregg, N., Bandalos, D. L., Davis, M., Coleman, C., Holdnack, J. A., et al. (2008). Latent mean and covariance differences with measurement equivalence in college students with developmental difficulties versus the Wechsler Adult Intelligence Scale-III/Wechsler Memory Scale-III normative sample. *Educational and Psychological Measurement, 68*(621 - 642).

Bowey, J. A. (1995). Socioeconomic status differences in preschool phonological sensitivity and first-grade reading achievement. *Journal of Educational Psychology, 87*(3), 476-487.

Box, G. E. P. (1949). A general distribution theory for a class of likelihood criteria. *Biometrika, 36*(3/4), 317-346.

Buss, A. R., & Royce, J. R. (1975). Detecting cross-cultural commonalities and differences: Intergroup factor analysis. *Psychological Bulletin, 82*(1), 128-136.

Butler, S. R., Marsh, H. W., Sheppard, M. J., & Sheppard, J. L. (1985). Seven-year longitudinal study of the early prediction of reading achievement. *Journal of Educational Psychology, 77*(3), 349-361.

Byrk, A. S., & Raudenbush, S. W. (1992). *Hierarchical linear models: Applications and data analysis methods*. Newbury Park, CA: Sage.

Byrne, B. (1992). Studies in the aquisition procedure for reading: Rationale, hypotheses and data. In P. B. Gough, L. Ehri, C. & R. Treiman (Eds.), *Reading Acquisition*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Byrne, B. (1998). *Structural equation modeling with LISREL, PRELIS, and SIMPLIS: basic concepts, applications, and programming*. Mahwah, NJ: Lawrence Erlbaum.

Byrne, B., Baron, P., & Balev, J. (1996). The Beck Depression Inventory: Testing for its factorial validity and invariance across gender for Bulgarian non-clinical adolescents. *Personality and Individual Differences, 21*(5), 641-651.

Campbell, H. L., Barry, C. L., Joe, J. N., & Finney, S. J. (2008). Configural, metric, and scalar invariance of the Modified Achievement Goal Questionnaire across African American and White university students. *Educational and Psychological Measurement, 68*(6), 988 - 1007.

Carle, A. C., Millsap, R. E., & Cole, D. A. (2008). Measurement bias across gender on the Children's Depression Inventory: Evidence for invariance from two latent variable models. *Educational and Psychological Measurement, 68*(2), 281-303.

Catts, H. W. (2001). Estimating the risk of future reading difficulties in kindergarten children: A research-based model and its clinical implementation. *Language, Speech, & Hearing Services in Schools, 32*(1), 38-50.

Catts, H. W., Fey, M. E., Zhang, X., & Tomblin, J. B. (1999). Language basis of reading and reading disabilities: Evidence from a longitudinal investigation. *Scientific Studies of Reading, 3*(4), 331-361.

Cauffman, E., & MacIntosh, R. (2006). A Rasch differential item functioning analysis of the Massachusetts Youth Screening Instrument: Identifying race and gender differential item functioning among juvenile offenders. *Educational and Psychological Measurement, 66*(3), 502-521.

Chall, J. S. (1996). *Learning to read : the great debate* (3rd ed.). Fort Worth, TX: Harcourt Brace College Publishers.

Chan, D. (2000). Detection of differential item functioning on the Kirton Adaption-Innovation Inventory using multiple-group mean and covariance structure analyses. *Multivariate Behavioral Research, 35*(2), 169 - 199.

Chatterji, M. (2006). Reading achievement gaps, correlates, and moderators of early reading achievement: Evidence from the Early Childhood Longitudinal Study (ECLS) kindergarten to first grade sample. *Journal of Educational Psychology, 98*(3), 489-507.

Cheung, G. W., & Rensvold, R. B. (1999). Testing factorial invariance across groups: a reconceptualization and proposed new method.(includes appendices). *Journal of Management, 25*(1), 1(2).

Cheung, G. W., & Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. [Article]. *Structural Equation Modeling, 9*, 233-255.

Chou, C., Bentler, P., & Pentz, M. A. (1998). Comparisons of two statistical approaches to study growth curves: the multilevel model and the latent curve analysis. *Structural Equation Modeling, 5*, 247 - 266.

Cizek, G. J., Rosenberg, S. L., & Koons, H. H. (2008). Sources of validity evidence for educational and psychological tests. *Educational and Psychological Measurement, 68*(3), 397-412.

Connor, C. M., & Craig, H. K. (2006). African American preschoolers' language, emergent literacy skills, and use of African American English: A complex relation. *Journal of Speech, Language, and Hearing Research, 49*(4), 771-792.

Cook, L. L., Eignor, D. R., & Taft, H. L. (1988). A comparative study of the effects of recency of instruction on the stability of IRT and conventional item parameter estimates. *Journal of Educational Measurement, 25*(1), 31-45.

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. Belmont, CA: Wadsworth Group.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16*(3), 297 - 334.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52*(4), 281-302.

Cunningham, A. E., & Stanovich, K. E. (1997). Early reading acquisition and its relation to reading experience and ability 10 years later. *Developmental Psychology, 33*(6), 934-945.

Cureton, E. E. (1951). Validity. In E. F. Lindquist (Ed.), *Educational measurement* (pp. 621 - 694). Washington, DC: American Council on Education.

Dally, K. (2006). The influence of phonological processing and inattentive behavior on reading acquisition. *Journal of Educational Psychology, 98*(2), 420-437.

Dancer, L. S., Anderson, A. J., & Derlin, R. L. (1994). Use of log-linear models for assessing differential item functioning in a measure of psychological functioning. (Special Section: New Methods in Mental Health Research). *Journal of Consulting and Clinical Psychology, v62*(n4), p710(718).

Darling-Hammond, L. (1998). Unequal opportunity: Race and education. (Black America: Progress & Prospects). *Brookings Review, v16*(n2), p28-32.

DeMars, C. (2001). Group differences based on IRT scores: Does the model matter? *Educational and Psychological Measurement, 61*(1), 60-70.

Dickenson, T., Habing, B., Rawls, A., & Johnson, R. L. (2008). *Vertical scaling and dimensionality of a primary reading assessment*. Paper presented at the the annual meeting of the National Council on Measurement in Education, New York, NY.

Dorans, N. J., & Holland, P. W. (1993). DIF detection and description: Mantel-Haenszel and standardization. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 35 - 66). Hillsdale, NJ: Lawrence Erlbaum Associates.

Downey, C. J., Steffy, B. E., Poston, W. K., & English, F. W. (2009). *50 ways to close the achievement gap* (3rd ed.). Johnston, IA: Curriculum Management Systems, Inc.

Draper, D. (1995). Inference and hierarchical modeling in the social sciences. *Journal of Educational and Behavioral Statistics, 20*(2), 115-147.

Drasgow, F. (1982). Biased test items and differential validity. *Psychological Bulletin, 92*(2), 526-531.

Drasgow, F. (1987). Study of the measurement bias of two standardized psychological tests. *Journal of Applied Psychology, 72*(1), 19-29.

Drasgow, F., & Kanfer, R. (1985). Equivalence of psychological measurement in heterogeneous populations. *Journal of Applied Psychology, 70*(4), 662-680.

Drasgow, F., Luecht, R., & Bennett, R. (2006). Technology and testing. In R. Brennan (Ed.), *Educational Measurement* (4th ed., pp. 471 - 515). Westport, CT: Praeger Publishers.

Embretson, S. (1983). Construct validity: Construct representation versus nomothetic span. *Psychological Bulletin, 93*(1), 179-197.

Embretson, S., & Reise, S. P. (2000). *Item response theory for psychologists*. Mahwah, NJ: Lawrence Erlbaum.

Fan, X. (1998). Item response theory and classical test theory: an empirical comparison of their item/person statistics. *Educational and Psychological Measurement, 58*(3), p357.

Ferrando, P. J., & Lorenzo-Seva, U. (2005). IRT-related factor analytic procedures for testing the equivalence of paper-and-pencil and internet-administered questionnaires. *Psychological Methods, 10*(2), 193-205.

Ferrara, S., & DeMauro, G. (2006). Standardized assessment of individual achievement in K-12. In R. Brennan (Ed.), *Educational Measurement* (4th ed., pp. 579 - 621). Westport, CT: Praeger Publishers.

Fidalgo, A. M., Hashimoto, K., Bartram, D., & Muniz, J. (2007). Empirical Bayes versus standard Mantel-Haenszel statistics for detecting differential item functioning under small sample conditions. *The Journal of Experimental Education, 75*(4), 293-315.

Finney, S. J., & DiStefano, C. (2006). Dealing with nonnormal and categorical data in structural equation models. In G. R. Hancock & R. O. Mueller (Eds.), *Structural equation modeling: A second course.* Greenwich, CT: Information Age.

Flippo, R. F. (Ed.). (2001). *Reading researchers in search of common ground*. Newark, DE: International Reading Association.

Florida Department of Education (2008). Student performance report: State reading demographic report Retrieved January 10, 2009, from https://app1.fldoe.org/FCATDemographics/Selections.aspx?level=State&subj=Reading

Foster, W. A., & Miller, M. (2007). Development of the literacy achievement gap: a longitudinal study of kindergarten through third grade. *Lang Speech Hear Serv Sch, 38*(3), 173-181.

Gamse, B. C., Bloom, H. S., Kemple, J. J., Jacob, R. T., & Institute of Education Sciences (2008). *Reading First Impact Study: Interim Report. NCEE 2008-4016*.

Washington, DC: National Center for Education Evaluation and Regional Assistance.

Goodwin, L. D., & Leech, N. L. (2003). The meaning of validity in the new standards for educational and psychological testing: implications for measurement courses.(Assessment In Action). *Measurement and Evaluation in Counseling and Development, 36*(3), 181 - 191.

Graham, J. M. (2006). Congeneric and (essentially) tau-equivalent estimates of score reliability: What they are and how to use them. *Educational and Psychological Measurement, 66*(6), 930 - 944.

Guion, R. M. (1977). Content Validity--The Source of My Discontent. *Applied Psychological Measurement, 1*(1), 1-10.

Guion, R. M. (1980). On trinitarian doctrines of validity. *Professional Psychology, 11*(3), 385-398.

Gustavsson, J. P., Eriksson, A.-K., Hilding, A., Gunnarsson, M., & Ã–Stensson, C.-G. R. (2008). Measurement invariance of personality traits from a five-factor model perspective: multi-group confirmatory factor analyses of the HP5 inventory. [Article]. *Scandinavian Journal of Psychology, 49*, 459-467.

Guttmannova, K., Szanyi, J. M., & Cali, P. W. (2008). Internalizing and externalizing behavior problem scores: Cross-ethnic and longitudinal measurement invariance of the behavior problem index. *Educational and Psychological Measurement, 68*(4), 676-694.

Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: Principles and applications*. Boston/Dordrecht/Lancaster: Kluwer Nijhoff Publishing a member of Kluwer Academic Publishers Group.

Hofmann, D. A., Griffin, M. A., & Gavin, M. A. (2000). The application of hierarchical linear modeling to organizational research. In K. J. Klein & S. W. J. Kozlowski (Eds.), *Multilevel theory, research, and methods in organizations: foundations, extensions and new directions*. San Francisco, CA: Jossey-Bass Inc.

Holland, P. W., & Thayer, D. T. (1988). Differential item performance and the Mantel-Haenszel procedure. In H. Wainer & H. I. Braun (Eds.), *Test Validity* (pp. 129 - 146). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Hong, S., Malik, M. L., & Lee, M.-K. (2003). Testing configural, metric, scalar, and latent mean invariance across genders in sociotropy and autonomy using a non-western sample. *Educational and Psychological Measurement, 63*(4), 636-654.

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: conventional criteria versus. *Structural Equation Modeling, 6*(1), 1.

Huynh, H., Gleaton, J., & Seaman, S. P. (1992). *Technical documentation for the South Carolina high school exit examination of reading and mathematics: Paper No. 2 (2nd ed.)*. Columbia, SC: University of South Carolina, College of Education.

Huynh, H., & Rawls, A. (2009). A comparison between robust z and 0.3-logit difference procedures in assessing stability of linking items for the Rasch model. In J. E. V. Smith & G. E. Stone (Eds.), *Applications of Rasch measurement in criterion-referenced testing: Practice analysis to score reporting.* Maple Grove, MN: JAM Press.

Innabi, H., & Dodeen, H. (2006). Content analysis of gender-related differential item functioning TIMSS items in mathematics in Jordan. (Trends in International Mathematics and Science Study). *School Science and Mathematics, 106*(8), 328-337.

International Reading Association (2001). Latest NAEP sees little change in past eight years. *Reading Today,* p. 6,

Jöreskog, K. (1969). A general approach to confirmatory maximum likelihood factor analysis. *Psychometrika, 34*(2), 183-202.

Jöreskog, K. (1971). Simultaneous factor analysis in several populations. *Psychometrika, 36*(4), 409-426.

Jöreskog, K. (1978). Structural analysis of covariance and correlation matrices. *Psychometrika, 43*(4), 443-477.

Joreskog, K., & Sorbom, D. (2001). Lisrel 8.5. Lincolnwood, IL: Scientific Software International.

Kamata, A. (2001). Item analysis by the hierarchical generalized linear model. *Journal of Educational Measurement, 38*(1), 79-93.

Kane, M. (2006). Validation. In R. Brennan (Ed.), *Educational Measurement* (4th ed., pp. 17 - 64). Westport, CT: Praeger Publishers.

Karazsia, B., van Dulmen, M., & Wildman, B. (2008). Confirmatory factor analysis of Arnold et al.'s Parenting Scale across race, age, and sex. *Journal of Child and Family Studies, 17*(4), 500-516.

Katsiyannis, A., Zhang, D., Ryan, J. B., & Jones, J. (2007). High-stakes testing and students with disabilities: Challenges and promises.(Report). *Journal of Disability Policy Studies, 18*(3), 160(168).

Kieffer, M. J. (2008). Catching up or falling behind? Initial English proficiency, concentrated poverty, and the reading growth of language minority learners in the United States. *Journal of Educational Psychology, 100*(4), 851-868.

Kim, D., & Huynh, H. (2008). Computer-based and paper-and-pencil administration mode effects on a statewide end-of-course english test. *Educational and Psychological Measurement, 68*(4), 554-570.

Kim, D., & Huynh, H. (2009). Transitioning from paper-and-pencil to computer-based testings: Examining stability of Rasch latent trait across gender and ethnicity. In J. E. V. Smith & G. E. Stone (Eds.), *Applications of Rasch measurement in criterion-referenced testing: Practice analysis to score reporting*. Maple Grove, MN: JAM Press.

Kim, K. H., Cramond, B., & Bandalos, D. L. (2006). The latent structure and measurement invariance of scores on the Torrance Tests of Creative Thinking-Figural. *Educational and Psychological Measurement, 66*(3), 459-477.

Koretz, D., & Hamilton, L. (2006). Testing for accountability in K-12. In R. Brennan (Ed.), *Educational Measurement* (4th ed., pp. 531 - 578). Westport, CT: Praeger Publishers.

Kristjansson, E., Aylesworth, R., & McDowell, I. (2005). A comparison of four methods for detecting differential item functioning in ordered response items. *Educational and Psychological Measurement, 65*(6), 935-953.

Kubiszyn, T. W., Meyer, G. J., Finn, S. E., Eyde, L. D., Kay, G. G., Moreland, K. L., et al. (2000). Empirical support for psychological assessment in clinical health care settings. *Professional Psychology: Research and Practice, 31*(2), 119-130.

Lederman, L. M., & Burnstein, R. A. (2006). Alternative approaches to high-stakes testing: Mr. Lederman and Mr. Burnstein propose a novel way to increase student engagement and counter the pressures of high-stakes testing. *Phi Delta Kappan, 87*(6), 429.

Leeuw, J. d., & Kreft, I. G. G. (1995). Questioning Multilevel Models. *Journal of Educational and Behavioral Statistics, 20*(2), 171-189.

Lievens, F., Anseel, F., Harris, M. M., & Eisenberg, J. (2007). Measurement invariance of the pay satisfaction questionnaire across three countries. *Educational and Psychological Measurement, 67*(6), 1042-1051.

Linacre, J. (2006). WINSTEPS Rasch measurement (Version 3.63.2). Chicago: Author.

Linacre, J., & Wright, B. (1999). *A user's guide to WINSTEPS MINISTEP: Rasch-model computer programs*. Chicago, IL: MESA Press.

Little, T. D. (1997). Mean and covariance structures (MACS) analyses of cross-cultural data: Practical and theoretical issues. *Multivariate Behavioral Research, 32*(1), 53.

Little, T. D., Cunningham, W. A., Shahar, G., & Widaman, K. F. (2002). To parcel or not to parcel: Exploring the question, weighing the merits. *Structural Equation Modeling, 9*(2), 151-173.

Lord, F. M. (1980). *Applications of item response to theory to practical testing problems*. Hillsdale, N.J.: Erlbaum Associates.

Lubke, G. H., Dolan, C. V., Kelderman, H., & Mellenbergh, G. J. (2003). On the relationship between sources of within- and between-group differences and measurement invariance in the common factor model. *Intelligence, 31*(6), 543-566.

Maller, S. J., & French, B. F. (2004). Universal nonverbal intelligence test factor invariance across deaf and standardization samples. *Educational and Psychological Measurement, 64*(4), 647-660.

Mantel, N., & Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute, 22*, 719 - 748.

Marsh, H. W., Balla, J. R., & McDonald, R. P. (1988). Goodness-of-fit indexes in confirmatory factor analysis: The effect of sample size. *Psychological Bulletin, 103*(3), 391-410.

Marsh, H. W., Hau, K., & Balla, J. R. (1995). *Is more ever to much: The number of indicators per factor in confirmatory factor analysis*: Educational Resources Information Center (ERIC Document Reproduction Service No. ED401329).

Marsh, H. W., & Hocevar, D. (1988). A new, more powerful approach to multitrait-multimethod analyses: Application of second-order confirmatory factor analysis. *Journal of Applied Psychology, 73*(1), 107-117.

McCoach, D. B., O'Connell, A. A., Reis, S. M., & Levitt, H. A. (2006). Growing readers: A hierarchical linear model of children's reading growth during the first 2 years of school. *Journal of Educational Psychology, 98*(1), 14-28.

McDonald, R. (1982). Linear versus models in item response theory. *Applied Psychological Measurement, 6*(4), 379-396.

Meece, J. L., & Kurtz-Costes, B. (2001). Introduction: The schooling of ethnic minority children and youth. *Educational Psychologist, 36*(1), 1 - 7.

Mellenbergh, G. J. (1989). Item bias and item response theory. *International Journal of Educational Research, 13*(2), 127-143.

Meredith, W. (1993). Measurement invariance, factor analysis and factorial invariance. *Psychometrika, 58*(4), 525-543.

Messick, S. (1980). Test validity and the ethics of assessment. *American Psychologist, 35*(11), 1012-1027.

Messick, S. (1988). The once and future issues of validity. Assessing the meaning and consequences of measurement. In H. Wainer & H. I. Braun (Eds.), *Test validity* (pp. 33 - 45). Hillsdale, N.J. :: Lawrence Erlbaum Associates.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational Measurement* (3rd ed., pp. 13 - 103). New York: Macmillan Publishing Company.

Meyer, G. J., Finn, S. E., Eyde, L. D., Kay, G. G., Moreland, K. L., Dies, R. R., et al. (2001). Psychological testing and psychological assessment: A review of evidence and issues. *American Psychologist, 56*(2), 128-165.

Mikulay, S. M., & Goffin, R. D. (1998). Measuring and predicting counterproductivity in the laboratory using integrity and personality testing. *Educational and Psychological Measurement, 58*(5), 768 - 791.

Miller, M. D., & Linn, R. L. (1988). Invariance of item characteristic functions with variations in instructional coverage. *Journal of Educational Measurement, 25*(3), 205-219.

Millsap, R. E. (1995). Measurement invariance, predictive invariance, and the duality paradox. *Multivariate Behavioral Research, 30*(4), 577.

Millsap, R. E. (1997). Invariance in measurement and prediction: Their relationship in the single-factor case. *Psychological Methods, 2*(3), 248-260.

Millsap, R. E. (2007). Invariance in measurement and prediction revisited. *Psychometrika, 72*(4), 461-473.

Millsap, R. E., & Everson, H. T. (1993). Methodology review: Statistical approaches for assessing measurement bias. *Applied Psychological Measurement, 17*(4), 297-334.

Mislevy, R. J. (1983). Item response models for grouped data. *Journal of Educational Statistics, 8*(4), 271-288.

Montgomery, J. K., & Hayes, L. L. (2005). Literacy transition strategies for upper elementary students with language-learning disabilities. *Communication Disorders Quarterly, 26*(2), 85(89).

Morris, C. N. (1995). Hierarchical models for educational data: An overview. *Journal of Educational and Behavioral Statistics, 20*(2), 190-200.

Morrison, D. F. (1976). *Multivariate statistical methods* (2nd ed.). New York, NY: McGraw-Hill.

Mullen, M. R. (1995). Diagnosing measurement equivalence in cross-national research. *Journal of International Business Studies, 26*(3), 573-596.

Nation, K. (2008). Learning to read words. *The Quarterly Journal of Experimental Psychology, 61*(8), 1121 - 1133.

National Forum on Education Statistics Race/Ethnicity Data Implementation Task Force (2008). Managing an identity crisis: Forum guide to implementing new federal race and ethnicity categories. (NFES 2008-802). Washington, DC: National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education.

National Institute of Child Health and Human Development (2000). *Teaching children to read: An evidence-based assessment of the scientific research literature on reading and its implications for reading instruction* (No. NIH Publication No. 00-4769). Washington, DC: U.S. Government Printing Office.

Netemeyer, R., Bearden, W., & Sharma, S. (2003). *Scaling procedures: Issues and applications*. Thousand Oaks, CA: Sage Publications, Inc.

No Child Left Behind Act (Reauthorization of Elementary and Secondary Education Act) Public Law 107-110 § Section 1202(c)(7)(A)(IV)(2) (2002).

North Carolina Department of Public Instruction (2008). Reports of disaggregated state, school system (LEA), and school performance data for 2006-2008. Retrieved January 10, 2009, from http://disag.ncpublicschools.org/2008/

Nunnally, J. (1978). *Psychometric theory.* (2nd ed.). New York: McGraw-Hill.

Pellegrino, J. W. (1988). Mental models and mental tests. In H. Wainer & H. I. Braun (Eds.), *Test Validity* (pp. 49 - 60). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Rasch, G. (1960/1980). *Probabilistic models for some intelligence and attainment tests (Copenhagen, Danish Institute for Educational Research), expanded edition (1980) with foreword and afterword by B.D. Wright*. Chicago: University of Chicago Press.

Raudenbush, S. W., & Bryk, A. (2002). *Hierarchical linear models* (2nd ed.). Thousand Oaks, CA: Sage Publications.

Raudenbush, S. W., Bryk, A., & Congdon, R. (2008). HLM 6.

Raykov, T. (1997). Estimation of composite reliability for congeneric measures. *Applied Psychological Measurement, 21*(2), 173 - 184.

Reise, S., Widaman, K., & Pugh, R. (1993). Confirmatory factor analysis and item response theory: Two approaches for exploring measurement invariance. *Psychological Bulletin, 114*(3), 552-566.

Richardson, C. G., Ratner, P. A., & Zumbo, B. D. (2007). A test of the age-based measurement invariance and temporal stability of Antonovsky's Sense of Coherence Scale. *Educational and Psychological Measurement, 67*(4), 679-696.

Risko, V. J., & Walker-Dalhouse, D. (2007). Tapping students' cultural funds of knowledge to address the achievement gap. (Reading research into the classroom). *The Reading Teacher, 61*(1), 98(93).

Ritz, C., & Brockhoff, P. B. (2005). Computing. Retrieved October 15, 2008, from http://www.imm.dtu.dk/~pbb/MAS/ST116

Rock, D. A., Werts, C. E., & Flaugher, R. L. (1978). The use of analysis of covariance structures for comparing the psychometric properties of multiple variables across populations [Article]. *Multivariate Behavioral Research, 13*, 403.

Sackett, P. R., Borneman, M. J., & Connelly, B. S. (2008). High stakes testing in higher education and employment: Appraising the evidence for validity and fairness. *American Psychologist, 63*(4), 215-227.

SAS Institute (2003). Statistical analysis software (Version 9.1.3). Cary, NC: SAS Institute, Inc.

Scroggins, W. A., Thomas, S. L., & Morris, J. A. (2008a). Psychological testing in personnel selection, part I: a century of psychological testing. *Public Personnel Management, 37*(1), 99(11).

Scroggins, W. A., Thomas, S. L., & Morris, J. A. (2008b). Psychological testing in personnel selection, part II: the refinement of methods and standards in employee selection. *Public Personnel Management, 37*(2), 185(114).

Shrestha, L. B. (2006). *The changing demographic profile of the United States*: Congressional Research Service, The Library of Congress.

Singh, J. (1995). Measurement issues in cross-national research. *Journal of International Business Studies, 26*(3), 597-619.

Snow, C. E., Burns, M. S., & Griffin, P. (1998). *Preventing reading difficulties in young children. [Report from the Committee on the Prevention of Reading Difficulties in Young Children]* (No. 030906418X). Washington, DC: National Academy Press.

South Carolina Department of Education (2008). State scores by demographic 2008 PACT Retrieved October 30, 2008, from http://ed.sc.gov/topics/assessment/scores/pact/2008/statescoresdemo.cfm

Sparks, R. L., Patton, J., Ganschow, L., Humbach, N., & Javorsky, J. (2008). Early first-language reading and spelling skills predict later second-language reading and spelling skills. *Journal of Educational Psychology, 100*(1), 162-174.

Stark, S., Chernyshenko, O., & Drasgow, F. (2006). Detecting differential item functioning with confirmatory factor analysis and item response theory: Toward a unified strategy. *Journal of Applied Psychology, 91*(6), 1292-1306.

Steenkamp, J., & Baumgartner, H. (1998). Assessing measurement invariance in cross-national consumer research. *The Journal of Consumer Research, 25*(1), 78-90.

Stone-Romero, E. F., Alliger, G. M., & Aguinis, H. (1994). Type II error problems in the use of moderated multiple regression for the detection of moderating effects of dichotomous variables. *Journal of Management, 20*(1), 167-178.

Supon, V. (2008). High-stakes testing: strategies by teachers and principals for student success. *Journal of Instructional Psychology, 35*(3), 306-308.

Suzuki, S., & Rancer, A. S. (1994). Argumentativeness and verbal aggressiveness: testing for conceptual and measurement equivalence across cultures. *Communication Monographs, 61*(3), 256.

Swaminathan, H., & Rogers, H. J. (1990). Detecting differential item functioning using logistic regression procedures. *Journal of Educational Measurement, 27*(4), 361-370.

Taylor, B. M., & Pearson, P. D. (Eds.). (2002). *Teaching reading : Effective schools, accomplished teachers*. Mahwah, N.J.: Lawrence Erlbaum Associates.

Thissen, D., Steinberg, L., & Gerrard, M. (1986). Beyond group-mean differences: The concept of item bias. *Psychological Bulletin, 99*(1), 118-128.

Thompson, G. B., & Fletcher-Flinn, C. M. (1993). A theory of knowledge sources and procedure for reading acquistion. In G. B. Thompson, W. E. Tunmer & T. Nicholson (Eds.), *Reading Acquisition Processes* (pp. 20 - 73). Cleveland, OH: Multilingual Matters LTD.

Topping, K. (2006). PISA/PIRLS data on reading achievement: Transfer into international policy and practice.(Programme for International Student Assessment, Progress in International Reading Literacy Study). *The Reading Teacher, 59*(6), 588(583).

United States Department of Education (2007). Growth models. Retrieved January 26, 2008, from http://www.ed.gov/admins/lead/account/growthmodel/index.html

United States Office of Management and Budget (1997). Revisions to the standards for the classification of federal data on race and ethnicity Retrieved October 30, 2008, from http://www.whitehouse.gov/omb/fedreg/1997standards.html

US Department of Education Center for Education Statistics (2006). *Digest of Education Statistics, 2005* (No. NCES 2006 - 030). Washington, DC: US Government Priniting Office.

Utsey, S. O., Brown, C., & Bolden, M. A. (2004). Testing the structural invariance of the africultural coping systems inventory across three samples of african descent populations. *Educational and Psychological Measurement, 64*(1), 185-195.

van der Linden, W. J., & Hambleton, R. K. (Eds.). (1997). *Handbook of modern item response theory*. New York: Springer.

Vandenberg, R., & Lance, C. (2000). A review and synthesis of the measurement invariance literature: Suggestions, practice, and recommendations for organizational research. *Organizational Research Methods, 3*, 4 - 69.

Vassend, O., & Skrondal, A. (1999). The problem of structural indeterminacy in multidimensional symptom report instruments. The case of SCL-90-R. *Behaviour Research and Therapy, 37*(7), 685-701.

Virginia Department of Education (2008). Virginia assessment results. Retrieved January 10, 2009, from https://p1pe.doe.virginia.gov/datareports/assess_test_result.do

Votruba-Drzal, E. (2006). Economic disparities in middle childhood development: Does income matter? *Developmental Psychology, 42*(6), 1154-1167.

Waller, N. G., Thompson, J. S., & Wenk, E. (2000). Using IRT to separate measurement bias from true group differences on homogeneous and heterogeneous scales: An illustration with the MMPI. *Psychological Methods, 5*(1), 125-146.

Wendorf, C. A. (2002). Comparisons of structural equation modeling and hierarchical linear modeling approaches to couples' data. *Structural Equation Modeling, 9*, 126 - 140.

Wicherts, J. M., Dolan, C. V., & Hessen, D. J. (2005). Stereotype threat and group differences in test performance: A question of measurement invariance. *Journal of Personality and Social Psychology, 89*(5), 696-716.

Willse, J. T., & Goodman, J. T. (2008). Comparison of multiple-indicators, multiple-causes and item response theory-based analyses of subgroup differences. *Educational and Psychological Measurement, 68*(4), 587-602.

Woods, C. M. (2008). Likelihood-ratio DIF testing: Effects of nonnormality. *Applied Psychological Measurement, 32*, 511 - 526.

Yeung, W. J., & Conley, D. (2008). Black-white achievement gap and family wealth. *Child Development, 79*(2), 303-324.

Yin, P., & Fan, X. (2003). Assessing The Factor Structure Invariance Of Self-Concept Measurement Across Ethnic And Gender Groups: Findings From A National Sample. *Educational and Psychological Measurement, 63*(2), 296-318.

Yoo, B. (2002). Cross-group comparisons: A cautionary note. *Psychology and Marketing, 19*(4), 357-368.

Yuan, K.-H., & Bentler, P. M. (2004). On chi-square difference and Z tests in mean and covariance structure analysis when the base model is misspecified. *Educational and Psychological Measurement, 64*(5), 737-757.

Zwick, R., Donoghue, J. R., & Grima, A. (1993). Assessment of differential item functioning for performance tasks. *Journal of Educational Measurement, 30*(3), 233-251.

Appendix A: HLM command file

```
#WHLM CMD FILE FOR ADMIN.MDM
nonlin:n
numit:5000000
stopval:0.0001000000
level1:RASCH=INTRCPT1+ITEM1+ITEM2+…+ITEM71+RANDOM
level2:INTRCPT1=INTRCPT2+random/
level3:INTRCPT2=INTRCPT3+random/
level2:ITEM1=INTRCPT2/
level3:INTRCPT2=INTRCPT3/
level2:ITEM2=INTRCPT2/
level3:INTRCPT2=INTRCPT3/
…
level2:ITEM71=INTRCPT2/
level3:INTRCPT2=INTRCPT3/
fixsigma2:1.000000
fixtau2:1
fixtau3:1
accel:5
level1weight:none
level2weight:none
level3weight:none
varianceknown:none
hypoth:n
resfiltype:sas
resfil1name:adminresidualsl1.sas
resfil1:y//RASCH/RASCH
resfil2name:adminresidualsl2.sas
resfil2:y/RASCH/RASCH
resfil3name:adminresidualsl3.sas
resfil3:y/RASCH
constrain:N
graphgammas:F:\ADMINgrapheq.geq
lvr-beta:n
title:ADMIN RESULTS
output:F:\ ADMINOUTPUT.txt
fulloutput:n
fishertype:2
```

Appendix B: Confirmatory factor analysis Lisrel/Simplis syntax for step 1

```
Spring 2005 Grade 3 - CFA
Group 1: Males
Observed Variables
ph vd rv rc
Latent Variables
Read
Correlation Matrix
1.00000
0.45098   1.00000
0.64999   0.45042   1.00000
0.64633   0.44899   0.70871   1.00000
Standard deviations:
4.06403 1.75353 2.30753 6.45518
Sample size: 1500
Relationships
ph = Read
vd = 1*Read
rv = Read
rc = Read

let the variance of Read be free

Group 2: Females
Correlation Matrix
1.00000
0.41670   1.00000
0.62464   0.45490   1.00000
0.62124   0.43636   0.65197   1.00000
Standard Deviations:
3.84698 1.66566 2.18666 6.22303
Sample size: 1403
Relationships
ph = Read
vd = 1*Read
rv = Read
rc = Read

let the error variance of ph be free
let the error variance of vd be free
let the error variance of rv be free
let the error variance of rc be free

let the variance of read be free

Method = Maximum Likelihood
lisrel output ND=5 ss rs ef tv se sc mi
PATH DIAGRAM
end of problem
```

Appendix C: Confirmatory factor analysis Lisrel/Simplis syntax for step 2

```
Spring 2005 Grade 3 - CFA
Group 1: Males
Observed Variables
ph vd rv rc
Latent Variables
Read
Correlation Matrix
1.00000
0.45098    1.00000
0.64999    0.45042    1.00000
0.64633    0.44899    0.70871    1.00000
Standard deviations:
4.06403 1.75353 2.30753 6.45518
Sample size: 1500
Relationships
ph = Read
vd = 1*Read
rv = Read
rc = Read

let the variance of Read be free

Group 2: Females
Correlation Matrix
1.00000
0.41670    1.00000
0.62464    0.45490    1.00000
0.62124    0.43636    0.65197    1.00000
Standard Deviations:
3.84698 1.66566 2.18666 6.22303
Sample size: 1403
Relationships
!ph = Read
!vd = 1*Read
!rv = Read
!rc = Read

let the error variance of ph be free
let the error variance of vd be free
let the error variance of rv be free
let the error variance of rc be free

let the variance of read be free

Method = Maximum Likelihood
lisrel output ND=5 ss rs ef tv se sc mi
PATH DIAGRAM
end of problem
```

Appendix D: Confirmatory factor analysis Lisrel/Simplis syntax for step 3

```
Spring 2005 Grade 3 - CFA
Group 1: Males
Observed Variables
ph vd rv rc
Latent Variables
Read
Correlation Matrix
1.00000
0.45098    1.00000
0.64999    0.45042    1.00000
0.64633    0.44899    0.70871    1.00000
Standard deviations:
4.06403 1.75353 2.30753 6.45518
Sample size: 1500
Relationships
ph = Read
vd = 1*Read
rv = Read
rc = Read

let the variance of Read be free

Group 2: Females
Correlation Matrix
1.00000
0.41670    1.00000
0.62464    0.45490    1.00000
0.62124    0.43636    0.65197    1.00000
Standard Deviations:
3.84698 1.66566 2.18666 6.22303
Sample size: 1403
Relationships
!ph = Read
!vd = 1*Read
!rv = Read
!rc = Read

!let the error variance of ph be free
!let the error variance of vd be free
```

```
!let the error variance of rv be free
!let the error variance of rc be free

let the variance of read be free

Method = Maximum Likelihood
lisrel output ND=5 ss rs ef tv se sc mi
PATH DIAGRAM
end of problem
```

Appendix E: Winsteps Control File

```
&INST
  TITLE = All Observations
 PERSON = Person ; persons are ...
   ITEM = Item ; items are ...
  ITEM1 = 4 ; column of response to first item in data record
     NI = 72 ; number of items
  NAME1 = 1 ; column of first character of person identifying label
NAMELEN = 3 ; length of person label
  XWIDE = 1 ; number of columns per item response
  CODES = 01 ; valid codes in data file
 UIMEAN = 0 ; item mean for local origin
 USCALE = 1 ; user scaling for logits
 UDECIM = 2 ; reported decimal places for user scaling

&END

END LABELS
```

Appendix F: Personal communication about Rasch linking protocols

From: HUYNH, HUYNH
Sent: Sunday, March 22, 2009 9:01 PM

Subject: The Golden Numbers in Rasch Linking Protocol

Dear TAC members:

Rasch linking protocols (used in SC, MD, AR, and other states too) call for four specific numbers (1.645 for robust z; correlation of .95; ratio of SD between .9 and 1.1; and deleting no more than 20% of all potential linking items. Although these operational benchmarks work well across the years, TAC members often ask for some documentation regarding these numbers. Here are some statistical and contextual reasons.

1. **Pool of "unstable" items**. These items are defined as having absolute robust z greater then 1.645 (10% two-tailed significance). The chapter by Huynh Huynh and Anita Rawls (2009) show that these items are almost identical to those identified by the time-tested Harcourt ".3 logit discrepancy" rule for several sets of SC assessment data.
2. **Correlation of .95**. Many studies on IRT item recovery of location parameter (Yen, 1987) show that the correlation between the true location parameters and their estimates is better than .97 in many simulated cases. This correlation is a "validity" coefficient. Taking the square (to get .94) will yield a "reliability" coefficient. So the correlation between independent replications is better than .94 in many situations. Rounded this number to something easier to remember, you will get .95.
3. **Ratio of SD between 0.9 and 1.1**. These bounds are a touch more "statistical" than the others. This has to do with the ML test for equality of two dependent variances. Assume that there are 30 linking items and the correlation between the Rasch values is .95 and the alpha level is 5%. Then the null hypothesis of equality of dependent variances is accepted if their observed ratio is in the range from 0.88 to 1.13. Rounding these to the nearest tenth, you will get 0.9 and 1.1.
4. **Bound for number of deleted items**. Assuming N = 30 potential linking items. In a number of testing programs, these items cover about six content strands, each with 5 items on the average. It seems reasonable that we should not delete more than one item per strand. So the 20% rule was born out of this content consideration.

So there are justifications for those four thresholds that underline the Rasch linking protocols used in Arkansas and other state testing programs.

Appendix G: Estimation of fixed effects for administration model

Fixed Effects	Coefficient	Standard Error	T-ratio	*df*	P-value
γ_{000}	0.71	0.01	64.55	2	0.00
γ_{100}	-0.12	1.00	-0.12	8230	0.91
γ_{200}	-0.01	1.00	-0.01	8230	1.00
γ_{300}	0.18	1.00	0.18	8230	0.86
γ_{400}	0.02	1.00	0.02	8230	0.98
γ_{500}	0.20	1.00	0.20	8230	0.84
γ_{600}	0.05	1.00	0.05	8230	0.96
γ_{700}	-0.01	1.00	-0.01	8230	1.00
γ_{800}	-0.14	1.00	-0.14	8230	0.89
γ_{900}	-0.01	1.00	-0.01	8230	1.00
γ_{1000}	-0.05	1.00	-0.05	8230	0.96
γ_{1100}	0.10	1.00	0.10	8230	0.92
γ_{1200}	0.04	1.00	0.04	8230	0.97
γ_{1300}	-0.02	1.00	-0.02	8230	0.99
γ_{1400}	-0.07	1.00	-0.07	8230	0.94
γ_{1500}	0.16	1.00	0.16	8230	0.87
γ_{1600}	-0.03	1.00	-0.03	8230	0.97
γ_{1700}	-0.08	1.00	-0.08	8230	0.93
γ_{1800}	0.22	1.00	0.22	8230	0.83
γ_{1900}	-0.07	1.00	-0.07	8230	0.94
γ_{2000}	0.13	1.00	0.13	8230	0.90
γ_{2100}	0.37	1.00	0.37	8230	0.72
γ_{2200}	-0.13	1.00	-0.13	8230	0.90
γ_{2300}	0.10	1.00	0.10	8230	0.92
γ_{2400}	0.04	1.00	0.04	8230	0.97
γ_{2500}	0.04	1.00	0.04	8230	0.97
γ_{2600}	-0.12	1.00	-0.12	8230	0.91

Fixed Effects	Coefficient	Standard Error	T-ratio	*df*	P-value
γ_{2800}	0.33	1.00	0.33	8230	0.74
γ_{2900}	0.05	1.00	0.05	8230	0.96
γ_{3000}	0.01	1.00	0.01	8230	0.99
γ_{3100}	-0.08	1.00	-0.08	8230	0.93
γ_{3200}	0.31	1.00	0.31	8230	0.76
γ_{3300}	-0.09	1.00	-0.10	8230	0.93
γ_{3400}	-0.02	1.00	-0.02	8230	0.99
γ_{3500}	0.20	1.00	0.20	8230	0.84
γ_{3600}	0.16	1.00	0.16	8230	0.87
γ_{3700}	0.16	1.00	0.16	8230	0.87
γ_{3800}	-0.24	1.00	-0.24	8230	0.81
γ_{3900}	-0.16	1.00	-0.16	8230	0.87
γ_{4000}	-0.11	1.00	-0.11	8230	0.92
γ_{4100}	0.11	1.00	0.11	8230	0.91
γ_{4200}	0.35	1.00	0.35	8230	0.73
γ_{4300}	0.05	1.00	0.05	8230	0.96
γ_{4400}	0.01	1.00	0.01	8230	0.99
γ_{4500}	0.10	1.00	0.10	8230	0.92
γ_{4600}	0.27	1.00	0.27	8230	0.79
γ_{4700}	0.08	1.00	0.08	8230	0.94
γ_{4800}	-0.16	1.00	-0.16	8230	0.87
γ_{4900}	-0.23	1.00	-0.23	8230	0.82
γ_{5000}	-0.01	1.00	-0.01	8230	1.00
γ_{5100}	-0.06	1.00	-0.06	8230	0.95
γ_{5200}	-0.18	1.00	-0.18	8230	0.86
γ_{5300}	0.04	1.00	0.04	8230	0.97
γ_{5400}	0.13	1.00	0.13	8230	0.90
γ_{5500}	-0.28	1.00	-0.28	8230	0.78
γ_{5600}	-0.13	1.00	-0.13	8230	0.90
γ_{5700}	-0.21	1.00	-0.21	8230	0.83
γ_{5800}	0.11	1.00	0.11	8230	0.91
γ_{5900}	0.07	1.00	0.07	8230	0.95

Fixed Effects	Coefficient	Standard Error	T-ratio	*df*	P-value
γ_{6100}	0.15	1.00	0.15	8230	0.88
γ_{6200}	-0.05	1.00	-0.05	8230	0.96
γ_{6300}	-0.13	1.00	-0.13	8230	0.90
γ_{6400}	-0.17	1.00	-0.17	8230	0.87
γ_{6500}	-0.16	1.00	-0.16	8230	0.87
γ_{6600}	0.01	1.00	0.01	8230	0.99
γ_{6700}	0.13	1.00	0.13	8230	0.90
γ_{6800}	-0.17	1.00	-0.17	8230	0.87
γ_{6900}	-0.11	1.00	-0.11	8230	0.92
γ_{7000}	-0.02	1.00	-0.02	8230	0.99
γ_{7100}	-0.11	1.00	-0.11	8230	0.92

Appendix H: Estimation of fixed effects for gender model

Fixed Effects	Coefficient	Standard Error	T-ratio	*df*	P-value
γ_{000}	0.71	0.02	41.69	Unable to	compute
γ_{100}	-0.10	1.00	-0.10	8230	0.92
γ_{200}	0.01	1.00	0.01	8230	0.99
γ_{300}	0.20	1.00	0.20	8230	0.85
γ_{400}	0.04	1.00	0.04	8230	0.97
γ_{500}	0.21	1.00	0.21	8230	0.83
γ_{600}	0.07	1.00	0.07	8230	0.95
γ_{700}	0.01	1.00	0.01	8230	0.99
γ_{800}	-0.12	1.00	-0.12	8230	0.90
γ_{900}	0.01	1.00	0.01	8230	0.99
γ_{1000}	-0.03	1.00	-0.03	8230	0.98
γ_{1100}	0.11	1.00	0.11	8230	0.91
γ_{1200}	0.05	1.00	0.05	8230	0.96
γ_{1300}	0.00	1.00	-0.01	8230	1.00
γ_{1400}	-0.06	1.00	-0.06	8230	0.96
γ_{1500}	0.18	1.00	0.18	8230	0.86
γ_{1600}	-0.02	1.00	-0.02	8230	0.99
γ_{1700}	-0.07	1.00	-0.07	8230	0.95
γ_{1800}	0.23	1.00	0.23	8230	0.82
γ_{1900}	-0.06	1.00	-0.06	8230	0.96
γ_{2000}	0.14	1.00	0.14	8230	0.89
γ_{2100}	0.38	1.00	0.38	8230	0.70
γ_{2200}	-0.11	1.00	-0.11	8230	0.91
γ_{2300}	0.11	1.00	0.11	8230	0.91
γ_{2400}	0.05	1.00	0.05	8230	0.96
γ_{2500}	0.05	1.00	0.05	8230	0.96
γ_{2600}	-0.10	1.00	-0.10	8230	0.92

Fixed Effects	Coefficient	Standard Error	T-ratio	*df*	P-value
γ_{2800}	0.34	1.00	0.34	8230	0.73
γ_{2900}	0.07	1.00	0.07	8230	0.95
γ_{3000}	0.02	1.00	0.02	8230	0.98
γ_{3100}	-0.07	1.00	-0.07	8230	0.95
γ_{3200}	0.32	1.00	0.32	8230	0.75
γ_{3300}	-0.08	1.00	-0.08	8230	0.94
γ_{3400}	0.00	1.00	-0.01	8230	1.00
γ_{3500}	0.21	1.00	0.21	8230	0.83
γ_{3600}	0.18	1.00	0.18	8230	0.86
γ_{3700}	0.18	1.00	0.18	8230	0.86
γ_{3800}	-0.23	1.00	-0.23	8230	0.82
γ_{3900}	-0.14	1.00	-0.14	8230	0.89
γ_{4000}	-0.09	1.00	-0.09	8230	0.93
γ_{4100}	0.13	1.00	0.13	8230	0.90
γ_{4200}	0.36	1.00	0.36	8230	0.72
γ_{4300}	0.07	1.00	0.07	8230	0.95
γ_{4400}	0.02	1.00	0.02	8230	0.98
γ_{4500}	0.11	1.00	0.11	8230	0.91
γ_{4600}	0.28	1.00	0.29	8230	0.78
γ_{4700}	0.10	1.00	0.10	8230	0.92
γ_{4800}	-0.14	1.00	-0.14	8230	0.89
γ_{4900}	-0.21	1.00	-0.21	8230	0.83
γ_{5000}	0.01	1.00	0.01	8230	0.99
γ_{5100}	-0.04	1.00	-0.04	8230	0.97
γ_{5200}	-0.16	1.00	-0.16	8230	0.87
γ_{5300}	0.05	1.00	0.05	8230	0.96
γ_{5400}	0.14	1.00	0.14	8230	0.89
γ_{5500}	-0.26	1.00	-0.26	8230	0.79
γ_{5600}	-0.11	1.00	-0.11	8230	0.91
γ_{5700}	-0.20	1.00	-0.20	8230	0.84
γ_{5800}	0.13	1.00	0.13	8230	0.90
γ_{5900}	0.08	1.00	0.08	8230	0.94
γ_{6000}	-0.19	1.00	-0.19	8230	0.85

Fixed Effects	Coefficient	Standard Error	T-ratio	*df*	P-value
γ_{6100}	0.16	1.00	0.16	8230	0.87
γ_{6200}	-0.03	1.00	-0.03	8230	0.98
γ_{6300}	-0.11	1.00	-0.11	8230	0.91
γ_{6400}	-0.15	1.00	-0.15	8230	0.88
γ_{6500}	-0.14	1.00	-0.14	8230	0.89
γ_{6600}	0.02	1.00	0.02	8230	0.98
γ_{6700}	0.14	1.00	0.14	8230	0.89
γ_{6800}	-0.15	1.00	-0.15	8230	0.88
γ_{6900}	-0.09	1.00	-0.09	8230	0.93
γ_{7000}	0.00	1.00	-0.01	8230	1.00
γ_{7100}	-0.09	1.00	-0.09	8230	0.93

Appendix I: Estimation of fixed effects for ethnicity model

Fixed Effects	Coefficient	Standard Error	T-ratio	*df*	P-value
γ_{000}	0.72	0.02	30.68	2	0.00
γ_{100}	-0.13	1.00	-0.13	8230	0.90
γ_{200}	-0.02	1.00	-0.02	8230	0.98
γ_{300}	0.17	1.00	0.17	8230	0.87
γ_{400}	0.01	1.00	0.01	8230	0.99
γ_{500}	0.18	1.00	0.18	8230	0.85
γ_{600}	0.04	1.00	0.04	8230	0.97
γ_{700}	-0.02	1.00	-0.02	8230	0.98
γ_{800}	-0.15	1.00	-0.15	8230	0.88
γ_{900}	-0.02	1.00	-0.02	8230	0.98
γ_{1000}	-0.06	1.00	-0.06	8230	0.95
γ_{1100}	0.08	1.00	0.08	8230	0.93
γ_{1200}	0.02	1.00	0.02	8230	0.98
γ_{1300}	-0.03	1.00	-0.03	8230	0.97
γ_{1400}	-0.08	1.00	-0.09	8230	0.93
γ_{1500}	0.15	1.00	0.15	8230	0.88
γ_{1600}	-0.05	1.00	-0.05	8230	0.96
γ_{1700}	-0.10	1.00	-0.10	8230	0.92
γ_{1800}	0.20	1.00	0.20	8230	0.84
γ_{1900}	-0.08	1.00	-0.09	8230	0.93
γ_{2000}	0.12	1.00	0.12	8230	0.91
γ_{2100}	0.35	1.00	0.35	8230	0.73
γ_{2200}	-0.14	1.00	-0.14	8230	0.89
γ_{2300}	0.08	1.00	0.08	8230	0.93
γ_{2400}	0.02	1.00	0.02	8230	0.98
γ_{2500}	0.02	1.00	0.02	8230	0.98
γ_{2600}	-0.13	1.00	-0.13	8230	0.90

Fixed Effects	Coefficient	Standard Error	T-ratio	*df*	P-value
γ_{2700}	-0.22	1.00	-0.22	8230	0.83
γ_{2800}	0.31	1.00	0.31	8230	0.75
γ_{2900}	0.04	1.00	0.04	8230	0.97
γ_{3000}	-0.01	1.00	-0.01	8230	1.00
γ_{3100}	-0.10	1.00	-0.10	8230	0.92
γ_{3200}	0.29	1.00	0.29	8230	0.77
γ_{3300}	-0.11	1.00	-0.11	8230	0.91
γ_{3400}	-0.03	1.00	-0.03	8230	0.97
γ_{3500}	0.18	1.00	0.18	8230	0.85
γ_{3600}	0.15	1.00	0.15	8230	0.88
γ_{3700}	0.15	1.00	0.15	8230	0.88
γ_{3800}	-0.26	1.00	-0.26	8230	0.80
γ_{3900}	-0.17	1.00	-0.17	8230	0.86
γ_{4000}	-0.12	1.00	-0.12	8230	0.91
γ_{4100}	0.10	1.00	0.10	8230	0.92
γ_{4200}	0.33	1.00	0.33	8230	0.74
γ_{4300}	0.04	1.00	0.04	8230	0.97
γ_{4400}	-0.01	1.00	-0.01	8230	1.00
γ_{4500}	0.08	1.00	0.08	8230	0.93
γ_{4600}	0.26	1.00	0.26	8230	0.80
γ_{4700}	0.07	1.00	0.07	8230	0.95
γ_{4800}	-0.17	1.00	-0.17	8230	0.86
γ_{4900}	-0.24	1.00	-0.24	8230	0.81
γ_{5000}	-0.02	1.00	-0.02	8230	0.98
γ_{5100}	-0.07	1.00	-0.07	8230	0.94
γ_{5200}	-0.19	1.00	-0.19	8230	0.85
γ_{5300}	0.02	1.00	0.02	8230	0.98
γ_{5400}	0.12	1.00	0.12	8230	0.91
γ_{5500}	-0.29	1.00	-0.29	8230	0.77
γ_{5600}	-0.14	1.00	-0.14	8230	0.89
γ_{5700}	-0.23	1.00	-0.23	8230	0.82
γ_{5800}	0.10	1.00	0.10	8230	0.92
γ_{5900}	0.05	1.00	0.05	8230	0.96

Fixed Effects	Coefficient	Standard Error	T-ratio	*df*	P-value
γ_{6100}	0.13	1.00	0.13	8230	0.90
γ_{6200}	-0.06	1.00	-0.06	8230	0.95
γ_{6300}	-0.14	1.00	-0.14	8230	0.89
γ_{6400}	-0.18	1.00	-0.18	8230	0.86
γ_{6500}	-0.17	1.00	-0.17	8230	0.86
γ_{6600}	-0.01	1.00	-0.01	8230	1.00
γ_{6700}	0.12	1.00	0.12	8230	0.91
γ_{6800}	-0.18	1.00	-0.18	8230	0.86
γ_{6900}	-0.12	1.00	-0.12	8230	0.91
γ_{7000}	-0.03	1.00	-0.03	8230	0.97
γ_{7100}	-0.12	1.00	-0.12	8230	0.91